EVIDENCE RESOLUTION: CERTIFYING HOW MUCH PERSONALIZATION FINITE EVIDENCE CAN SUPPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Flexible treatment-effect learners can propose action rules more finely than finite randomized evidence can defend. We formalize **evidence resolution**: the finest clinically worthwhile action granularity supportable by a representation and prespecified evidential criterion. A symmetric one-signal oracle yields the worst-case frontier $npd^2 = \Theta(\log M)$; instance-dependent theory separates exact localization from value-tolerant localization and yields a search-versus-multiplicity crossover against exhaustive certification. HonestER freezes representation choices upstream and independently certifies refinements with family-local backoff. A prospectively locked Criteo model for exact reference-best recovery plus certification failed as specified. The decomposition explains why: at the finest depth, selected regions certified in 94% of runs but matched the empirical reference-best bin only 40%, while the leading bins were only 0.35 standard errors apart. Post hoc, a zero-fit certification-resolution prediction from frozen reference geometry matches the observed R_{80} at 8 of 11 sample sizes and misses the other three by one bit. This supports a central distinction: certifiable resolution, exact localization, and deployed value are different statistical objects.

1 INTRODUCTION

Modern causal learners can produce individualized treatment rankings at a granularity far beyond the local information available in a trial. The resulting question is not only whether a policy has high estimated welfare or whether a discovered subgroup is statistically significant. It is: **how fine an action distinction can the current representation and sample actually defend?**

We call this quantity **evidence resolution**. It separates representation expressivity, certifiable resolution, exact candidate localization, and deployed policy value. These objects can diverge: a representation may omit useful actions; finite data may be unable to identify a unique best refinement; several near-tied refinements may nevertheless have essentially the same value; and even a high-value candidate may still lack enough independent evidence for deployment.

A prospectively locked real-outcome experiment makes the distinction concrete. On the randomized Criteo benchmark, a prespecified quantitative model for **exact reference-best recovery plus certification** failed on held-out depths: adding a raw-cardinality localization term worsened prediction relative to certification alone. At $n = 2^{20}$ and depth 8, however, the selected bin certified in 94% of runs while exact reference-best recovery was only 40%. The empirical best and runner-up differed by only 0.35 standard errors and by 2.9×10^{-5} in population-value units. The failure therefore decomposes into a localization-identity problem rather than a failure of candidate certification. We preserve the locked negative endpoint and use the decomposition to motivate a sharper distinction among certifiable resolution, exact localization, and value-tolerant localization. A zero-fit prediction of certifiable resolution computed from frozen reference geometry, with no fitted constants, then matches the observed resolution at 8 of 11 sample sizes (10 of 11 when conditioned on the upstream-selected bin).

The theory explains why resolution is an information-allocation problem. In a symmetric unknown-region experiment, reliable localization incurs a logarithmic search cost. In a richer passive Gaussian model, exact localization is governed by the ordered spectrum of near-best gaps. Independent action certification has a separate information requirement. Comparing these costs with simultaneous Bonferroni certification yields a **search-versus-multiplicity phase boundary**: exhaustive certification is favored over an intermediate range of difficult spectra; on very hard near-tie plateaus

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.

both architectures become selection-limited and their difference vanishes, whereas the advantage of select-then-certify grows on sufficiently separated spectra. A value-tolerant extension further shows that exact identity can be statistically difficult while the decision consequence of choosing a near tie is negligible.

Evidence resolution is adjacent to uncertainty-aware policy learning, abstention, safe policy improvement, valid subgroup discovery, and best-arm identification, but the statistical object is different. Policy-risk methods ask which policy is valuable or safe; subgroup methods ask whether a discovered group has an effect above a threshold; best-arm identification asks which alternative is optimal under a sampling model. We study the **granularity of treatment action** supportable under passive trial evidence, separate representation loss from downstream inference, and require independent deployment certification after upstream adaptation.

The final **HonestER** method is deliberately simple. Complementary representations are learned on training data. Independent validation data rank one entry per representation and freeze deployment scores. A fixed error budget is allocated across representation families. Each path is then certified independently on untouched data using entry-then-shell fixed-sequence backoff. A failure in one family does not eliminate alternatives in another.

Our contributions are: **(i)** evidence resolution as a target distinct from exact candidate identity and policy value; **(ii)** a worst-case resolution frontier, a matching exact-localization theorem in a passive Gaussian oracle, and a search-versus-multiplicity crossover against exhaustive certification; **(iii)** a value-tolerant localization layer showing when exact near-tie resolution is unnecessary; **(iv)** HonestER, a representation-aware certification architecture with conditional portfolio-level error control and empirical evidence for representation diversity; and **(v)** on randomized real outcomes, a no-fit prediction of certifiable resolution from frozen reference geometry, together with a disclosed failure mode of the Welch certificate for rare binary outcomes.

