EFFICIENT INFERENCE FOR DISTRIBUTIONAL TREATMENT EFFECTS WITH PARTIALLY OBSERVED OUTCOMES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the distributional treatment effect when a surrogate outcome is fully observed, but the primary outcome is observed only for a subsample. This setting arises naturally in model or agent evaluation, where automated judges make it inexpensive to score every high-dimensional response while human ratings are often available for only a small subsample. When evaluation scores are discrete or ordinal, a difference between cumulative distribution functions (CDFs) directly measures changes in the probability of low-quality outputs at each rating cutpoint. Building on prediction-powered inference and surrogate-assisted treatment-effect estimation, we develop a cross-fitted estimator that combines conditional CDFs with and without surrogate information. We establish a Gaussian-process limit and uniform confidence bands over outcome thresholds, identify the semiparametric efficiency bound, and characterize variance gains over estimators that reweight the labeled subsample by its known sampling probabilities or adjust for covariates without the surrogate. The construction does not require continuous output density or agreement between the surrogate and primary-outcome scales. Through simulation results and four real data applications, we demonstrate the effectiveness of our approach.

1 INTRODUCTION

Randomized controlled trials (RCTs) provide a direct way to compare a new system with an existing one. In LLM applications, the intervention might be a change of model, prompt, or response-generation policy. Obtaining human ratings is costly: each output must be read and judged, sometimes by a subject matter expert. LLM judges can score responses at much larger scale (Zheng et al., 2023; Kim et al., 2024), making it practical to score every response automatically and obtain human ratings for a small subset. The challenge is to use automated scores to estimate an intervention's effect on *human-rated quality*, even when the judge makes systematic errors.

A growing body of research uses predictions and imperfect annotations to support statistical inference. Prediction-powered inference (PPI), for example, combines predictions with a smaller labeled sample for inference on quantities such as means and quantiles (Angelopoulos et al., 2023a;b). The observed outcomes provide a correction for prediction errors. Related approaches include design-based inference with imperfect LLM annotations (Egami et al., 2023) and surrogate-assisted treatment-effect estimation when primary outcomes are only partially observed (Kallus & Mao, 2025). These methods provide a basis for using automated ratings without treating them as substitutes for human observations.

The choice of estimand also matters. Although the average treatment effect (ATE) is a common target, an average can conceal changes elsewhere in the outcome distribution (Jiang et al., 2023; Byambadalai et al., 2024). Many LLM evaluations use *ordinal ratings*: MT-Bench uses a 1–10 judge scale, Prometheus's Feedback Collection uses integer scores from 1 to 5, and HelpSteer2 provides human ratings from 0 to 4 (Zheng et al., 2023; Kim et al., 2024; Wang et al., 2024). An evaluator may want to know whether an update reduces poor responses or increases the fraction meeting an acceptable-quality threshold. A *distributional treatment effect* (DTE) answers this question by comparing the probability of a rating at or below a cutpoint between treatment arms. Taken over

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.

all nontrivial cutpoints, these contrasts describe the entire change in the ordinal distribution and are invariant to increasing recodings of the rating scale. Quantile treatment effects (QTEs) offer another view of distributional change, but regression-adjusted methods such as Jiang et al. (2023) assume continuous outcomes. DTEs accommodate discrete and mixed outcomes directly. They can also reveal changes that a particular quantile misses: two arms may both have median rating 3 even if an update reduces the share rated at most 1 from 20% to 10%. The median effect is zero; the DTE at 1 records a 10-percentage-point reduction. No continuous outcome density is needed.

We develop a cross-fitted DTE estimator for experiments with sparsely observed primary outcomes, allowing labeling to depend on observed information and the surrogate and primary-outcome scales to differ. To quantify uncertainty across thresholds, we establish a functional central limit theorem and uniform confidence bands. The efficiency bound and variance decompositions characterize the precision available from surrogate information beyond covariate adjustment. We assess inferential validity in simulations with known truth, comparing against thresholdwise PPI++ and labeled-only regression adjustment. An offline randomized model comparison on BiGGen Bench then evaluates performance with human ratings, while HelpSteer2, wine reviews, and ACTG 175 provide further precision comparisons.

2 Related Work

Prediction-powered and surrogate-assisted inference. Prediction-powered inference (PPI) combines a small labeled sample with predictions available for a larger sample, using the observed outcomes to correct prediction errors and construct confidence intervals for population quantities (Angelopoulos et al., 2023a). PPI++ tunes the predictions' contribution for efficiency (Angelopoulos et al., 2023b); when predictors must be learned from the labeled sample, cross-PPI uses cross-fitting (Zrnic & Candès, 2024). In LLM applications, design-based supervised learning corrects imperfect annotations (Egami et al., 2023), while hybrid evaluation combines prediction-powered inference with stratified designs (Boyeau et al., 2025; Fisch et al., 2024). The statistical use of unlabeled observations has a longer history in semi-supervised inference (Chakrabortty & Cai, 2018; Zhang et al., 2019). Recent work connects this literature to predictions as surrogate outcomes (Ji et al., 2025) and general two-phase estimation (Kluger & Bates, 2026).

The closest causal-inference foundation is Kallus & Mao (2025), who derive efficient surrogate-assisted ATE estimation with limited outcome data. Applying a PPI mean procedure thresholdwise also yields distributional estimators, motivating our baselines. Our focus is joint distributional inference in one population under positive MAR labeling probabilities. This differs from transported DTE/QTE estimation (Wang, 2026) and long-term QTE data fusion (Liu et al., 2026), which connect source and target populations through transportability assumptions. We build on the pointwise foundations to establish a functional limit and uniform bands, with variance decompositions that isolate the surrogate's contribution.

Distributional and quantile treatment effects. Distributional estimations such as DTEs and QTEs describe treatment effects beyond the mean (Doksum, 1974; Lehmann & D'Abrera, 1975; Bitler et al., 2006). Identification and testing have been studied for complier outcome distributions (Imbens & Rubin, 1997; Abadie, 2002) and counterfactual distributions under nonlinear difference-in-differences designs (Athey & Imbens, 2006). For inference, Firpo (2007) establishes efficient QTE estimation, while Chernozhukov et al. (2013) develop functional-limit and bootstrap theory for counterfactual distributions. Distribution-function bands can also be used for quantile-effect inference with discrete outcomes (Chernozhukov et al., 2020). High-dimensional and localized debiased methods accommodate flexible nuisance estimation (Belloni et al., 2017; Kallus et al., 2024). Our labeled-only baseline follows the regression-adjusted DTE estimator of Byambadalai et al. (2024). We extend the scope of distributional inference to partially observed primary outcomes, using fully observed surrogates to account for selective labeling and improve precision.

3 Observation Model and Estimand

For each unit, observe $O = (X, W, \widetilde{Y}, R, RY)$, where X denotes pre-treatment covariates, $W \in \{0, 1\}$ the evaluation arm, $\widetilde{Y}$ a surrogate available for everyone, Y the primary outcome of interest,

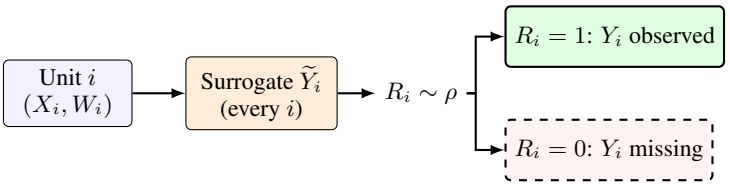


Figure 1: Observation model. Every unit contributes covariates X, arm W, and a cheap surrogate $\widetilde{Y}$; the expensive primary outcome Y is observed only on the labeled subsample ($R = 1$), with labeling probability $\rho(X, W, \widetilde{Y})$ (MAR, Assumption 2). The proposed estimator combines both subsamples through a cross-fitted score that uses two conditional CDFs of the primary outcome, $\mu_w(y, X)$ and $\nu_w(y, X, \widetilde{Y})$, together with the labeling and treatment probabilities.

and R its observation indicator. The notation RY is only an observation convention: no imputed value is treated as an observed primary outcome. We present two arms for clarity; the construction applies to any pair of arms with overlap. We adopt the potential-outcomes framework (Rubin, 1974), with consistency $Y = Y(W)$. The surrogate may be affected by treatment, for example a judge score of the generated response. Treatment exchangeability and the propensity score condition on pre-treatment X only; the surrogate is not included in the treatment propensity score. The estimand concerns Y alone; $\widetilde{Y}$ is an auxiliary measurement for recovering missing primary outcomes, rather than a second outcome whose treatment effect is targeted.

Fix a nonrandom threshold set $\mathcal{Y} \subseteq \mathbb{R}$, which may be an interval, the whole line, or a finite grid $\{y_1 < \cdots < y_J\}$. Our target is

$$F_w(y) = P\{Y(w) \le y\}, \qquad \Delta(y) = F_1(y) - F_0(y), \quad y \in \mathcal{Y}. \tag{1}$$

For higher-is-better ratings, a negative contrast means fewer low-rated outcomes in arm 1. This is a contrast of marginal distributions, not the distribution of individual effects $Y(1) - Y(0)$.

Assumption 1 (Sampling, exchangeability, and overlap)**.** *The full unit records are i.i.d.;* $Y(w) \perp W \mid X$ *for each arm; and* $\pi_w(X) := P(W = w \mid X) \ge \epsilon > 0$ *almost surely.*

Independent randomized assignment implies exchangeability. In an observational application this assumption is substantive and cannot be verified by computing outcomes on a fully labeled sample.

Assumption 2 (Label observability and selection)**.** *The surrogate is observed for every unit,* $R \perp Y \mid (X, W, \widetilde{Y})$*, and* $\rho(X, W, \widetilde{Y}) := P(R = 1 \mid X, W, \widetilde{Y}) \ge \rho_{\min} > 0$ *almost surely.*

Random annotation with a fixed positive sampling fraction is a special case. The more general MAR model permits oversampling particular surrogate scores. Neither accurate surrogate predictions nor equality of the primary-outcome and surrogate scales is assumed. Positivity is fixed as $n \to \infty$; the results do not cover a labeling fraction vanishing with n.

Define the two conditional CDFs of the primary outcome

$$\mu_w(y, x) = P(Y \le y \mid X = x, W = w), \quad \nu_w(y, x, s) = P(Y \le y \mid X = x, W = w, \widetilde{Y} = s). \tag{2}$$

Under MAR, ν_w is identified from labeled observations conditional on $(X, W, \widetilde{Y})$. Thus the surrogate can support identification under selective labeling as well as improve precision; it is not a treatment-adjustment covariate. The tower property and exchangeability then give

$$F_w(y) = \mathbb{E}_X[\mu_w(y, X)] = \mathbb{E}_X\left[\mathbb{E}\{\nu_w(y, X, \widetilde{Y}) \mid X, W = w\}\right]. \tag{3}$$

In empirical observational examples we use the standardized functional $\mathbb{E}_X[\mu_1(y, X) - \mu_0(y, X)]$ without claiming that exchangeability makes it causal. This functional is defined directly by the observed distribution and retains the same estimation formula under MAR.

4 SURROGATE-AUGMENTED DISTRIBUTION REGRESSION

For $Z_y = \mathbf{1}\{Y \le y\}$, define the score

$$\Psi_w(y; O, \eta) = \mu_w(y, X) + \frac{\mathbf{1}\{W = w\}}{\pi_w(X)}\{\nu_w(y, X, \widetilde{Y}) - \mu_w(y, X)\} + \frac{\mathbf{1}\{W = w\}R}{\pi_w(X)\rho(X, W, \widetilde{Y})}\{Z_y - \nu_w(y, X, \widetilde{Y})\}. \tag{4}$$

Here $\eta = (\mu, \nu, \pi, \rho)$ collects the nuisances. Only the last term uses observed primary outcomes. The first term standardizes over the target covariate distribution; the second integrates the surrogate-conditional prediction over the observed arm; the third corrects its prediction errors with labeled data. No comparison of $\widetilde{Y}$ and a threshold on the primary-outcome scale is needed.

Split the observations into a fixed number K of approximately equal folds. Train all nuisances outside each evaluation fold and let $\widehat{\eta}^{-k(i)}$ denote the nuisance fits for unit i. The estimator is

$$\widehat{F}_w(y) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Psi}_{wi}(y), \qquad \widehat{\Delta}(y) = \widehat{F}_1(y) - \widehat{F}_0(y), \qquad \widehat{\Psi}_{wi}(y) = \Psi_w(y; O_i, \widehat{\eta}^{-k(i)}). \tag{5}$$

When randomization or annotation probabilities are known by design, they can be inserted directly into the augmented score, eliminating the corresponding nuisance-estimation requirements. Cross-fitting avoids evaluating an observed primary outcome with a nuisance fit trained on that observation.

Algorithm 1: Distributional inference with partially observed primary outcomes
Input: $(X_i, W_i, \widetilde{Y}_i, R_i, R_iY_i)_{i=1}^n$, primary-outcome cutpoints, and level $1 - \alpha$.

1. Split units into K folds. Outside each evaluation fold, fit π from (X, W) and ρ from $(X, W, \widetilde{Y}, R)$, or use supplied design probabilities.
2. For each arm and cutpoint, use labeled training units to regress $\mathbf{1}\{Y \le y\}$ on $(X, \widetilde{Y})$ for $\widehat{\nu}_w$, and on X with weights $1/\widehat{\rho}$ for $\widehat{\mu}_w$.
3. Evaluate scores (4) on held-out units. Pool all folds and average arm-score differences to obtain $\widehat{\Delta}(y)$ in (5).
4. Center these differences and compute SEs by (9). Use common multipliers across cutpoints to compute $T^{(b)}$, its empirical $(1 - \alpha)$-quantile, and bands by (10).

Output: DTE estimates and simultaneous bands over the evaluated cutpoints. Implementation safeguards are in Appendix A.

Estimating the outcome CDF conditional on covariates under MAR. An unweighted regression of Z_y on X in the labeled arm generally estimates $P(Y \le y \mid X, W = w, R = 1)$, not μ_w. We instead fit $\widehat{\mu}_w$ using inverse-label weights $1/\widehat{\rho}(X, w, \widetilde{Y})$ on labeled training units. With the true weights,

$$\mathbb{E}\left[\frac{R}{\rho}\{Z_y - m(X)\}\,\middle|\, X, W = w\right] = \mu_w(y, X) - m(X).$$

A sufficiently rich weighted regression can therefore recover μ_w. Alternatively, regress a suitably cross-fitted estimate of ν_w on X using *all* training observations in arm w. Algorithm 1 uses weighted regression for μ_w and unweighted labeled regression for ν_w.

Robustness and nuisance learning. For fixed candidate nuisances $\bar{\eta}$ with positive weights, the score bias has the exact form

$$\mathbb{E}[\Psi_w(y; O, \bar{\eta})] - F_w(y) = \mathbb{E}\left[(\bar{\mu}_w - \mu_w)\frac{\bar{\pi}_w - \pi_w}{\bar{\pi}_w}\right] + \mathbb{E}\left[\frac{\mathbf{1}\{W = w\}}{\bar{\pi}_w}(\bar{\nu}_w - \nu_w)\frac{\bar{\rho} - \rho}{\bar{\rho}}\right]. \tag{6}$$

Arguments are suppressed for readability. This proves Neyman orthogonality and specifies the robustness combinations: each of the two products must vanish. For example, correct design weights make the score unbiased for any fixed outcome regressions, and correct outcome regressions make it unbiased for any positive candidate weights. Efficiency additionally requires learning the true conditional CDFs. Arbitrary surrogate quality is permitted.

5 Uniform Inference and Efficiency

Write $\varphi_w(y;O) = \Psi_w(y;O,\eta) - F_w(y)$ and $\varphi = \varphi_1 - \varphi_0$. Norms integrate over a fresh $O \sim P$, holding the training sample fixed; fitted-function norms are therefore random. Let $\mathbb{G}_n f = \sqrt{n}(P_n - P)f$ and $\Omega(y,y') = \mathbb{E}[\varphi(y;O)\varphi(y';O)]$.

Assumption 3 (Uniform nuisance rates and score complexity)**.** *For every fold and arm, fitted CDF values lie in* $[0,1]$*,* $\widehat{\pi}_w \geq \epsilon/2$*, and* $\widehat{\rho} \geq \rho_{\min}/2$*, with fitted probabilities at most one. All nuisances are* $L^2(P)$ *consistent uniformly over* $\mathcal{Y}$*, and*

$$\sup_{y\in\mathcal{Y}} \|\widehat{\mu}_w - \mu_w\|_2 \|\widehat{\pi}_w - \pi_w\|_2 + \sup_{y\in\mathcal{Y}} \|\widehat{\nu}_w - \nu_w\|_2 \|\widehat{\rho} - \rho\|_2 = o_P(n^{-1/2}). \tag{7}$$

The number of folds is fixed, with positive limiting fold proportions. Conditional on each training sample, the fitted score class $\widehat{\mathcal{F}}_{kw} = \{\Psi_w(y;\cdot,\widehat{\eta}^{-k}) : y \in \mathcal{Y}\}$ *is pointwise measurable and, with probability tending to one, satisfies*

$$\sup_Q N(uM, \widehat{\mathcal{F}}_{kw}, L^2(Q)) \leq (A/u)^v, \qquad 0 < u \leq 1, \tag{8}$$

for fixed constants $A \geq e, v, M < \infty$*, with* M *also a constant envelope;* Q *ranges over finitely supported probability measures. The indexed processes are separable, with measurable suprema.*

Rates $o_P(n^{-1/4})$ suffice, but known weights remove the corresponding product restrictions. The entropy condition is automatic on a fixed finite grid and holds for fixed-dimensional logistic score classes; it is not a consistency guarantee for misspecified learners. Appendix C.2 verifies these complexity claims and proves the true-score class is Donsker.

Theorem 1 (Gaussian-process limit and uniform bands)**.** *Under Assumptions 1–3,*

$$\sqrt{n}(\widehat{\Delta} - \Delta) = \mathbb{G}_n\varphi + o_P(1) \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{Y}),$$

where the remainder vanishes in supremum norm and $\mathbb{G}$ *is a tight centered Gaussian process with covariance kernel* Ω*. The fitted covariance kernel is uniformly consistent, and the Rademacher multiplier process below converges conditionally to* $\mathbb{G}$*. On a fixed nonempty set* $\mathcal{Y}_0 \subseteq \mathcal{Y}$ *with* $\inf_{y\in\mathcal{Y}_0} \Omega(y,y) > 0$*, its studentized bands have asymptotic simultaneous coverage* $1-\alpha$*, provided the limiting supremum distribution is continuous and strictly increasing at its* $(1-\alpha)$*-quantile.*

Let $\widehat{\varphi}_i(y) = \widehat{\Psi}_{1i}(y) - \widehat{\Psi}_{0i}(y) - \widehat{\Delta}(y)$. Set

$$\widehat{\Omega}(y,y') = \frac{1}{n}\sum_i \widehat{\varphi}_i(y)\widehat{\varphi}_i(y'), \qquad \widehat{\mathrm{SE}}(y) = \sqrt{\widehat{\Omega}(y,y)/n}. \tag{9}$$

Taking the arm-score difference retains their covariance. For independent Rademacher multipliers $\zeta_i^{(b)}$, independent of the data, compute

$$T^{(b)} = \sup_{y\in\mathcal{Y}_0} \left| \frac{n^{-1/2}\sum_i \zeta_i^{(b)}\widehat{\varphi}_i(y)}{\sqrt{\widehat{\Omega}(y,y)}} \right|, \qquad C_n(y) = \left[\widehat{\Delta}(y) \pm \widehat{c}_{1-\alpha}\widehat{\mathrm{SE}}(y)\right], \tag{10}$$

Here $\widehat{c}_{1-\alpha}$ is the conditional $(1-\alpha)$-quantile of $T^{(b)}$:

$$\widehat{c}_{1-\alpha} = \inf\{t : P_*(T^{(b)} \leq t) \geq 1-\alpha\},$$

where P_* holds the observed data fixed and randomizes only the multipliers. In computation, we approximate this quantile by the empirical $(1-\alpha)$-quantile of $T^{(1)},\ldots,T^{(B)}$ from B independent multiplier draws. The same critical value is used at every threshold in $C_n(y)$; pointwise intervals instead use $z_{1-\alpha/2}$. The unstudentized process limit permits $\mathcal{Y} = \mathbb{R}$; studentization excludes thresholds with vanishing variance. Unstudentized bands are given in Appendix C.2.

Ordinal outcomes and finite grids. For an outcome with a fixed finite number of ordered categories, all nontrivial cutpoints form a finite set and determine the full CDF contrast. The entropy condition in Assumption 3 is then automatic; the nuisance consistency and product-rate requirements

remain. On $\mathcal{Y}_0 = \{y_1, \ldots, y_J\}$, the limit is $N_J(0, (\Omega(y_j, y_\ell))_{j\ell})$ and the supremum is a maximum. Positive coordinate variances suffice for quantile continuity, even with singular covariance, so Theorem 1 gives simultaneous coverage of the full ordinal CDF contrast under its remaining assumptions. Algorithm 1 fits binary outcome regressions at each cutpoint, using X for μ_w and $(X, \widetilde{Y})$ for ν_w; a 1–5 judge score can inform a 0–4 human-rating scale without a separate calibration map. For continuous outcomes, evaluating a fixed grid establishes grid coverage; approximating a continuum supremum requires separate discretization control.

Theorem 2 (Semiparametric efficiency bound and oracle IPW comparison)**.** *For any fixed finite collection $y_1, \ldots, y_J \in \mathcal{Y}$, in the nonparametric observed-data model satisfying Assumptions 1–2, $\boldsymbol{\varphi}(O) = (\varphi(y_j; O))_{j=1}^J$ is the efficient influence function of the DTE vector. Its semiparametric efficiency bound, with $\Omega_{j\ell} = \Omega(y_j, y_\ell)$, is*

$$\Omega = \mathbb{E}[\boldsymbol{\varphi}(O)\boldsymbol{\varphi}(O)^\top],$$

which lower-bounds the limiting covariance of regular estimators scaled by $\sqrt{n}$. Under Assumption 3, our estimator attains this bound, giving efficient asymptotic variance Ω_{jj}/n at cutpoint y_j. For oracle IPW, with arm score $H_w = \mathbf{1}\{W = w\}RZ_y/(\pi_w\rho)$,

$$\mathrm{Var}(H_1 - H_0) = \mathrm{Var}\{\varphi(y; O)\} + \mathrm{Var}(A_1 - A_0 + B_1 - B_0), \tag{11}$$

where $A_w = \mu_w(\mathbf{1}\{W = w\}/\pi_w - 1)$ and $B_w = \nu_w\mathbf{1}\{W = w\}(R/\rho - 1)/\pi_w$.

Thus the feasible efficient estimator asymptotically weakly improves on *known-weight* unadjusted IPW. Plug-in IPW with estimated weights has additional first-order terms and need not have the oracle variance (Hirano et al., 2003). Its uncertainty must be computed for its actual fitting procedure. The comparison in (11) does not transfer by asserting equality of feasible and oracle IPW variances.

Isolating the surrogate's contribution. Part of the gain against unnormalized IPW is available even without a useful surrogate. A more informative comparator uses the same labeling weights and a covariate-only outcome regression:

$$C_w = \mu_w(y, X) + \frac{\mathbf{1}\{W = w\}R}{\pi_w\rho}\{Z_y - \mu_w(y, X)\}.$$

An orthogonal decomposition gives

$$\mathrm{Var}(C_1 - C_0) - \mathrm{Var}\{\varphi(y; O)\} = \sum_{w=0}^{1} \mathbb{E}\left[\frac{\mathbf{1}\{W = w\}}{\pi_w^2}\frac{1-\rho}{\rho}\{\nu_w(y, X, \widetilde{Y}) - \mu_w(y, X)\}^2\right]. \tag{12}$$

This is strictly positive when the surrogate changes a conditional cutpoint probability on a set with missing labels of positive probability. It vanishes for a conditionally uninformative surrogate or fully observed primary outcomes. Under MAR the comparator may still use the surrogate in its labeling model; "covariate-only" describes its outcome regression.

6 Experiments

We first assess inferential validity in simulations with known population truth. We then compare empirical precision in BiGGen Bench, HelpSteer2, wine, and ACTG 175 under their stated sampling and modeling assumptions. Additional BiGGen repetitions assess error against its fully rated prompt pool. Benchmark coverage and application SEs have different interpretations from population coverage in the simulations; implementation details are in Appendix D.

6.1 Simulation: continuous and ordinal primary outcomes

We generate $X \sim N(0, I_4)$, independent $W \sim \mathrm{Bernoulli}(1/2)$, and

$$Y(0) = 0.3X_1 - 0.2X_2 + \varepsilon, \qquad Y(1) = 0.6X_1 + 0.4X_2^2 + \varepsilon, \qquad \varepsilon \sim N(0, 0.5^2).$$

The continuous surrogate is $\widetilde{Y} = Y + 0.3(X_1 - 0.5X_2) + \eta$, with independent $\eta \sim N(0, 0.6^2)$. The arms differ in shape as well as location. In the continuous experiment, labeling depends on covariates

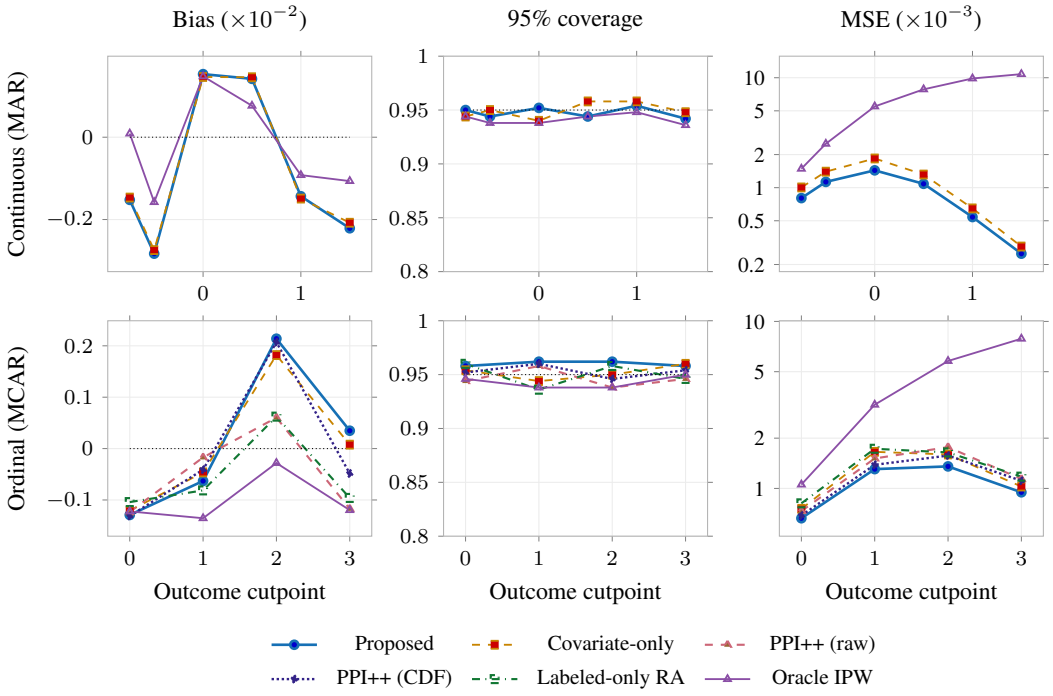


Figure 2: Known-truth simulations at $n = 2000$ with 500 repetitions. Top: a continuous primary outcome with MAR labeling. Bottom: a five-category ordinal primary outcome with MCAR labeling. The panels report bias, pointwise 95% coverage, and MSE. MSE uses a logarithmic vertical axis, with tick values in units of 10^{-3}. The ordinal panels include raw-score PPI++, learned-CDF PPI++, and labeled-only RA; Table 3 also reports the simpler baselines.

and the surrogate through a fixed bounded MAR mechanism, with expected labeled fraction 0.2. In the ordinal experiment, both the primary outcome and surrogate are discretized into five categories using boundaries $(-0.75, -0.25, 0.25, 0.75)$ and labeling is MCAR with probability 0.2. The latter mirrors a five-level judge rubric. We evaluate continuous thresholds $\{-0.75, -0.5, 0, 0.5, 1, 1.5\}$ and ordinal cutpoints $\{0, 1, 2, 3\}$. The continuous grid starts near the 10th percentile in each arm. The true DTE is evaluated by one-dimensional numerical integration, avoiding Monte Carlo error in the reference CDF.

We use five-fold cross-fitting and logistic distribution regression, including X_2^2 among covariates. Comparators are known-weight IPW, the covariate-only outcome-adjustment ablation, labeled-only empirical CDFs, and surrogate-only CDFs. The ordinal MCAR experiment adds untuned thresholdwise PPI, PPI++ (Angelopoulos et al., 2023b) with raw-score and learned-CDF predictions, and labeled-only regression-adjusted DTE estimation (Byambadalai et al., 2024). Labeled-only RA uses labeled records for both fitting and averaging; PPI++ uses disjoint labeled and unlabeled arm samples with training-only tuning. These methods are compared under randomized MCAR sampling, where they identify the same target; they are not applied unchanged to MAR or observational data. Appendix D.2 specifies the implementations.