2 Relation to policy uncertainty, subgroup inference, and identification

Confidence-aware policy learning optimizes welfare under estimation risk (Chernozhukov et al., 2025); safe policy learning constrains deterioration relative to a baseline (Ben-Michael et al., 2025); abstaining policy learning permits deferral (Sawarni et al., 2026); and valid subgroup-selection methods control false discoveries after flexible subgroup refinement (Cheng et al., 2025; Muller et al., 2025). These methods answer whether a policy or subgroup is valuable, safe, or inferentially valid. Evidence resolution instead asks whether the evidence could defend a **finer action distinction**.

Heterogeneous-effect and policy-learning methods supply candidate representations and policies (Athey & Imbens, 2016; Wager & Athey, 2018; Athey et al., 2019; Athey & Wager, 2021), while model selection for treatment choice already formalizes holdout selection among policy classes (Mbakop & Tabord-Meehan, 2021). Thus sample splitting itself is not novel. The localization component is also close to best-arm identification. Fixed-budget work studies the finite-sample search cost of identifying a unique best arm, including Successive Rejects and lower bounds based on hard alternatives/permutation constructions (Audibert, Bubeck, & Munos, 2010; Carpentier & Locatelli, 2016). Fixed-confidence adaptive identification provides complementary information-theoretic characterizations (Kaufmann, Cappé, & Garivier, 2016; Garivier & Kaufmann, 2016), and structured settings can materially reduce the effective competitor set (Poiani et al., 2025). We therefore do **not** claim best-arm complexity itself as novel. Our contribution is the coupling of passive localization with a separate action-certification requirement and the resolution target induced by that coupling.

The post hoc reference-geometry audit is also related to ranking-and-selection and **multiple comparisons with the best** (MCB). Gupta's subset-selection framework and Hsu's MCB intervals exploit the fact that only comparisons against the best are scientifically relevant, avoiding an unnecessary all-pairs multiplicity penalty (Gupta, 1965; Hsu, 1984). We use that framework to report plausible reference winners rather than treating an empirically selected best bin as known ground truth. Combining heterogeneous learners is likewise related to Super Learner (van der Laan, Polley, & Hubbard, 2007); HonestER's novelty is not ensemble prediction, but independent certification and failure isolation after upstream representation learning.

HonestER therefore does not claim a new generic holdout, sequential test, or abstention principle. The contribution is the **resolution target and decomposition**: representation, localization, and certification are separately measurable constraints on how much personalization can be defended.

3 EVIDENCE RESOLUTION AND ITS INFORMATION LIMITS

Let $X \sim P_X$, and let $\mathcal{H}$ be a rooted family of treatment policies. Node v changes its parent's action on

$$C_v = \{x : \pi_v(x) \neq \pi_{\mathrm{pa}(v)}(x)\}.$$

For favorable-oriented utility $U(a)$, define parent-relative gain

$$g_v(P) = \mathbb{E}[U(\pi_v(X)) - U(\pi_{\mathrm{pa}(v)}(X)) \mid X \in C_v].$$

An edge is locally valid when its clinically relevant gain exceeds the prespecified margin and deployably valid when every required action-changing ancestor is valid. For nested treated regions S_v, define retained-population resolution

$$r_P(v) = -\log_2 P_X(S_v).$$

Policy value remains primary; resolution is a granularity descriptor.

3.1 REPRESENTATION AND DOWNSTREAM LOSSES

Let V^* be unrestricted oracle value, V_H the best valid value expressible by the representation, V_D the best valid value among proposed candidates, and V_F the best valid value surviving certification. With supported deployed value V_{sup},

$$V^* - V_{\mathrm{sup}} = (V^* - V_H) + (V_H - V_D) + (V_D - V_F) + (V_F - V_{\mathrm{sup}}),$$

separating representation, proposal/localization, certification, and post-certification deployment loss. These terms are simulation diagnostics rather than separately identifiable clinical estimands.

3.2 SYMMETRIC SEARCH-SUPPORT FRONTIER

Consider M disjoint equal-mass regions of probability p. At the global boundary every region is exactly at the clinical margin; under alternative j, only region j has standardized excess d. Uniform localization and false-personalization control require

$$npd^2 = \Theta(\log M)$$

up to fixed-error constants. For a balanced hierarchy $M = 2^r$, $p = 2^{-r}$,

$$r = \log_2(nd^2) - \log_2 \log(nd^2) + O(1).$$

The $\log \log$ correction is the worst-case price of having to find the fine action region instead of being told where it is. The derivation is in Appendix A.