Results. Figure 2 shows that, in the ordinal experiment, the proposed method reduces MSE across cutpoints by 30–45% relative to labeled-only, 34–48% relative to untuned PPI, and 8–21% relative to covariate-only adjustment. Its pointwise coverage ranges from 0.958 to 0.962. Against the stronger baselines in Table 3, the reductions are 9–23% relative to raw-score PPI++, 4–15% relative to learned-CDF PPI++, and 18–24% relative to labeled-only RA. For continuous primary outcomes under MAR, MSE reductions across cutpoints are 46–98% relative to oracle IPW and 14–22% relative to covariate-only adjustment. Continuous pointwise coverage ranges from 0.942 to 0.954.

For each repetition, we approximate $\widehat{c}_{0.95}$ with 500 Rademacher realizations of $T^{(b)}$ in (10). With $n = 2000$, 5000 repetitions per outcome, and both π and ρ estimated, nominal 95% bands attain simultaneous coverage over all evaluated cutpoints of 0.951 for ordinal outcomes and 0.938 for continuous outcomes. The mean ratios of band to pointwise-interval widths are 1.26 and 1.32, respectively.

6.2 Applications

BiGGen Bench. We compare Mistral-7B-Instruct-v0.2 with Mixtral-8x7B-Instruct-v0.1 on 695 shared prompts from BiGGen Bench (Kim et al., 2025). Each prompt is assigned one model's released response with probability $1/2$; exactly 200 responses are sampled uniformly for human rating. The primary outcome is the human 1–5 task-specific rubric rating, and the surrogate is the cached GPT-4 Turbo score. We adjust for user- and system-prompt lengths, character diversity, and capability indicators. The DTE measures how switching to Mixtral changes the probability of a rating at or below each cutpoint, without assuming equal distances between rating categories. The application uses the first archived assignment and annotation sample, with every nontrivial cutpoint $y \in \{1, 2, 3, 4\}$.

HelpSteer2. We use 20,324 prompt–response pairs with human helpfulness ratings from HelpSteer2 (Wang et al., 2024). The contrast compares released complexity ratings at least 2 against simpler responses, adjusting for prompt length and character diversity. Human ratings and cached judge scores use a 0–4 scale, with 399 human ratings retained under a 2% MCAR mask. The judge applies the helpfulness rubric without being fitted to the released human ratings. Cutpoint contrasts describe how the frequency of low helpfulness ratings differs by response complexity.

Wine reviews. We compare red and white varieties among 10,000 US wine reviews, adjusting for log price, its square, and state (Ji et al., 2025). The primary outcome is the professional points rating; the surrogate is an LLM score of the tasting description on a 1–5 scale. This application illustrates unequal primary-outcome and surrogate scales: the outcome regressions predict primary-outcome cutpoint probabilities directly from the auxiliary rating. A 20% MCAR mask retains 2,049 professional ratings.

ACTG 175. We compare zidovudine–didanosine with zidovudine monotherapy (Hammer et al., 1996) in 1,054 patients, using week-20 CD4 as a surrogate for week-96 CD4, observed for 654 patients. Here missing primary outcomes arise in follow-up rather than from an imposed annotation mask. We adjust for fourteen baseline covariates, including baseline CD4, age, and weight, and fit a labeling score using these covariates, treatment, and week-20 CD4.

Uncertainty and results. Table 1 reports each cutpoint separately. Every SE is the standard deviation of 200 record-bootstrap estimates, refitting nuisances and repeating cross-fitting with common resamples across methods. Assignment and observed labels are fixed before bootstrapping. These are precision diagnostics, without corrections for additional prompt, reviewer, finite-population, or trial-design dependence. Across the 18 dataset–cutpoint combinations in Table 1, proposed bootstrap SEs are lower than IPW in all 18 (reductions of 5–64%) and lower than covariate-only adjustment in 16 (reductions up to 28%). SE gains depend on the cutpoint and comparator: proposed SE slightly exceeds covariate-only SE for BiGGen at $y = 3$ and ACTG at $y = 100$. In BiGGen's repeated-assignment benchmark (Appendix Table 4), the proposed estimator reduces MSE relative to covariate-only adjustment and performs comparably to thresholdwise PPI++; neither method dominates at every cutpoint. Our theoretical results additionally cover selective MAR labeling and uniform inference across outcome thresholds, settings beyond this MCAR, finite-grid benchmark.

7 Discussion and Limitations

A model update can affect different parts of the human-rating distribution in different ways. Distributional treatment effects make these changes visible without assigning numerical distances between rating categories. For ordinal outcomes, simultaneous inference across all nontrivial cutpoints describes uncertainty in the entire distribution contrast. This supports evaluation decisions

			Bootstrap standard error			SE ratio		MSE ratio
Dataset	y	$\widehat{\Delta}$	Proposed	Covariate	IPW	/ IPW	/ Cov.	/ Cov.
BiGGen	1	$+0.025$	0.042	0.045	0.045	0.95	0.95	0.80
	2	$+0.012$	0.055	0.056	0.063	0.87	0.99	0.80
	3	-0.058	0.066	0.064	0.080	0.82	1.03	0.78
	4	-0.052	0.068	0.069	0.112	0.61	0.99	0.90
HelpSteer2	0	-0.054	0.027	0.030	0.032	0.85	0.91	—
	1	-0.096	0.036	0.041	0.046	0.80	0.89	—
	2	-0.193	0.047	0.049	0.055	0.85	0.96	—
	3	-0.037	0.050	0.052	0.073	0.69	0.97	—
Wine	85	$+0.073$	0.012	0.017	0.021	0.58	0.72	—
	87	$+0.106$	0.018	0.022	0.030	0.59	0.81	—
	89	$+0.098$	0.019	0.022	0.037	0.52	0.87	—
	91	$+0.068$	0.020	0.021	0.044	0.46	0.96	—
	93	$+0.062$	0.018	0.018	0.050	0.36	1.00	—
ACTG 175	100	-0.107	0.027	0.027	0.032	0.86	1.00	—
	200	-0.173	0.032	0.033	0.042	0.75	0.97	—
	300	-0.153	0.033	0.035	0.052	0.64	0.95	—
	400	-0.119	0.032	0.032	0.058	0.55	0.99	—
	500	-0.056	0.024	0.025	0.061	0.40	0.98	—

Table 1: Application estimates and bootstrap SEs at each primary-outcome cutpoint (200 refitted replicates). BiGGen uses 200 human ratings. SE ratios divide proposed SE by the indicated comparator's SE. BiGGen MSE ratios compare proposed to covariate-only adjustment over 500 repeated assignments against the fully rated prompt-pool reference; dashes indicate that no truth-based MSE is available for the other applications. IPW uses known probabilities in BiGGen and fitted treatment or labeling probabilities where needed in the other applications.

that consider several quality thresholds together, rather than relying on an average score or selecting a favorable cutpoint.

The experiments show that surrogate information can improve precision beyond covariate adjustment, although the benefit varies across thresholds. In BiGGen's repeated-assignment benchmark, the proposed estimator reduces MSE relative to covariate-only adjustment and performs comparably to thresholdwise PPI++. These results support using automated ratings alongside a limited human-rating sample, while showing that the additional benefit depends on the information those ratings provide. The theoretical framework also accommodates selective labeling under MAR, extending beyond the randomized labeling used in BiGGen.

Finite-sample calibration remains an important follow-up. Ordinal simultaneous coverage is close to nominal in our simulation, whereas continuous-outcome bands slightly under-cover. Replacing fitted conditional CDFs with their exact values improves coverage, suggesting that outcome estimation deserves particular attention in further work. The asymptotic guarantees therefore do not settle how much labeled data is sufficient for reliable bands in a given application.

AI USE STATEMENT

Generative AI was used to review and revise the manuscript, discuss the statistical formulation and proofs, identify relevant literature, implement and check experimental procedures, generate figures and tables from computed results. Cached LLM-generated surrogate ratings are used in BiGGen, HelpSteer2, and wine; their construction is described in Appendix D. AI assistance does not replace author responsibility for the final mathematical claims, code, empirical results, or text.

ETHICS STATEMENT

This work uses simulations and secondary analyses of existing datasets. Human ratings serve as reference outcomes for specific evaluation rubrics and should not be interpreted as universal measures of quality. LLM-generated surrogate scores may reflect model and dataset biases; improved statisti-

cal efficiency does not itself establish fairness or causal validity. The empirical applications illustrate uncertainty estimation, and the ACTG 175 analysis is not intended to support clinical recommendations. Dataset sources and preprocessing are documented in Appendix D and the supplementary materials.

Reproducibility statement

Algorithm 1 gives the estimation workflow; Appendices A–D provide implementation details, proofs, and experimental specification. The accompanying implementation saves replication-level estimates, standard errors, coverage indicators, bootstrap draws of empirical estimates, and the generated tables and figures. Judge scores for the experiments are also included in the reproduction package.

References

Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292, 2002. doi: 10.1198/016214502753479419.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Susan Athey and Guido W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006. doi: 10.1111/j.1468-0262.2006.00668.x.

Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, 1993.

Marianne P. Bitler, Jonah B. Gelbach, and Hilary W. Hoynes. What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012, 2006. doi: 10.1257/aer.96.4.988.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5276–5290. PMLR, 2025. URL `https://proceedings.mlr.press/v267/boyeau25a.html`.

Undral Byambadalai, Tatsushi Oka, and Shota Yasui. Estimating distributional treatment effects in randomized experiments: Machine learning for variance reduction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5082–5113. PMLR, 2024.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *Annals of Statistics*, 46(4):1541–1572, 2018.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Victor Chernozhukov, Iván Fernández-Val, Blaise Melly, and Kaspar Wuthrich. Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, 115(529):123–137, 2020.

Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2(2):267–277, 1974.

Naoki Egami, Musashi Hinck, Brandon M Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68589–68601, 2023. doi: 10.52202/075280-3000. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d862f7f5445255090de13b825b880d59-Abstract-Conference.html.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75 (1):259–276, 2007.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. *arXiv preprint arXiv:2406.04291*, 2024.

Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335 (15):1081–1090, 1996.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Guido W. Imbens and Donald B. Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997. doi: 10.2307/2971731.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *arXiv preprint arXiv:2501.09731*, 2025.

Liang Jiang, Peter C. B. Phillips, Yubo Tao, and Yichong Zhang. Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics*, 234(2):758–776, 2023.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society: Series B*, 87(2):480–509, 2025.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25 (16):1–59, 2024.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.08491.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. The BiGGen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5877–5919. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.303. URL https://aclanthology.org/2025.naacl-long.303/.

Dan M. Kluger and Stephen Bates. M-estimation under two-phase multiwave sampling with applications to prediction-powered inference. *arXiv preprint arXiv:2602.16933*, 2026.

Erich L Lehmann and Howard J D'Abrera. *Nonparametrics: statistical methods based on ranks*. Holden-Day, 1975.

Ziyang Liu, Niwen Zhou, Peng Wu, and Xu Guo. Double/debiased machine learning of quantile treatment effects on long-term outcomes in clinical trials. *arXiv preprint arXiv:2605.14275*, 2026.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Aad W van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Aad W. van der Vaart and Jon A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. doi: 10.1214/11-EJS605.

Pengyun Wang. Efficient transported distributional and quantile treatment effects with surrogate-assisted missing primary outcomes. *arXiv preprint arXiv:2605.02154*, 2026.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501, 2024. doi: 10.52202/079017-0047. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/02fd91a387a6a5a5751e81b58a75af90-Abstract-Datasets_and_Benchmarks_Track.html.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47(5):2538–2566, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

Notation used in the main text and proofs is collected in Appendix B.

A Implementation Details

Algorithm 1 uses inverse-label weights $1/\widehat{\rho}(X, w, \widetilde{Y})$ when fitting $\widehat{\mu}_w$ on labeled training units under MAR. Bound fitted inverse weights as in Assumption 3. If estimating μ_w by integrating $\widehat{\nu}_w$ over all arm-w training units instead, use an additional training split or other conditions sufficient to establish the stated nuisance rates.

Train all propensity, labeling, outcome, and feature transformations outside the evaluation fold. Observed primary outcomes in that fold cannot be used for surrogate calibration or preprocessing in its training pipeline. In the experiments the surrogate scores are fixed existing measurements; no calibration to the primary-outcome scale is learned.

An all-zero or all-one threshold target in a training arm is represented by its constant fitted probability. For ordinal outcomes, the all-inclusive final category is omitted from studentization because its DTE is identically zero.

B SUMMARY OF MATHEMATICAL NOTATION

Arguments are suppressed when unambiguous. $\mathcal{Y}$ is the theoretical threshold index set; experiments evaluate fixed finite grids within it.

Symbol	Definition or interpretation
Observations and targets	
O_i, n	Observed record $(X_i, W_i, \widetilde{Y}_i, R_i, R_iY_i)$; number of independent records.
X, W, w	Pre-treatment covariates; binary arm assignment; an arm in $\{0, 1\}$.
Y, $Y(w)$, $\widetilde{Y}$	Primary outcome; potential primary outcome under arm w; surrogate observed for every unit.
R, I_w, Z_y	Primary-outcome observation indicator; $\mathbf{1}\{W = w\}$; $\mathbf{1}\{Y \le y\}$. Empirical uses of the latent Z_y are multiplied by R.
$\mathcal{Y}$, $\mathcal{Y}_0$, J	Fixed threshold index set; subset used for studentized bands; number of thresholds in a finite-grid specialization.
$F_w(y)$, $\Delta(y)$	$P\{Y(w) \le y\}$ and $F_1(y) - F_0(y)$. The standardized interpretation is $F_w(y) = \mathbb{E}_X \mu_w(y, X)$; the causal interpretation requires exchangeability.
Nuisance functions and estimation	
$\pi_w(x)$, $\rho(x, w, s)$	$P(W = w \mid X = x)$ and $P(R = 1 \mid X = x, W = w, \widetilde{Y} = s)$.
ϵ, $\rho_{\min}$	Positive population lower bounds for π_w and ρ.
$\mu_w(y, x)$	$P(Y \le y \mid X = x, W = w)$.
$\nu_w(y, x, s)$	$P(Y \le y \mid X = x, W = w, \widetilde{Y} = s)$; under MAR, also the conditional CDF of the primary outcome among labeled records.
η, $\bar{\eta}$, $\widehat{\eta}^{-k}$	True nuisance collection (μ, ν, π, ρ); fixed candidate collection; collection fitted outside fold k.
K, $\mathcal{I}_k$, n_k, $k(i)$	Number of folds; evaluation indices in fold k; $\lvert\mathcal{I}_k\rvert$; fold containing observation i.
Ψ_w, $\widehat{\Psi}_{wi}$	Augmented arm score in (4); its evaluation at O_i using $\widehat{\eta}^{-k(i)}$.
$\widehat{F}_w$, $\widehat{\Delta}$	Average fitted arm score and difference between the two arm averages, as in (5).
Influence functions and inference	
$\varphi_w(y; O)$, $\varphi(y; O)$	$\Psi_w(y; O, \eta) - F_w(y)$ and $\varphi_1(y; O) - \varphi_0(y; O)$.
$\mathbb{G}$, $\Omega(y, y')$	Tight Gaussian limit; covariance kernel $\mathbb{E}[\varphi(y; O)\varphi(y'; O)]$. On a finite grid, $\boldsymbol{\varphi} = (\varphi(y_j; O))_j$ and $\Omega_{j\ell} = \Omega(y_j, y_\ell)$.
$\widehat{\varphi}_i(y)$, $\widehat{\Omega}(y, y')$	Centered fitted contrast score; its empirical covariance kernel in (9).
$\sigma(y)$, $V(y)$	$\sqrt{\Omega(y, y)}$; studentized limit $\mathbb{G}(y)/\sigma(y)$ on $\mathcal{Y}_0$. Hats denote plug-in versions.
$\zeta_i^{(b)}$, $T^{(b)}$, $\widehat{c}_{1-\alpha}$	Rademacher multipliers; supremum statistic in (10); its conditional $(1-\alpha)$-quantile, approximated by the empirical quantile across multiplier draws.

Proof conventions. $Pf = \int f\, dP$, $P_n f = n^{-1} \sum_i f(O_i)$, and $P_{n,k} f = n_k^{-1} \sum_{i \in \mathcal{I}_k} f(O_i)$. The training sigma-field is $\mathcal{A}_k$; $\|f\|_2 = (P[f^2])^{1/2}$ holds its training sample fixed and is generally random for fitted f. In Appendix C.2, $\delta_f = \widehat{f} - f$, $\xi_w = \nu_w - \mu_w$, and $\eta_w = Z_y - \nu_w$; this subscripted residual η_w is distinct from the nuisance collection η. $\mathbb{G}_n = \sqrt{n}(P_n - P)$ and $\mathbb{G}_{n,k} = \sqrt{n_k}(P_{n,k} - P)$ are empirical processes. $N(u, \mathcal{F}, L^2(Q))$ is the covering number at radius u. $P_*, \mathbb{E}_*$ denote multiplier probability and expectation conditional on the data. In Appendix C.3, $L_0^2(P)$ is the space of square-integrable centered functions, $\mathcal{T}$ is the observed-data tangent space, and $\mathcal{T}_X, \mathcal{T}_W, \mathcal{T}_S, \mathcal{T}_R, \mathcal{T}_Y$ are its orthogonal factor components; S in these subscripts denotes the surrogate $\widetilde{Y}$.

C PROOFS

Throughout, $I_w = \mathbf{1}\{W = w\}$, $Z_y = \mathbf{1}\{Y \le y\}$, and all nuisance arguments are implicit when unambiguous. Statements involving latent Z_y are expectations; its empirical contribution is always multiplied by R.

C.1 IDENTIFICATION AND THE SCORE BIAS IDENTITY

MAR gives $\mathbb{E}[Z_y \mid X, W = w, \widetilde{Y}, R = 1] = \nu_w$. The tower property gives $\mathbb{E}[\nu_w \mid X, W = w] = \mu_w$, and consistency and exchangeability give $F_w = \mathbb{E}_X \mu_w$. Thus each arm CDF is identified from the observed-data law as

$$F_w(y) = \mathbb{E}_X\Big[\mathbb{E}\Big\{P(Y \le y \mid X, W = w, \widetilde{Y}, R = 1) \mid X, W = w\Big\}\Big].$$

The inner expectation averages over the surrogate distribution among all records in arm w conditional on X, not only labeled records. Treatment and labeling positivity ensure that the required conditional distributions are identified on the target support.

For fixed candidate fits, conditional expectation of the final score correction is

$$\mathbb{E}\left[\frac{I_w R}{\bar{\pi}_w \bar{\rho}}(Z_y - \bar{\nu}_w)\middle| X, W, \widetilde{Y}\right] = \frac{I_w \rho}{\bar{\pi}_w \bar{\rho}}(\nu_w - \bar{\nu}_w).$$

Combining with the second term yields

$$\mathbb{E}[\Psi_w(\bar{\eta})] = \mathbb{E}\bar{\mu}_w + \mathbb{E}\left[\frac{I_w}{\bar{\pi}_w}(\nu_w - \bar{\mu}_w)\right] + \mathbb{E}\left[\frac{I_w}{\bar{\pi}_w}(\bar{\nu}_w - \nu_w)\frac{\bar{\rho} - \rho}{\bar{\rho}}\right].$$

Condition the middle term on X and subtract $F_w = \mathbb{E}\mu_w$. The first two terms become $\mathbb{E}[(\bar{\mu}_w - \mu_w)(\bar{\pi}_w - \pi_w)/\bar{\pi}_w]$, proving (6). This also proves the asserted robustness combinations without presuming consistency of all four candidate functions. At the true nuisance functions, the bias vanishes, giving the score representation of the identified targets:

$$F_w(y) = \mathbb{E}[\Psi_w(y; O, \eta)], \qquad \Delta(y) = \mathbb{E}[\Psi_1(y; O, \eta) - \Psi_0(y; O, \eta)].$$

C.2 FUNCTIONAL LIMIT, COVARIANCE CONSISTENCY, AND UNIFORM BANDS

We prove Theorem 1. The partition $\{\mathcal{I}_k\}_{k=1}^K$ is fixed or drawn independently of the observations. Conditional on the partition, let $n_k = |\mathcal{I}_k|$, with $n_k/n \to \lambda_k \in (0, 1)$, and define

$$\mathcal{A}_k = \sigma(O_i : i \notin \mathcal{I}_k), \qquad P_{n,k} f = \frac{1}{n_k}\sum_{i \in \mathcal{I}_k} f(O_i), \qquad P_n f = \frac{1}{n}\sum_{i=1}^n f(O_i).$$

Any independent algorithmic randomization is included in $\mathcal{A}_k$. For an $\mathcal{A}_k$-measurable random function f, $Pf = \int f(o)\, dP(o)$ holds the training sample fixed; hence $\mathbb{E}[P_{n,k} f \mid \mathcal{A}_k] = Pf$. Each argument below conditions on one fold's training sample, not on all fitted nuisance functions simultaneously.

Step 0: entropy, measurability, and a local maximal inequality. All suprema below are measurable by Assumption 3; equivalently, one may work with separable versions and outer probabilities. Choose conditional distributions as CDF kernels, so $y \mapsto \mu_w(y, x)$ and $y \mapsto \nu_w(y, x, s)$ are nondecreasing and right-continuous. Each class indexed by y is bounded by one and has nested subgraphs. Such a class is VC-subgraph, as is $\{\mathbf{1}\{Y \le y\} : y \in \mathcal{Y}\}$; restricting the index set preserves this property. The true influence function can be written

$$\varphi_w(y) = \left(1 - \frac{I_w}{\pi_w}\right)\mu_w(y) + \frac{I_w}{\pi_w}\left(1 - \frac{R}{\rho}\right)\nu_w(y) + \frac{I_w R}{\pi_w \rho} Z_y - F_w(y).$$

The coefficients are fixed bounded functions of the observation, independent of y. For a bounded weight a, $\|a(f - g)\|_{Q,2} \le \|a\|_\infty \|f - g\|_{Q,2}$; finite sums admit covers obtained by multiplying the component covering numbers. Consequently the true arm-score and contrast-influence classes have constant envelopes and polynomial uniform covering numbers. The class of constants $\{F_w(y) : y \in$

$\mathcal{Y}\}$ has the same property. Their entropy integrals are finite, giving the P-Donsker property by the uniform-entropy CLT (van der Vaart & Wellner, 1996). This uses CDF ordering, not continuity of the outcome distribution or a density assumption. The resulting Gaussian process has sample paths uniformly continuous in the $L^2(P)$ semimetric; Euclidean continuity across outcome atoms is not asserted.

Condition on one training field $\mathcal{A}_k$. On the event in (8), the score-error class $\mathcal{D}_{kw} = \{d_{kw}(y) : y \in \mathcal{Y}\}$, its squared class, and finite sums and products of bounded component classes inherit polynomial entropy with deterministic constants. Here one can cover the difference class by the Cartesian product of covers for the fitted and true score classes; no smoothness of the map $y \mapsto \widehat{\eta}(y)$ is assumed. Its bounded envelope depends only on the overlap and entropy constants. For any such class $\mathcal{D}$ on m independent evaluation observations, write $s = \sup_{d\in\mathcal{D}} \|d\|_2$. The local maximal inequality of van der Vaart & Wellner (2011, Thm. 2.1), after rescaling the envelope, implies for $0 < a < 1$ and $s \le a$,

$$\mathbb{E}\|\mathbb{G}_m\|_{\mathcal{D}} \le C\left\{a\sqrt{\log(C/a)} + \frac{\log(C/a)}{\sqrt{m}}\right\}, \qquad \mathbb{G}_m = \sqrt{m}(P_m - P), \tag{13}$$

with C independent of the realized training data on that event. Indeed the uniform entropy integral is bounded by $Ca\sqrt{\log(C/a)}$, and substitution into that theorem gives (13). Enlarging C absorbs the constant envelope. If $s = o_P(1)$, conditional Markov's inequality first with fixed a, then $m \to \infty$, then $a \downarrow 0$, proves $\|\mathbb{G}_m\|_{\mathcal{D}} = o_P(1)$. The exceptional entropy event has probability tending to zero. Thus no deterministic rate for s beyond consistency is needed for this centered remainder.

For a fixed J-point grid, any fitted score class has at most J elements, so (8) holds without restricting its learner. For the implemented fixed-dimensional logistic regressions, conditional on training, standardization is an affine change of the fixed features, and each CDF fit belongs to the class of logistic transforms of linear predictors, including the two constant fits. This is a bounded VC-subgraph class. The fitted inverse weights are bounded and do not depend on y, so the preceding covering argument gives (8) with constants depending on feature dimension and overlap, not on the training sample. This verifies complexity only: correct approximation and the uniform rates in (7) remain substantive requirements. For more flexible learners one must verify (8) or supply an alternative vanishing local-entropy bound; cross-fitting alone is insufficient.

Step 1: exact score expansion and conditional bounds. Fix k, w and suppress $(-k)$ on fitted nuisances. For each $f \in \{\mu_w, \nu_w, \pi_w, \rho\}$, write $\delta_f = \widehat{f} - f$. Also write $\xi_w = \nu_w - \mu_w$ and $\eta_w = Z_y - \nu_w$. The difference $d_{kw}(y, O) = \Psi_w(y; O, \widehat{\eta}^{-k}) - \Psi_w(y; O, \eta)$ decomposes exactly as $\sum_{a=1}^{4} T_{a,kw}(y, O)$, where

$$\begin{aligned} T_{1,kw} &= \delta_{\mu_w}\left(1 - \frac{I_w}{\widehat{\pi}_w}\right), & T_{2,kw} &= \frac{I_w\delta_{\nu_w}}{\widehat{\pi}_w}\left(1 - \frac{R}{\widehat{\rho}}\right), \\ T_{3,kw} &= I_w\xi_w\left(\frac{1}{\widehat{\pi}_w} - \frac{1}{\pi_w}\right), & T_{4,kw} &= I_wR\eta_w\left(\frac{1}{\widehat{\pi}_w\widehat{\rho}} - \frac{1}{\pi_w\rho}\right). \end{aligned} \tag{14}$$

The population residuals satisfy

$$\mathbb{E}[\xi_w \mid X, W = w] = 0, \qquad \mathbb{E}[\eta_w \mid X, W = w, \widetilde{Y}, R = 1] = 0. \tag{15}$$

Consequently, conditioning successively on X, $(X, W, \widetilde{Y})$, and $(X, W, \widetilde{Y}, R)$ gives

$$PT_{1,kw} = P\left[\frac{\delta_{\mu_w}\delta_{\pi_w}}{\widehat{\pi}_w}\right], \qquad PT_{2,kw} = P\left[\frac{I_w\delta_{\nu_w}\delta_\rho}{\widehat{\pi}_w\widehat{\rho}}\right], \qquad PT_{3,kw} = PT_{4,kw} = 0. \tag{16}$$

These are conditional expectations with respect to a fresh evaluation record, with training fits held fixed. In particular, the exact bias identity and Cauchy–Schwarz imply

$$|Pd_{kw}(y)| \le \frac{2}{\epsilon}\|\delta_{\mu_w}\|_2\|\delta_{\pi_w}\|_2 + \frac{4}{\epsilon\rho_{\min}}\|\delta_{\nu_w}\|_2\|\delta_\rho\|_2. \tag{17}$$

To control second moments, the overlap bounds and probability-valued estimates give

$$\begin{aligned} \left|\frac{1}{\widehat{\pi}_w} - \frac{1}{\pi_w}\right| &\le \frac{2}{\epsilon^2}|\delta_{\pi_w}|, \\ \left|\frac{1}{\widehat{\pi}_w\widehat{\rho}} - \frac{1}{\pi_w\rho}\right| &\le \frac{4}{\epsilon^2\rho_{\min}^2}(|\delta_{\pi_w}| + |\delta_\rho|). \end{aligned}$$

Using $|\xi_w|, |\eta_w| \le 1$, we obtain the four explicit bounds

$$\begin{aligned}
\|T_{1,kw}\|_2 &\le (1 + 2/\epsilon)\|\delta_{\mu_w}\|_2, \\
\|T_{2,kw}\|_2 &\le (2/\epsilon)(1 + 2/\rho_{\min})\|\delta_{\nu_w}\|_2, \\
\|T_{3,kw}\|_2 &\le (2/\epsilon^2)\|\delta_{\pi_w}\|_2, \\
\|T_{4,kw}\|_2 &\le \{4/(\epsilon^2\rho_{\min}^2)\}(\|\delta_{\pi_w}\|_2 + \|\delta_\rho\|_2).
\end{aligned}$$

Thus, for a constant C depending only on the overlap bounds,

$$\sup_{y\in\mathcal{Y}} \|d_{kw}(y)\|_2 \le C \sum_{f\in\{\mu_w,\nu_w,\pi_w,\rho\}} \sup_{y\in\mathcal{Y}} \|\widehat{f} - f\|_2 = o_P(1). \tag{18}$$

Equation (17) and Assumption 3 separately yield $\sup_y |Pd_{kw}(y)| = o_P(n^{-1/2})$. Consistency controls the centered empirical error; the product rates control its conditional mean.