3.3 EXACT LOCALIZATION AND THE SEARCH-VERSUS-MULTIPLICITY BOUNDARY

In the balanced passive Gaussian model, let the unique best candidate have gaps $0 < \Delta_{(1)} \leq \cdots \leq \Delta_{(M-1)}$ from its competitors. For $0 < \delta \leq 1/4$, let $m^\star_{\mathrm{sel}}(\mu, \delta)$ denote the smallest balanced per-candidate validation sample size for which a selector is δ-correct uniformly over the stated local permutation class, and define

$$H_{\mathrm{sel}}(\mu, \delta) = \max\left\{ \frac{\log(1/\delta)}{\Delta_{(1)}^2}, \max_{k \geq 2} \frac{\log k}{\Delta_{(k)}^2} \right\}.$$

Theorem 1 (instance-dependent exact localization). On this local permutation class,

$$\frac{1}{4} H_{\mathrm{sel}} \leq m_{\mathrm{sel}}^{\star} \leq 14 H_{\mathrm{sel}} + 1.$$

The proof is in Appendix B, Theorem B.1.

Thus one near tie incurs a nearest-competitor cost, while a broad near-tie plateau restores a logarithmic multiplicity penalty.

Now compare honest sample splitting with exhaustive reuse. Certifying one fixed Gaussian candidate at level α and power $1 - \beta$ requires

$$B_1 = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{d_c^2},$$

whereas Bonferroni certification of all M fixed claims requires

$$B_M = \frac{(z_{1-\alpha/M} + z_{1-\beta})^2}{d_c^2}.$$

The exhaustive arm must also localize which certified candidate to deploy, but it can reuse the full sample. Hence the idealized per-candidate information costs are

$$C_{\mathrm{hon}} = m_{\mathrm{sel}}^{\star} + B_1, \qquad C_{\mathrm{exh}} = \max\{m_{\mathrm{sel}}^{\star}, B_M\}.$$

Proposition 1 (search-versus-multiplicity crossover).

$$C_{\mathrm{hon}} < C_{\mathrm{exh}} \quad \Longleftrightarrow \quad m_{\mathrm{sel}}^{\star} < B_M - B_1.$$

The derivation is in Appendix B, Proposition B.1.

When $m_{\mathrm{sel}}^{\star} \geq B_M$, exhaustive reuse is cheaper by the fixed certification increment B_1, although the ratio tends to one as selection becomes dominant. Theorem 1 yields sufficient constant-factor regions around this boundary; the unresolved band reflects the theorem's unidentified constants rather than a claim of universal dominance.

We checked this crossover numerically using exact one-dimensional quadrature in the same balanced Gaussian model. The honest split is optimized per instance, which favors HonestER and makes regions where exhaustive still wins conservative. Exhaustive certification is cheaper at intermediate gaps, by up to about 1.6–$1.8\times$ in required sample size. On very hard near-tie plateaus the ratio returns to approximately one because both arms are selection-limited. Conversely, on separated spectra select-then-certify can require only 42–62% of the exhaustive sample. The spectrum dependence predicted by H_{sel} is also visible: from $M = 8$ to 1024, the plateau crossover moves only from 1.61 to $1.22\, d_c$, whereas the single-near-tie crossover moves from 1.05 to $0.40\, d_c$. The order-level theory gets the direction and contrast right but not the full magnitude, as expected from universal constant-factor bounds.

3.4 Value-tolerant localization

Exact identity can be stronger than the decision problem requires. For tolerance $\epsilon \geq 0$, define $\mathcal{G}_\epsilon = \{j : \mu_\star - \mu_j \leq \epsilon\}$ and retain only unacceptable candidates with gaps $\Delta_j > \epsilon$. Ordering those gaps as $\Gamma_{(k)}^{\epsilon}$, define

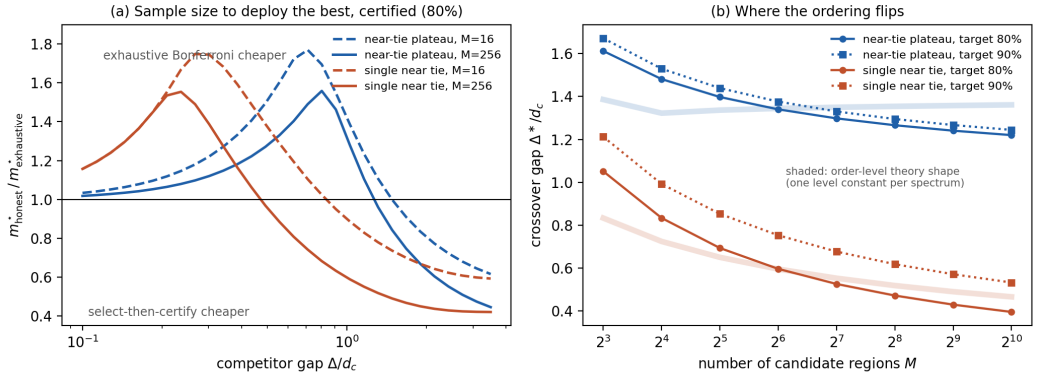

Figure 1: Search-versus-multiplicity phase boundary in the balanced passive Gaussian model, computed by quadrature. The honest split is optimized per instance. Left: ratio of per-candidate budgets needed for 80% successful deployment. Right: crossover gap versus M for 80% and 90% targets. The theory captures the qualitative dependence on the competitor spectrum but not the exact finite-budget constants.