Step 2: uniform remainder and functional central limit theorem. Adding and subtracting the population score gives the exact expansion

$$\begin{aligned}
\sqrt{n}\{\widehat{F}_w(y) - F_w(y)\} &= \mathbb{G}_n\varphi_w(y) + R_{1,nw}(y) + R_{2,nw}(y), \\
R_{1,nw}(y) &= \sum_{k=1}^{K} \sqrt{n_k/n}\,\mathbb{G}_{n,k}d_{kw}(y), \\
R_{2,nw}(y) &= \sqrt{n}\sum_{k=1}^{K}(n_k/n)Pd_{kw}(y), \qquad \mathbb{G}_{n,k} = \sqrt{n_k}(P_{n,k} - P).
\end{aligned} \tag{19}$$

By (18), $s_{kw} := \sup_y \|d_{kw}(y)\|_2 = o_P(1)$. To make the conditional-to-unconditional step explicit, let E_{kw} be the training-measurable event on which the entropy bound holds. For fixed $t > 0$ and $0 < a < 1$, (13) gives

$$\begin{aligned}
P\{\|\mathbb{G}_{n,k}\|_{\mathcal{D}_{kw}} > t\} \le{}& P(E_{kw}^c) + P(s_{kw} > a) \\
&+ \frac{C}{t}\left\{a\sqrt{\log(C/a)} + n_k^{-1/2}\log(C/a)\right\}.
\end{aligned}$$

First let $n \to \infty$ and then $a \downarrow 0$. Since K is fixed, $\sup_y |R_{1,nw}(y)| = o_P(1)$. Equation (17) and the uniform product rates similarly give

$$\sup_y |R_{2,nw}(y)| \le \sqrt{n}\sum_k (n_k/n)\sup_y |Pd_{kw}(y)| = o_P(1).$$

No independence between foldwise remainders is used. Subtracting arms yields

$$\left\|\sqrt{n}(\widehat{\Delta} - \Delta) - \mathbb{G}_n\varphi\right\|_{\mathcal{Y}} = o_P(1), \qquad \|h\|_{\mathcal{Y}} := \sup_{y\in\mathcal{Y}} |h(y)|. \tag{20}$$

Step 0 shows that $\mathcal{F}_\varphi = \{\varphi(y;\cdot) : y \in \mathcal{Y}\}$ is P-Donsker. Thus $\mathbb{G}_n\varphi \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{Y})$, with covariance $P[\varphi(y)\varphi(y')]$ because $P\varphi(y) = 0$. Functional Slutsky applied to (20) proves the process limit, including asymptotic tightness. It also gives $\|\widehat{\Delta} - \Delta\|_{\mathcal{Y}} = O_P(n^{-1/2})$. All these statements hold at each fixed law P satisfying the assumptions; uniformity over thresholds is distinct from uniformity over a class of data-generating laws.

Step 3: uniform covariance consistency and centering. Set $q_i(y) = \widehat{\Psi}_{1i}(y) - \widehat{\Psi}_{0i}(y)$, $b_n(y) = \widehat{\Delta}(y) - \Delta(y)$, and $e_i(y) = q_i(y) - \{\varphi(y; O_i) + \Delta(y)\}$. On fold k, $e_i(y) = e_k(y, O_i) := d_{k1}(y, O_i) - d_{k0}(y, O_i)$. The conditional squared-error class has a bounded envelope and polynomial entropy. Its uniform empirical-process expectation is bounded by a deterministic constant, using (13) with a fixed radius after envelope rescaling. Consequently, conditional Markov and the vanishing probability of the exceptional entropy event give

$$\sup_y |(P_{n,k} - P)e_k(y)^2| = o_P(1).$$

Since $\sup_y Pe_k(y)^2 = o_P(1)$ by Step 1, and there are finitely many folds,

$$r_n^2 := \sup_{y\in\mathcal{Y}} \frac{1}{n}\sum_i e_i(y)^2 = o_P(1). \tag{21}$$

This uniform law replaces the finite union bound; pointwise conditional Markov bounds alone would not establish it.

Let $\widetilde{\varphi}_i(y) = q_i(y) - \Delta(y) = \varphi(y; O_i) + e_i(y)$ and let C_φ bound the true influence functions. Expansion and Cauchy–Schwarz yield

$$\sup_{y,y'\in\mathcal{Y}} \left| \frac{1}{n}\sum_i \widetilde{\varphi}_i(y)\widetilde{\varphi}_i(y') - \frac{1}{n}\sum_i \varphi(y; O_i)\varphi(y'; O_i) \right| \le 2C_\varphi r_n + r_n^2 = o_P(1). \tag{22}$$

Empirical centering has the exact effect

$$\widehat{\Omega}(y,y') = \frac{1}{n}\sum_i \widetilde{\varphi}_i(y)\widetilde{\varphi}_i(y') - b_n(y)b_n(y'), \qquad \sup_{y,y'} |b_n(y)b_n(y')| = O_P(n^{-1}). \tag{23}$$

The bounded true-score product class is Glivenko–Cantelli: covers for each factor give polynomial entropy for their products. Its uniform LLN, (22), and (23) prove

$$\sup_{y,y'\in\mathcal{Y}} |\widehat{\Omega}(y,y') - \Omega(y,y')| = o_P(1). \tag{24}$$

In particular, with $\sigma(y) = \sqrt{\Omega(y,y)}$ and $\widehat{\sigma}(y) = \sqrt{\widehat{\Omega}(y,y)}$, $\|\widehat{\sigma} - \sigma\|_{\mathcal{Y}} = o_P(1)$; on $\mathcal{Y}_0$ with $\inf_{\mathcal{Y}_0} \sigma > 0$, also $\|1/\widehat{\sigma} - 1/\sigma\|_{\mathcal{Y}_0} = o_P(1)$.

Step 4: conditional multiplier process convergence. Let ζ_i be independent Rademacher variables, independent of the observations, and define

$$G_n^*(y) = n^{-1/2}\sum_i \zeta_i \widehat{\varphi}_i(y), \qquad G_{n,0}^*(y) = n^{-1/2}\sum_i \zeta_i \varphi(y; O_i).$$

For the true P-Donsker class, the conditional multiplier CLT (van der Vaart & Wellner, 1996, Thm. 2.9.6) gives $G_{n,0}^* \rightsquigarrow_* \mathbb{G}$ in $\ell^\infty(\mathcal{Y})$, in probability. Applying the theorem to empirically centered true scores gives the same result: their difference is $(P_n\varphi)(n^{-1/2}\sum_i \zeta_i)$, whose supremum norm is $O_P(n^{-1/2})O_{P_*}(1)$.

We verify that replacing true scores by cross-fitted scores is conditionally negligible. On fold k, conditional on $\mathcal{A}_k$, the pairs (O_i, ζ_i) are i.i.d. under $P \otimes P_\zeta$. The class $\{(o,z) \mapsto z e_k(y,o) : y \in \mathcal{Y}\}$ is centered, has the same L^2 radii as the error class, and inherits its entropy bound since $|z| = 1$. Therefore (13) and the fixed-radius argument in Step 2 give

$$H_{n,k}^*(y) := n_k^{-1/2}\sum_{i\in\mathcal{I}_k} \zeta_i e_k(y, O_i), \qquad \|H_{n,k}^*\|_{\mathcal{Y}} = o_P(1)$$

under the joint law of data and multipliers. This also gives conditional negligibility in probability: for any $t, \delta > 0$,

$$P\{P_*(\|H_{n,k}^*\|_{\mathcal{Y}} > t) > \delta\} \le \delta^{-1} P(\|H_{n,k}^*\|_{\mathcal{Y}} > t) \longrightarrow 0,$$

Finally,

$$G_n^* - G_{n,0}^* = \sum_k \sqrt{n_k/n}\, H_{n,k}^* - b_n\, n^{-1/2}\sum_i \zeta_i.$$

The last term is conditionally negligible because $\|b_n\|_{\mathcal{Y}} = O_P(n^{-1/2})$ and $\mathbb{E}_*(n^{-1/2}\sum_i \zeta_i)^2 = 1$. Hence

$$G_n^* \rightsquigarrow_* \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{Y}), \text{ in probability.} \tag{25}$$

Here conditional convergence means convergence, in probability, of conditional expectations against bounded Lipschitz functionals. We never condition simultaneously on all training fields to claim independence of fitted scores.

Step 5: studentization, quantiles, and coverage. On $\mathcal{Y}_0$, Steps 2–4 and continuity of multiplication by the bounded function $1/\sigma$ imply

$$\frac{\sqrt{n}(\widehat{\Delta}-\Delta)}{\widehat{\sigma}} \rightsquigarrow V := \frac{\mathbb{G}}{\sigma}, \qquad \frac{G_n^*}{\widehat{\sigma}} \rightsquigarrow_* V \quad \text{in } \ell^\infty(\mathcal{Y}_0).$$

For the multiplier statement, the error in replacing $\widehat{\sigma}$ by σ is bounded by $\|G_n^*\|_{\mathcal{Y}_0}\|1/\widehat{\sigma} - 1/\sigma\|_{\mathcal{Y}_0} = o_{P_*}(1)$ in probability, using conditional tightness and Step 3. The supremum norm is a continuous functional. Let $H(t) = P(\|V\|_{\mathcal{Y}_0} \le t)$ and $c = H^{-1}(1-\alpha)$. The theorem's quantile condition means H is continuous at c and $H(c-\delta) < 1-\alpha < H(c+\delta)$ for every sufficiently small $\delta > 0$. Conditional distribution convergence at continuity points on either side of c implies $\widehat{c}_{1-\alpha} \to_P c$. A sandwich argument then gives

$$P\{\Delta(y) \in C_n(y) \text{ for every } y \in \mathcal{Y}_0\} = P\left\{\left\|\frac{\sqrt{n}(\widehat{\Delta}-\Delta)}{\widehat{\sigma}}\right\|_{\mathcal{Y}_0} \le \widehat{c}_{1-\alpha}\right\} \longrightarrow 1-\alpha. \tag{26}$$

Only quantile regularity at the selected level is needed; it is stated explicitly for a continuum rather than inferred from a finite-dimensional argument.

For a fixed finite $\mathcal{Y}_0$, every coordinate of V has variance one. Thus $P(\max_j |V_j| = t) \le \sum_j P(|V_j| = t) = 0$. If its correlation matrix has rank $r \ge 1$, write $V = AU$ for $U \sim N_r(0, I_r)$ with A of full column rank. The map $u \mapsto \|Au\|_\infty$ is a norm, and every shell $\{u : a < \|Au\|_\infty < b\}$, $0 < a < b$, is nonempty and open. Its Gaussian probability is positive, proving strict increase of H and hence the finite-grid specialization even when the covariance is singular. At finite J, one can alternatively use conditional Chebyshev and a finite union bound in Step 2, so no learner-complexity restriction is necessary.

For unstudentized bands, let $\widehat{q}_{1-\alpha}$ be the conditional quantile of $\|G_n^*\|_{\mathcal{Y}}$. If the distribution of $\|\mathbb{G}\|_{\mathcal{Y}}$ satisfies the same quantile condition, the identical argument gives coverage $1-\alpha$ for $[\widehat{\Delta}(y) \pm \widehat{q}_{1-\alpha}/\sqrt{n}]$ simultaneously over $\mathcal{Y}$, including $\mathbb{R}$ when the assumptions hold there. No variance lower bound is needed for this unstudentized construction. In either construction, replacing the exact conditional quantile by B_n independent draws preserves coverage when $B_n \to \infty$; the fixed 500 draws used in the experiments are a numerical approximation.

A numerical grid $\mathcal{Y}_{0,n}$ approximates continuum studentization only if its supremum error is negligible. A sufficient condition is that it is asymptotically dense in the Gaussian L^2 semimetric on $\mathcal{Y}_0$, together with stochastic equicontinuity of both studentized processes in that metric; then their suprema on $\mathcal{Y}_{0,n}$ differ from those on $\mathcal{Y}_0$ by $o_P(1)$ (conditionally for multipliers). The experiments do not invoke such a discretization limit: their reported event is coverage at every point of the stated fixed grid. □

C.3 Observed-data efficient influence function

We prove the efficiency assertion of Theorem 2 in the observed-data model. The bound is stated for every fixed finite collection of thresholds; it does not assert a separate infinite-dimensional minimax bound. Fix w and y. All conditional densities below may be probability mass functions or Radon–Nikodym derivatives relative to appropriate dominating measures, so the argument includes discrete and mixed primary outcomes.

Step 1: observed likelihood and tangent space. Write $p_1(v \mid x,w,s) = p_{Y|X,W,\widetilde{Y},R=1}(v \mid x,w,s)$. With the outcome coordinate present only when $r=1$, the observed likelihood is

$$p_O(o) = p_X(x)p_{W|X}(w \mid x)p_{\widetilde{Y}|X,W}(s \mid x,w)p_{R|X,W,\widetilde{Y}}(r \mid x,w,s)\, p_1(v \mid x,w,s)^r. \tag{27}$$

In particular, the final factor is the *observed* conditional outcome distribution. Under MAR, its CDF satisfies

$$\nu_w^{R=1}(y,x,s) := \int \mathbf{1}\{v \le y\} p_1(v \mid x,w,s)\, dv = \nu_w(y,x,s). \tag{28}$$

Thus the CDF defining the outcome tangent direction is the same ν_w that identifies the target. MAR and exchangeability impose no additional equality restrictions on the observed likelihood: any observed law with overlap admits a MAR full-data extension and an exchangeable potential-outcome

extension. The positive lower bounds ensure finite inverse weights and allow sufficiently small bounded tilts to remain in a model with overlap. We impose no restriction such as a treatment-invariant surrogate distribution.

Let $L_0^2(P)$ be the Hilbert space of square-integrable, mean-zero functions with inner product $\langle f, g\rangle = P(fg)$. Consider a dominated one-dimensional submodel $\{P_t : |t| < t_0\}$ through $P_0 = P$, differentiable in quadratic mean, with score $\dot{\ell}_O$. Differentiating the factors of (27) gives

$$\dot{\ell}_O = \dot{\ell}_X + \dot{\ell}_W + \dot{\ell}_S + \dot{\ell}_R + R\dot{\ell}_Y. \tag{29}$$

Normalization of each factor implies conditional mean zero given its conditioning variables; for the last factor, $\mathbb{E}[\dot{\ell}_Y \mid X, W, \widetilde{Y}, R = 1] = 0$. Define the closed component spaces

$$\begin{aligned}
\mathcal{T}_X &= \{a(X) : Pa = 0,\ Pa^2 < \infty\},\\
\mathcal{T}_W &= \{a(X, W) : \mathbb{E}[a \mid X] = 0,\ Pa^2 < \infty\},\\
\mathcal{T}_S &= \{a(X, W, \widetilde{Y}) : \mathbb{E}[a \mid X, W] = 0,\ Pa^2 < \infty\},\\
\mathcal{T}_R &= \{a(X, W, \widetilde{Y}, R) : \mathbb{E}[a \mid X, W, \widetilde{Y}] = 0,\ Pa^2 < \infty\},\\
\mathcal{T}_Y &= \{Ra(Y, X, W, \widetilde{Y}) : \mathbb{E}[a \mid X, W, \widetilde{Y}, R = 1] = 0,\ P(Ra^2) < \infty\}.
\end{aligned}$$

For example, $a_X \in \mathcal{T}_X$ and $a_W \in \mathcal{T}_W$ satisfy $P(a_X a_W) = P\{a_X \mathbb{E}[a_W \mid X]\} = 0$. The same conditioning argument handles all pairs; $\mathbb{E}[Ra \mid X, W, \widetilde{Y}, R] = 0$ handles the outcome component against the first four. The component spaces are therefore mutually orthogonal.

Bounded, conditionally centered functions generate factor submodels through tilts $p_t = p(1 + ta)$; centering preserves normalization and $|t|\|a\|_\infty < 1$ ensures nonnegativity. Truncation followed by conditional centering approximates square-integrable component functions by bounded ones. Consequently the tangent space is

$$\mathcal{T} = \mathcal{T}_X \oplus \mathcal{T}_W \oplus \mathcal{T}_S \oplus \mathcal{T}_R \oplus \mathcal{T}_Y. \tag{30}$$

It is the nonparametric observed-data tangent space: for any $h \in L_0^2(P)$, successive conditional-expectation differences along X, (X, W), $(X, W, \widetilde{Y})$, and $(X, W, \widetilde{Y}, R)$ give the first four components. The remaining innovation vanishes when $R = 0$, because no outcome is then observed, and belongs to $\mathcal{T}_Y$. This establishes the completeness of (30), not only pairwise orthogonality.

Step 2: pathwise derivative of the target. Along the submodel, write

$$F_{w,t}(y) = \int \mathbf{1}\{v \le y\} p_{1,t}(v \mid x, w, s) p_{S,t}(s \mid x, w) p_{X,t}(x)\, dv\, ds\, dx. \tag{31}$$

Differentiating the three factors at $t = 0$ gives

$$\begin{aligned}
\dot{F}_w(y) = {} & \mathbb{E}[\mu_w(y, X)\dot{\ell}_X(X)] + \mathbb{E}_X\mathbb{E}[\nu_w(y, X, \widetilde{Y})\dot{\ell}_S \mid X, W = w]\\
& + \mathbb{E}_X\mathbb{E}[Z_y\dot{\ell}_Y \mid X, W = w].
\end{aligned} \tag{32}$$

The last expectation uses the MAR extension of p_1; equivalently, its inner integral is with respect to $p_1(\cdot \mid X, w, \widetilde{Y})$ and its outer surrogate integral uses $p_S(\cdot \mid X, w)$. Differentiation is immediate for bounded factor tilts. The resulting functional is continuous in the observed score's $L^2(P)$ norm, by the bounded inverse weights below, and hence extends to the closure of regular scores. There is no derivative in the treatment or labeling factor. In particular, throughout (32) the outer X law is marginal. The identity used for standardization is

$$\mathbb{E}\left[\frac{I_w}{\pi_w(X)} h(Y, X, W, \widetilde{Y})\right] = \mathbb{E}_X\mathbb{E}[h(Y, X, W, \widetilde{Y}) \mid X, W = w], \tag{33}$$

which follows by conditioning first on X. It cannot generally be replaced by $\mathbb{E}[h \mid W = w]$.

Step 3: membership of the proposed gradient. Decompose $\varphi_w = M_w + N_w + L_w$, where

$$M_w = \mu_w - F_w, \qquad N_w = \frac{I_w}{\pi_w}(\nu_w - \mu_w), \qquad L_w = \frac{I_w R}{\pi_w \rho}(Z_y - \nu_w).$$

All three are bounded, with $|M_w| \le 1$, $|N_w| \le 1/\epsilon$, and $|L_w| \le 1/(\epsilon\rho_{\min})$. Since $P\mu_w = F_w$, $M_w \in \mathcal{T}_X$. Also,

$$\mathbb{E}[N_w \mid X, W] = \frac{I_w}{\pi_w}\{\mathbb{E}[\nu_w \mid X, W = w] - \mu_w\} = 0,$$

where the displayed expression is zero off the arm by $I_w = 0$, and the tower property applies on the arm. Hence $N_w \in \mathcal{T}_S$.

For the outcome component take

$$h_Y = \frac{I_w}{\pi_w \rho}(Z_y - \nu_w), \qquad \mathbb{E}[h_Y \mid X, W, \widetilde{Y}] = 0. \tag{34}$$

On $\{W = w\}$, this conditional expectation equals $(\pi_w\rho)^{-1}\mathbb{E}[Z_y - \nu_w \mid X, W = w, \widetilde{Y}] = 0$; on $\{W \neq w\}$ it vanishes because $I_w \equiv 0$. By (28), the same two-case argument gives $\mathbb{E}[h_Y \mid X, W, \widetilde{Y}, R = 1] = 0$ for the observed outcome factor. Thus its observed-data score is $Rh_Y = L_w$, and $L_w \in \mathcal{T}_Y$.

To exhibit the submodel explicitly, perturb only the observed outcome factor:

$$p_{1,t}(v \mid x, u, s) = p_1(v \mid x, u, s)\left[1 + t\frac{\mathbf{1}\{u = w\}}{\pi_w(x)\rho(x, u, s)}\{\mathbf{1}\{v \le y\} - \nu_w(y, x, s)\}\right].$$

Its conditional integral is one, and it is nonnegative for $|t| < \epsilon\rho_{\min}$. Boundedness permits local density tilts with observed score Rh_Y. Analogous tilts of p_X by M_w and of p_S by N_w realize the other two components. Therefore $\varphi_w \in \mathcal{T}_X \oplus \mathcal{T}_S \oplus \mathcal{T}_Y \subset \mathcal{T}$.

Step 4: derivative representation and the bound. Orthogonality reduces the derivative representation to three inner products. For the covariate factor,

$$P(\varphi_w \dot{\ell}_X) = P(M_w \dot{\ell}_X) = P(\mu_w \dot{\ell}_X),$$

since $P\dot{\ell}_X = 0$. For the surrogate factor, (33) and conditional centering give

$$\begin{aligned} P(\varphi_w \dot{\ell}_S) &= P\left[\frac{I_w}{\pi_w}(\nu_w - \mu_w)\dot{\ell}_S\right] \\ &= \mathbb{E}_X\mathbb{E}[\nu_w \dot{\ell}_S \mid X, W = w]. \end{aligned}$$

For the outcome factor, $R^2 = R$, MAR, and $\mathbb{E}[\dot{\ell}_Y \mid X, W, \widetilde{Y}, R = 1] = 0$ imply

$$\begin{aligned} P(\varphi_w R\dot{\ell}_Y) &= P\left[\frac{I_w R}{\pi_w \rho}(Z_y - \nu_w)\dot{\ell}_Y\right] \\ &= P\left[\frac{I_w}{\pi_w}\mathbb{E}[(Z_y - \nu_w)\dot{\ell}_Y \mid X, W, \widetilde{Y}, R = 1]\right] \\ &= \mathbb{E}_X\mathbb{E}[Z_y \dot{\ell}_Y \mid X, W = w]. \end{aligned}$$

Finally $P(\varphi_w \dot{\ell}_W) = P(\varphi_w \dot{\ell}_R) = 0$ by (30). Combining these identities with (32) proves

$$\dot{F}_w(y) = P(\varphi_w \dot{\ell}_O) \qquad \text{for every regular submodel score.} \tag{35}$$

The representation extends to $\mathcal{T}$ by Cauchy–Schwarz. Together with membership in $\mathcal{T}$, it identifies φ_w as the canonical gradient (Bickel et al., 1993; Tsiatis, 2006). Linearity makes $\varphi_1 - \varphi_0$ the canonical gradient of $\Delta(y)$, and Ω the efficient covariance matrix for the finite target vector. For any regular asymptotically linear estimator with influence vector g, $g - \boldsymbol{\varphi}$ is orthogonal to $\mathcal{T}$ coordinatewise; hence

$$P(gg^\top) = \Omega + P[(g - \boldsymbol{\varphi})(g - \boldsymbol{\varphi})^\top] \succeq \Omega.$$

The convolution theorem gives the corresponding lower bound for regular estimators more generally (van der Vaart, 2000, Thm. 25.20). Identification of the canonical gradient above follows from its tangent-space membership and derivative representation; the convolution theorem supplies the lower bound once that gradient is identified.

Step 5: regularity and attainment. To make attainment precise, fix a differentiable-in-quadratic-mean submodel with scalar score s, information $I = Ps^2$, and local sequence $t_n = h/\sqrt{n}$. Local asymptotic normality (van der Vaart, 2000, Thm. 7.2) gives

$$\log \frac{dP_{t_n}^{\otimes n}}{dP^{\otimes n}} = \frac{h}{\sqrt{n}} \sum_i s(O_i) - \frac{h^2 I}{2} + o_P(1).$$

The joint CLT for $(\boldsymbol{\varphi}, s)$ has cross-covariance $P(\boldsymbol{\varphi} s) = \dot{\boldsymbol{\Delta}}$. Contiguity transfers the asymptotic-linear remainder of Appendix C.2 to $P_{t_n}^{\otimes n}$. Le Cam's third lemma (van der Vaart, 2000, Ex. 6.7) shifts the Gaussian mean of $n^{-1/2} \sum_i \boldsymbol{\varphi}(O_i)$ by $h\dot{\boldsymbol{\Delta}}$. Since $\sqrt{n}(\boldsymbol{\Delta}_{t_n} - \boldsymbol{\Delta}) = h\dot{\boldsymbol{\Delta}} + o(1)$,

$$\sqrt{n}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}_{t_n}) \rightsquigarrow N_J(0, \Omega) \qquad \text{under } P_{t_n}^{\otimes n}.$$

The limit does not depend on h, establishing local regularity and attainment of the efficient covariance bound. If treatment or labeling probabilities are known, the corresponding score spaces are removed from (30); the gradient remains in the reduced tangent space and the same conclusion follows. No stronger restriction on the surrogate law is used. □

C.4 Variance decompositions

We prove the variance comparison in Theorem 2 and the surrogate-gain identity (12). Fix y and evaluate all scores at the true nuisances. First,

$$PH_w = P\left[\frac{I_w R Z_y}{\pi_w \rho}\right] = P\left[\frac{I_w \nu_w}{\pi_w}\right] = P\mu_w = F_w.$$

Expanding (4) gives the exact difference

$$H_w - F_w = \varphi_w + A_w + B_w, \qquad A_w = \mu_w(I_w/\pi_w - 1), \quad B_w = I_w \nu_w (R/\rho - 1)/\pi_w. \tag{36}$$

Let $\mathcal{C} = \sigma(X, W, \widetilde{Y})$ and $\mathcal{D} = \sigma(X, W, \widetilde{Y}, R)$. The following conditional identities collect the needed orthogonality statements:

$$\begin{aligned} &\mathbb{E}[A_w \mid X] = 0, \qquad \mathbb{E}[B_w \mid \mathcal{C}] = 0, \\ &M_v \in \sigma(X), \quad \mathbb{E}[N_v \mid X, W] = 0, \quad N_v \in \mathcal{C}, \quad \mathbb{E}[L_v \mid \mathcal{D}] = 0. \end{aligned}$$

Here $M_v \in \sigma(X)$ denotes measurability. For either $v, w \in \{0, 1\}$, these give all six cross-inner-products:

$$\begin{aligned} P(M_v A_w) &= P\{M_v \mathbb{E}[A_w \mid X]\} = 0, & P(M_v B_w) &= P\{M_v \mathbb{E}[B_w \mid \mathcal{C}]\} = 0, \\ P(N_v A_w) &= P\{A_w \mathbb{E}[N_v \mid X, W]\} = 0, & P(N_v B_w) &= P\{N_v \mathbb{E}[B_w \mid \mathcal{C}]\} = 0, \\ P(L_v A_w) &= P\{A_w \mathbb{E}[L_v \mid \mathcal{D}]\} = 0, & P(L_v B_w) &= P\{B_w \mathbb{E}[L_v \mid \mathcal{D}]\} = 0. \end{aligned}$$

Therefore $P\{(\varphi_1 - \varphi_0)(A_1 - A_0 + B_1 - B_0)\} = 0$. Taking variances in the arm difference of (36) proves (11). Moreover $\mathrm{Cov}(A_v, B_w) = 0$ by conditioning on $\mathcal{C}$, so neither the treatment-weight nor label-weight residual can cancel the other in that variance gap. This proves the known-weight oracle comparison directly, without treating oracle IPW as the influence function of an estimator that fits its weights.

For the covariate-only outcome comparator, direct subtraction gives

$$C_w - \Psi_w = D_w := \frac{I_w}{\pi_w}\left(\frac{R}{\rho} - 1\right)(\nu_w - \mu_w). \tag{37}$$

Here $\mathbb{E}[D_w \mid \mathcal{C}] = 0$, so $PD_w = 0$. The previous calculations with D_w in place of B_w yield $P(M_v D_w) = P(N_v D_w) = P(L_v D_w) = 0$ for both arms. In addition $D_1 D_0 = 0$ by disjoint arm support and thus $\mathrm{Cov}(D_1, D_0) = 0$. As $R \mid \mathcal{C}$ is Bernoulli with parameter ρ,

$$\mathbb{E}[(R/\rho - 1)^2 \mid \mathcal{C}] = \frac{\rho}{\rho^2} - 2\frac{\rho}{\rho} + 1 = \frac{1 - \rho}{\rho}.$$

Consequently

$$\mathrm{Var}(C_1 - C_0) - \mathrm{Var}(\varphi_1 - \varphi_0) = \mathrm{Var}(D_1 - D_0) = \sum_{w=0}^{1} PD_w^2$$
$$= \sum_{w=0}^{1} P\left[\frac{I_w}{\pi_w^2}\frac{1-\rho}{\rho}(\nu_w - \mu_w)^2\right],$$

which is (12). Since each integrand is nonnegative and π_w, ρ are bounded away from zero, the gap is zero if and only if $I_w(1-\rho)(\nu_w - \mu_w)^2 = 0$ almost surely for both arms. It is strictly positive if this quantity is positive with positive probability for at least one arm. The identities apply at each specified threshold; usefulness for a conditional mean alone does not imply a gain at every cutpoint. □

D Experimental Specification and Additional Results

Implementation and recorded outputs. All manuscript figures and tables are embedded directly from saved replication-level results by the figure-generation script. Every simulation uses seeds $1000 + b$ for repetition b; bootstrap multipliers use independent seeds $900000 + b$. HelpSteer2 and wine masks use seed 0; application record bootstraps use seed 707. BiGGen assignment and annotation seeds are specified below. Folds are generated by shuffled five-fold splitting with seed 123. All empirical methods in a comparison share bootstrap resamples.

Conditional CDF learning. We use logistic regression with an intercept, L^2 regularization parameter `C=1`, a maximum of 1000 iterations, and standardization fitted within the relevant training data. μ_w uses inverse-label weights normalized to have training-sample mean one; ν_w uses $(X, \widetilde{Y})$ without label weights. The ablation uses the same fitted μ_w for both prediction slots. Known probabilities are used for the simulation comparisons in Figure 2, empirical MCAR masks, and the ACTG working treatment probability. The 5000-repetition band experiment estimates the simulation probabilities: π from X and ρ from $(X, W, \widetilde{Y})$, using the same cross-fitting folds. Estimated ρ also supplies the weights for fitting μ_w. The logistic labeling model contains the constant ordinal MCAR probability but only approximates the bounded continuous MAR mechanism below. Unknown scores use logistic regression, with fitted probabilities clipped to $[0.02, 0.98]$. Working outcome regressions need not equal the true conditional CDFs, even with the simulation's quadratic feature.

D.1 Simulation design and diagnostics

Oracle IPW normalization. The simulation's oracle IPW baseline uses known weights and the Horvitz–Thompson arm estimator $n^{-1}\sum_i I_{wi}R_iZ_{yi}/(\pi_w\rho_i)$. It does not divide by the realized sum of weights. Under MCAR with label probability r and $\pi_0 = \pi_1 = 1/2$, its exact DTE mean squared error is

$$\mathrm{MSE}\{\widehat{\Delta}_{\mathrm{HT}}(y)\} = \frac{1}{n}\left[\frac{2}{r}\{F_1(y) + F_0(y)\} - \Delta(y)^2\right].$$

This follows from unbiasedness, disjoint arm indicators, and $Z_y^2 = Z_y$. As both CDFs approach one, the MSE approaches $4/(nr)$ rather than zero: the estimated arm totals still fluctuate. Thus increasing right-tail MSE is expected for this unnormalized estimator. Hájek IPW instead divides each arm's weighted outcome total by its weight total; under randomized MCAR sampling it equals the labeled-only empirical CDF comparator. The variance guarantee in Theorem 2 concerns oracle IPW, and gains over it should not be interpreted as gains over normalized IPW.

Continuous MAR labeling. Let $f_W(X)$ be the arm-specific mean in Section 6.1, and $b(X) = 0.3(X_1 - 0.5X_2)$. We set

$$\rho(X, W, \widetilde{Y}) = 0.05 + 0.30\,\mathrm{expit}\{0.4X_1 + 0.4[\widetilde{Y} - f_W(X) - b(X)]\}.$$

This is a fixed function of observed variables, bounded between 0.05 and 0.35. Since the logit argument is centered and symmetric under the DGP, $\mathbb{E}\rho = 0.20$. Label selection depends on the surrogate

and is MAR rather than MCAR. All simulated records remain i.i.d.; the probability is not normalized by a sample-dependent empirical mean. Continuous thresholds are $\{-0.75, -0.5, 0, 0.5, 1, 1.5\}$. At the left boundary, $F_0(-0.75) = 0.1119$ and $F_1(-0.75) = 0.1016$; both arms are evaluated at the same outcome threshold. Ordinal cutpoints are $\{0, 1, 2, 3\}$ after assigning categories 0–4 using the four stated boundaries.

Known truth and simulation summaries. The control CDF is $\Phi\{y/\sqrt{0.3^2 + 0.2^2 + 0.5^2}\}$. The treated CDF is

$$F_1(y) = \int_{-\infty}^{\infty} \Phi\left(\frac{y - 0.4x^2}{\sqrt{0.6^2 + 0.5^2}}\right) \phi(x)\, dx.$$

Numerical quadrature over $[-10, 10]$ with absolute tolerance 10^{-11} supplies the reference. The ordinal CDFs equal these continuous CDFs at the category boundaries. Bias is the replication mean error, MSE is the replication mean squared error, and coverage is the fraction of intervals containing this reference. Reported SE calibration compares the mean estimated SE to the empirical standard deviation across repetitions. Neither bias nor coverage is inferred from a single real-data point estimate. Figure 2 and the ordinal baseline comparisons use 500 repetitions at $n = 2000$ with known probabilities. The band experiment uses 5000 repetitions for each outcome at $n = 2000$, estimating both π and ρ.

Diagnosing continuous-band undercoverage. We use 5000 paired repetitions at $n = 2000$, replacing both fitted outcome CDFs with their exact DGP values. The $n = 2000$ run uses the same data seeds as the main band experiment but retains known treatment and labeling probabilities to isolate outcome fitting. Writing $f_w(X)$ for the arm mean and $b(X)$ for the surrogate bias, Gaussian conditioning gives

$$\mu_w(y, X) = \Phi\left\{\frac{y - f_w(X)}{0.5}\right\}, \qquad \nu_w(y, X, s) = \Phi\left\{\frac{y - f_w(X) - \frac{0.25}{0.61}[s - b(X) - f_w(X)]}{\sqrt{0.25 \cdot 0.36/0.61}}\right\}.$$

Treatment and labeling probabilities, evaluation grids, and multiplier seeds are unchanged. With $B = 500$ Rademacher draws, replacing the fitted conditional CDFs by their true functions raises coverage from 0.929 to 0.944, a paired gain of 0.015 (Monte Carlo SE 0.003). Increasing to $B = 5000$, retaining the first 500 draws, yields coverage 0.931 with fitted CDFs and 0.944 with exact CDFs. Thus outcome fitting contributes to undercoverage, but replacing it does not fully restore nominal coverage.

Misses are dispersed across cutpoints, with the largest coordinate miss rate (0.015) shared by $y = -0.75$ and $y = -0.5$; mean estimated SE divided by empirical SD ranges from 0.981 to 1.007. Thus average SE calibration can coexist with a poorly calibrated maximum studentized error. Since design weights are known, the conditional bias identity is zero for arbitrary training-fold outcome regressions whenever the fits exist. This diagnostic implicates the fitted-score approximation, but does not distinguish limited training data from the working logistic link. Supplementary outputs also report conditional Gaussian calibration and a separate population-covariance check.

Thresholdwise PPI baseline. For arm w, let $Q_i(y) = \mathbf{1}\{\widetilde{Y}_i \le y\}$. Under MCAR, the baseline estimates its primary-outcome CDF as the average of $Q_i(y)$ over all arm-w records plus the average of $Z_{iy} - Q_i(y)$ over labeled arm-w records. These are nested samples. Writing their respective empirical means as $\bar{Q}_w$ and $\bar{B}_w$, its estimated influence function is

$$\frac{I_{wi}}{\widehat{p}_w}(Q_i - \bar{Q}_w) + \frac{I_{wi}R_i}{\widehat{p}_{wR}}(Z_{iy} - Q_i - \bar{B}_w),$$

where $\widehat{p}_w = n_w/n$ and $\widehat{p}_{wR} = n_{wR}/n$. We compute the SE of the difference of these arm functions on the same records. This accounts for overlap between labeled and full samples; an independent-sample PPI variance formula would be inappropriate. No power-tuning parameter is fitted.

D.2 Ordinal baseline implementations and paired comparisons

All added comparisons use the same 500 ordinal datasets at $n = 2000$, generated with seeds $1000+b$. Their outcome-observation masks, targets, and cutpoints coincide with the original simulation. The

Outcome	y	Bias	SD	Mean SE	Mean SE / SD	Coverage
Continuous	-0.75	-0.0015	0.0283	0.0285	1.00	0.950
	-0.5	-0.0028	0.0335	0.0332	0.99	0.944
	0	+0.0015	0.0379	0.0375	0.99	0.952
	0.5	+0.0014	0.0329	0.0326	0.99	0.944
	1	-0.0014	0.0232	0.0237	1.02	0.954
	1.5	-0.0022	0.0157	0.0159	1.01	0.942
Ordinal	0	-0.0013	0.0257	0.0265	1.03	0.958
	1	-0.0006	0.0362	0.0352	0.97	0.962
	2	+0.0021	0.0368	0.0367	1.00	0.962
	3	+0.0003	0.0308	0.0306	0.99	0.958

Table 2: Proposed-estimator pointwise diagnostics at $n = 2000$ (500 repetitions). SD is the empirical standard deviation across repetitions; mean SE averages the score-based DTE SE retaining arm covariance.

algorithms below use only observed primary outcomes. They require randomized treatment and MCAR labeling and are not presented as solutions to selective labeling or observational treatment assignment.

Labeled-only regression-adjusted DTE. We implement the regression-adjusted CDF in Byambadalai et al. (2024, Eq. (1)) on the labeled subset $\mathcal{L} = \{i : R_i = 1\}$. With $m = |\mathcal{L}|$, $m_w = \sum_{i\in\mathcal{L}} I_{wi}$, and five-fold cross-fitted logistic CDF predictions $\widehat{g}_{wi}(y)$ based only on X, the estimator is

$$\widehat{F}_w^{\text{labeled-RA}}(y) = \frac{1}{m}\sum_{i\in\mathcal{L}} \widehat{g}_{wi}(y) + \frac{1}{m_w}\sum_{i\in\mathcal{L}} I_{wi}\{Z_{yi} - \widehat{g}_{wi}(y)\}.$$

The nuisance learner and covariate features match the proposed method. Unlabeled covariates and all surrogate scores are excluded. The centered arm influence estimates are

$$\widehat{g}_{wi} - \overline{g}_w + \frac{I_{wi}}{m_w/m}\{Z_{yi} - \widehat{g}_{wi} - \overline{e}_w\},$$

where $\overline{g}_w$ averages predictions over $\mathcal{L}$ and $\overline{e}_w$ averages residuals over its arm-w records. We take the arm difference before estimating its variance and scale its SE by $\sqrt{m}$.

Thresholdwise PPI++. For each arm and threshold, write q_y for a prediction of Z_y. Raw-score PPI++ uses $q_y = \mathbf{1}\{\widetilde{Y} \le y\}$; learned-CDF PPI++ uses logistic regression on $(X, \widetilde{Y})$, with the same learner as $\widehat{\nu}_w$. For outer fold k, estimate

$$\widehat{\lambda}_{kwy} = \text{clip}_{[0,1]}\left[\widehat{u}_{kw}\,\frac{\widehat{\text{Cov}}_{kw}(Z_y, q_y)}{\widehat{\text{Var}}_{kw}(q_y)}\right],$$

where $\widehat{u}_{kw}$ is the unlabeled fraction in the training arm. Covariance and variance use its labeled records; learned predictions for tuning come from an inner three-fold split. A zero proxy variance sets $\widehat{\lambda} = 0$. The learned predictor is then refitted on the outer training arm's labels and evaluated on its held-out records. This is the disjoint-sample mean power-tuning formula of Angelopoulos et al. (2023b, Sec. 6), applied to threshold indicators; the factor $\widehat{u}_{kw}$ accounts for the finite unlabeled sample.

Let $v_{wi} = \widehat{\lambda}_{k(i)wy}\widehat{q}_{wi}(y)$, and let u_w be the number of unlabeled arm-w records. We compute

$$\widehat{F}_w^{++}(y) = \frac{1}{m_w}\sum_{i:W_i=w,R_i=1}(Z_{yi} - v_{wi}) + \frac{1}{u_w}\sum_{i:W_i=w,R_i=0} v_{wi}.$$

For SEs and bands, its centered influence estimates on all n records are

$$\frac{I_{wi}R_i}{m_w/n}(Z_{yi} - v_{wi} - \overline{a}_w) + \frac{I_{wi}(1-R_i)}{u_w/n}(v_{wi} - \overline{b}_w),$$

Method	MSE $\times 10^3$	Reduction (%)	SE/SD	Pointwise coverage	Band	Width
Proposed	0.663–1.355	0	1.00	0.958–0.962	0.946	0.159
Covariate-only	0.755–1.656	8–21	0.99	0.944–0.960	—	—
PPI++ (raw)	0.732–1.767	9–23	1.00	0.938–0.958	0.958	0.173
PPI++ (CDF)	0.688–1.573	4–15	1.00	0.946–0.960	0.936	0.167
Labeled-only RA	0.814–1.728	18–24	0.99	0.936–0.960	0.942	0.177
Threshold PPI	1.276–2.174	34–48	0.97	0.934–0.944	—	—
Labeled only	0.946–2.448	30–45	0.99	0.940–0.956	—	—
Oracle IPW	1.055–7.880	37–88	0.99	0.938–0.950	—	—
Surrogate only	0.470–5.492	−177–82	1.02	0.076–0.956	—	—

Table 3: Paired ordinal MCAR comparisons: $n = 2000$, 500 repetitions. MSE, reduction, and pointwise coverage report ranges across the four cutpoints. Reduction is $100(1-\mathrm{MSE}_{\mathrm{proposed}}/\mathrm{MSE}_{\mathrm{row}})$ at each cutpoint; negative values indicate higher proposed MSE. SE/SD averages the ratio of mean reported SE to empirical SD across cutpoints. Band is simultaneous 95% coverage and Width is the mean full band width. Dashes indicate bands not computed for the simpler baselines. Monte Carlo SE at coverage 0.95 is approximately 0.010.

where $\overline{a}_w$ and $\overline{b}_w$ are the two corresponding sample averages. The arm difference retains their joint covariance. Tuning and preprocessing exclude the outer evaluation fold. Ordinary PPI retains the original nested-sample construction with unit weight; PPI++ uses disjoint samples and its matching tuning and SE formulas.

Evaluation. All new methods receive 500 shared Rademacher draws per repetition. Labeled-only RA uses the same draws restricted to labeled records. We include every prespecified cutpoint in the band event, including zero-SE cases. Table 3 reports empirical MSE, SE calibration, pointwise coverage, and simultaneous coverage and width.

D.3 Additional application details

Shared estimation and uncertainty. For each application, Table 1 compares the proposed method, covariate-only outcome adjustment, and IPW on the same records and cutpoints. All use the five-fold nuisance-learning procedure specified above. Each reported SE is the standard deviation of 200 nonparametric record-bootstrap estimates, using seed 707 and common resamples across methods. Every replicate refits estimated nuisances and repeats cross-fitting; known design probabilities remain fixed. The observed assignment and labeling indicators are resampled with their records, rather than redrawn. This includes the fitting variability of estimated IPW weights. Results are reported separately at each cutpoint; score-based proposed SEs are saved separately from bootstrap SEs. These calculations do not adjust for additional prompt, reviewer, finite-population, or trial-design dependence. They describe empirical precision; population coverage is assessed in the simulations.

D.3.1 BiGGen Bench

Data, contrast, and surrogate. We use the public `human_eval` split of BiGGen Bench (Kim et al., 2025); the pinned source revision is recorded in the reproduction package. It contains 2,780 responses from four models on 695 prompts. The primary contrast assigns $W = 0$ to Mistral-7B-Instruct-v0.2 and $W = 1$ to Mixtral-8x7B-Instruct-v0.1. The primary outcome is the released integer human rating on the 1–5 task-specific rubric; the surrogate is the cached GPT-4-Turbo-2024-04-09 score. We evaluate all nontrivial cutpoints $y \in \{1, 2, 3, 4\}$ without rounding or calibrating human ratings. Four missing human ratings, encoded as -1, belong to Llama-2-13b-hf, which is excluded from the reported comparisons. The models used below have complete human, GPT-4 Turbo, and Claude ratings.

Labeling and covariates. Each prompt contributes one response, selected by independent Bernoulli$(1/2)$ model assignment; exactly 200 selected responses are then sampled uniformly for annotation. The fixed application uses repetition $b = 0$ from `SeedSequence([20260920,b])`, without searching across assignments. Known probabilities are $\pi_w = 1/2$ and marginal $\rho =$

$200/695$. Covariates are log character and word counts for both user and system prompts, user-prompt character diversity, and seven capability indicators. Response features do not enter the treatment model. Unobserved primary outcomes are represented by zero before fitting; counterfactual ratings are withheld from every estimator. Using one response per prompt also prevents its two model outputs from entering different folds.

Estimation and uncertainty. The fixed application resamples its 695 observed prompt records under the shared 200-replicate bootstrap above. Both probabilities remain known; scaling and outcome fitting occur within training folds. In addition to the three main comparators, Table 4 reports labeled-only, labeled-only RA, raw-score and learned-CDF PPI++, thresholdwise PPI, and surrogate-only estimates. Their implementations use the same learners and sampling conventions described in Appendix D.2. Bootstrap SEs for the fixed application are distinct from the score-based SEs and empirical SDs in the repeated-assignment experiment below.

Results and interpretation. The target compares generation policies on released responses, not a deployed user A/B test. Rubrics vary by task, so ratings represent task-specific levels rather than a universal quality scale. BiGGen's annotator qualification used agreement with GPT-4-1106 scores; human–surrogate agreement is therefore not wholly independent validation. Cached scores are reused without judge training or new API calls. Table 4 separates fixed-application precision from error against the fully rated benchmark. Differences depend on the cutpoint and comparator and do not establish uniform finite-sample superiority over PPI++.

Additional benchmark comparisons. For repetitions $b = 0, \ldots, 499$, `SeedSequence([20260920,b])` generates the assignment and 200-label sample. Assignments and labels are common across baselines and surrogate sensitivities. For the fixed prompt pool $\mathcal{P}$ and the two released outputs, the reference DTE is

$$\Delta_{\mathcal{P}}(y) = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \big[\mathbf{1}\{Y_i(1) \le y\} - \mathbf{1}\{Y_i(0) \le y\}\big].$$

At each cutpoint separately, we average squared error against this reference over 500 repetitions. No MSE or SE is pooled across cutpoints. The uncertainty calculation uses the i.i.d. arm-score difference and 500 common Rademacher draws from `SeedSequence([20260920,b,1])`. Labeled-only RA restricts these draws to labeled units and uses its labeled sample size. Labeling is without replacement and the prompt pool is fixed; the reported SEs do not include finite-population corrections.

The model sensitivity replaces Mixtral with gpt-3.5-turbo-0125 on the same 695 prompts; the surrogate sensitivity replaces GPT-4 Turbo with Claude-3-Opus in the primary comparison. These choices were fixed before the 500-repetition runs. No repetition or cutpoint is removed, and all three 200-label conditions finish without an empty-arm fitting failure or zero-SE cutpoint. Table 5 reports the additional comparisons. The submission contains the three 200-label conditions; exploratory 100-label runs remain outside the submission package.

D.3.2 HELPSTEER2

Data, contrast, and surrogate. We use the 20,324 prompt–response pairs in the HelpSteer2 training split (Wang et al., 2024); the pinned source revision is recorded in the reproduction package. The primary outcome is released human helpfulness on the 0–4 scale. The contrast indicator is $W = \mathbf{1}\{\text{complexity} \ge 2\}$. The surrogate is the stored `gpt-5.6-terra` batch rating under the HelpSteer2 helpfulness rubric, without retraining on released human ratings. We evaluate $y \in \{0, 1, 2, 3\}$. Missing judge scores trigger an error rather than imputation based on human ratings; the exact rubric and prompt template are included in the reproduction package.

Labeling and covariates. A Bernoulli MCAR mask with probability 0.02 and NumPy seed 0 retains 399 human ratings. Hidden labels are replaced by zero before fitting. Covariates are $\log(1 +$ prompt characters$)$, $\log(1 +$ prompt words$)$, and distinct-character count divided by prompt length. The prepared record order and mask are fixed across methods.

		Fixed assignment		Repeated assignment and labeling			
Method	y	$\widehat{\Delta}$	Boot. SE	MSE	SD	Mean SE	Coverage
Proposed	1	$+0.025$	0.042	1.430	0.038	0.042	0.980
	2	$+0.012$	0.055	2.746	0.052	0.054	0.952
	3	-0.058	0.066	3.332	0.058	0.062	0.964
	4	-0.052	0.068	4.205	0.065	0.067	0.960
Covariate-only	1	$+0.023$	0.045	1.793	0.042	0.045	0.960
	2	-0.002	0.056	3.421	0.058	0.060	0.952
	3	-0.081	0.064	4.272	0.065	0.069	0.950
	4	-0.071	0.069	4.696	0.069	0.070	0.958
Labeled only	1	$+0.009$	0.043	1.654	0.041	0.043	0.952
	2	-0.023	0.057	3.288	0.057	0.058	0.950
	3	-0.105	0.065	4.269	0.065	0.067	0.952
	4	-0.042	0.070	4.414	0.066	0.068	0.958
Labeled-only RA	1	$+0.024$	0.045	1.751	0.042	0.044	0.952
	2	$+0.001$	0.057	3.332	0.057	0.059	0.956
	3	-0.090	0.066	4.100	0.064	0.067	0.950
	4	-0.048	0.068	4.479	0.067	0.069	0.960
PPI++ (raw)	1	$+0.029$	0.042	1.345	0.037	0.040	0.960
	2	-0.045	0.055	2.570	0.051	0.053	0.956
	3	-0.090	0.063	3.352	0.058	0.061	0.958
	4	-0.020	0.071	4.311	0.066	0.067	0.952
PPI++ (CDF)	1	$+0.018$	0.044	1.435	0.038	0.042	0.958
	2	-0.006	0.054	2.718	0.052	0.055	0.942
	3	-0.074	0.063	3.386	0.058	0.062	0.964
	4	-0.051	0.068	4.381	0.066	0.067	0.954
Threshold PPI	1	$+0.028$	0.043	1.453	0.038	0.041	0.968
	2	-0.052	0.057	2.855	0.053	0.055	0.958
	3	-0.079	0.069	3.844	0.062	0.065	0.962
	4	$+0.005$	0.080	5.034	0.071	0.073	0.962
Oracle IPW	1	-0.020	0.045	1.815	0.043	0.045	0.950
	2	-0.080	0.063	3.917	0.063	0.065	0.962
	3	-0.190	0.080	6.186	0.079	0.082	0.958
	4	-0.200	0.112	11.109	0.105	0.112	0.974
Surrogate only	1	-0.026	0.021	0.378	0.016	0.019	0.952
	2	-0.055	0.028	0.560	0.023	0.027	0.984
	3	-0.025	0.035	1.441	0.028	0.035	0.942
	4	-0.006	0.030	3.503	0.025	0.029	0.566

Table 4: Pointwise BiGGen results with 200 human ratings. The first two numerical columns describe the fixed application in Table 1, including its refitted-bootstrap SE. Remaining columns use 500 repeated assignments and annotation samples: MSE ($\times 10^{-3}$), empirical SD, mean score-based SE, and pointwise benchmark coverage, each computed separately at the indicated cutpoint. Surrogate-only targets automated ratings, not the human-rating distribution.

Estimation and uncertainty. We estimate $\pi(X)$ by cross-fitted logistic regression and use the known labeling probability $\rho = 0.02$. Proposed, covariate-only, and IPW estimates use the shared 200-replicate record bootstrap, refitting propensity and outcome regressions as applicable. Table 1 reports each cutpoint's estimate and bootstrap SE; Figure 3 visualizes the same contrasts with pointwise SE bars.

Results and interpretation. Complexity is an observed response attribute, so the target is a descriptive contrast standardized over prompt covariates rather than the causal effect of a model or prompt intervention. Responses may share prompts; record-bootstrap SEs are not cluster-adjusted deployment uncertainty. We do not claim absence of pretraining contamination in the judge. Smaller SEs alone establish neither accuracy nor coverage.

Setting	y	MSE	Mean SE	/ Cov.	/ RA	/ PPI++	/ CDF	Coverage
GPT-3.5 comparison	1	1.443	0.040	0.842	0.883	1.005	0.989	0.960
	2	2.789	0.054	0.799	0.802	0.967	0.967	0.952
	3	3.342	0.062	0.763	0.782	0.993	0.978	0.950
	4	3.922	0.067	0.860	0.885	0.960	0.928	0.962
Claude surrogate	1	1.509	0.042	0.842	0.862	1.070	1.024	0.958
	2	2.809	0.056	0.821	0.843	1.016	0.994	0.958
	3	3.343	0.063	0.783	0.815	0.959	0.964	0.964
	4	4.439	0.068	0.945	0.991	1.009	0.992	0.962

Table 5: Pointwise BiGGen sensitivity results with 200 human ratings. Each row reports proposed MSE ($\times 10^{-3}$), mean score-based SE, four MSE ratios against the indicated baselines, and pointwise benchmark coverage over 500 repetitions. GPT-3.5 replaces Mixtral as the treatment model; Claude replaces GPT-4 Turbo as the surrogate. No metric is averaged across cutpoints.

D.3.3 WINE REVIEWS

Data, contrast, and surrogate. The prepared sample contains 10,000 US wine reviews from `spawn99/wine-reviews`, filtered to nonmissing points, price, description, variety, and state, and restricted to unambiguously red or white varieties (Ji et al., 2025). We set $W = 1$ for red and $W = 0$ for white wine. The primary outcome is the professional points rating; the surrogate is the cached `gpt-5.6-terra` rating of the tasting description on a 1–5 rubric. We use raw judge ratings without calibration on professional ratings, evaluating primary-outcome thresholds $y \in \{85, 87, 89, 91, 93\}$. The prompt template and score order are recorded in the reproduction package.

Labeling and covariates. A Bernoulli MCAR mask with probability 0.20 and NumPy seed 0 retains 2,049 professional ratings. Unobserved primary outcomes are represented by zero before fitting. Covariates are log price, its square, and indicators for California, Washington, Oregon, New York, and Virginia, with other states as the reference.

Estimation and uncertainty. We estimate $\pi(X)$ by cross-fitted logistic regression and use known $\rho = 0.20$. Conditional regressions predict primary-outcome threshold probabilities from the auxiliary score directly, allowing the primary-outcome and surrogate scales to differ. Proposed, covariate-only, and IPW SEs use the shared 200-replicate record bootstrap, refitting all estimated nuisances. Table 1 and Figure 3 report the threshold-specific results.

Results and interpretation. Price and state standardization defines a descriptive target, not a manipulable wine-color intervention. Description-based surrogate ratings are auxiliary measurements of existing reviews rather than blinded assessments of new wines. The record bootstrap does not account for reviewer or winery dependence. As in HelpSteer2, these results compare precision without known population truth for validating coverage.

D.3.4 ACTG 175

Data, contrast, and surrogate. We retain arms 0 and 1 of the released ACTG 175 data (Hammer et al., 1996): 1,054 patients comparing zidovudine–didanosine ($W = 1$) with zidovudine monotherapy ($W = 0$). The primary outcome is week-96 CD4 (`cd496`), the surrogate is week-20 CD4 (`cd420`), and observation of week-96 CD4 is recorded by the supplied `r` indicator. We evaluate thresholds $y \in \{100, 200, 300, 400, 500\}$.

Labeling and covariates. There are 654 observed primary outcomes; no synthetic annotation mask is imposed. The baseline covariates are `cd40`, `age`, `wtkg`, `karnof`, `hemo`, `homo`, `drugs`, `race`, `gender`, `symptom`, `oprior`, `z30`, `preanti`, and `str2`. Missing primary-outcome values are masked before estimation, while week-20 CD4 enters the labeling and surrogate-augmented outcome regressions.

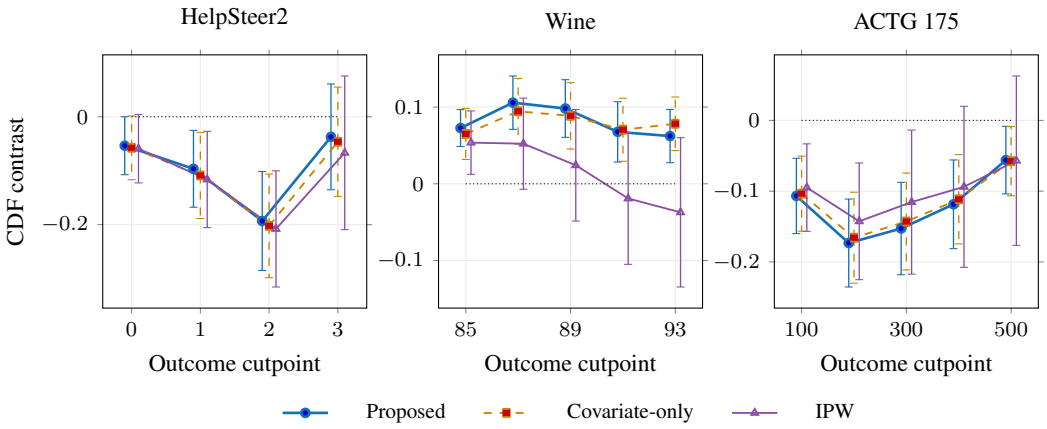


Figure 3: Empirical standardized distributional contrasts with ± 1.96 bootstrap-SE bars. Methods are horizontally offset around each primary-outcome cutpoint for visibility; all estimates and SEs are evaluated at the original cutpoints. The same nuisance-refitting bootstrap is used for proposed, covariate-only outcome adjustment, and IPW. Bars visualize estimated precision; they are not validated coverage statements or simultaneous bands.

Estimation and uncertainty. We use the working treatment probability $\pi = 1/2$ and estimate $\rho(X, W, \widetilde{Y})$ by cross-fitted logistic regression. Under the stated MAR assumption, this permits label observation to depend on baseline covariates, treatment, and week-20 CD4. The shared 200-replicate record bootstrap refits labeling and outcome models. Table 1 and Figure 3 report the results at each CD4 threshold.

Results and interpretation. A fitted labeling model is not evidence of MAR or noninformative censoring. The i.i.d. working analysis does not reproduce the trial's original stratified assignment, so these bootstrap SEs are not a coverage claim under that design. No clinical decision or general stochastic-dominance conclusion is inferred from the illustrative estimates.