$$H_\epsilon^{\mathrm{pass}} = \max\left\{\frac{\log(1/\delta)}{(\Gamma^\epsilon_{(1)})^2}, \max_{k\geq 2}\frac{\log k}{(\Gamma^\epsilon_{(k)})^2}\right\}.$$

Empirical-mean selection satisfies

$$\Pr(\mu_\star - \mu_{\widehat{J}} > \epsilon) \leq \sum_{j:\Delta_j>\epsilon} e^{-m\Delta_j^2/4},$$

so $m \geq 14H_\epsilon^{\mathrm{pass}} + 1$ is sufficient. If $\epsilon < \Delta_{(1)}$, this reduces to exact localization and the lower bound above matches up to constants. When multiple candidates are already acceptable, we do not claim a matching general converse: the original single-swap lower-bound class can become degenerate.

For equal-mass disjoint regions of mass p, a local effect gap Δ_j corresponds to population-value gap $p\Delta_j$. Thus with population-value tolerance η, use $\epsilon = \eta/p$. At fine resolution $p = 1/M$, exact identity can become harder with M even while the value loss of confusing two nearby regions shrinks as Δ/M.

4 HONESTER: PRESERVE ALTERNATIVES, CERTIFY INDEPENDENTLY

Split data into $D_{\mathrm{tr}}, D_{\mathrm{val}}, D_{\mathrm{cert}}$. On training data, fit a small prespecified set of complementary representations. On validation data, the frozen **expected certified value (ECV)** construction ranks one entry per family and supplies empirical-Bayes shrunk value-by-depth scores. ECV is an upstream planning surrogate, not a certificate and not a calibrated utility model. Appendix D gives two limited guarantees appropriate to this role. Propagating a joint validation confidence region through the deterministic ECV map yields an observable bound on **ECV-score selection ambiguity**, but no policy-value bridge is claimed. In a Gaussian projection model the certification noncentrality error induced by validation estimation has variance n_c/n_v. Writing a for the true standardized certification noncentrality and $\widehat{q} = \Phi(\widehat{a})$ for the plug-in pass probability,

$$\mathbb{E}\,\widehat{q} = \Phi\left(a/\sqrt{1 + n_c/n_v}\right)$$

rather than $\Phi(a)$. Thus projected pass probabilities need not calibrate near the decision boundary when n_c/n_v stays non-negligible; validity comes only from the untouched certification sample.

The final experiments use clinician/HGB/Ridge family weights

$$(0.25, 0.10, 0.65), \qquad \alpha_b = \alpha w_b,$$

fixed before final confirmation. Within each family, certify the entry first, then shell exclusions in fixed order, stopping that family at its first failure. This uses standard intersection–union / sequentially rejective ideas for structured multiple testing (Berger & Hsu, 1996; Bretz et al., 2009), while the hierarchical backoff perspective is related to tree-structured FWER control (Meinshausen, 2008). Independent downstream testing after upstream adaptation is also aligned with learn-then-test risk-control logic (Angelopoulos et al., 2021). Among certified outputs, deploy the one with the largest score frozen before certification; otherwise use the baseline.

HonestER algorithm (planning frozen before certification). **1** split into training/validation/certification; **2** fit complementary representations; **3** rank one entry per family by ECV and freeze depth/deployment scores; **4** allocate $\alpha_b = \alpha w_b$;
5 test entry then shells in fixed sequence; **6** stop/back off at the first family failure; **7** deploy the certified output with the highest frozen score; **8** otherwise deploy baseline.

Theorem 2 (conditional portfolio validity). Conditional on training and validation, suppose family b's true-null component tests are valid at $\alpha_b = \alpha w_b$, with $w_b \geq 0$ and $\sum_b w_b \leq 1$. If each family deploys only consecutive certified prefixes, then

$$\Pr(\text{invalid final deployment} \mid D_{\text{tr}}, D_{\text{val}}) \leq \sum_b \alpha_b \leq \alpha.$$

The proof is in Appendix C.2. No independence is required within or across families. The implementation uses Welch-Normal component tests, so this theorem is paired with empirical boundary calibration rather than an exact finite-sample claim.

5 Experiments

5.1 Real randomized outcomes: certifiable resolution separates from winner identity

We use the 13.98M-row randomized Criteo uplift benchmark (Diemert et al., 2018) with deterministic disjoint representation, reference, and experiment pools. A LightGBM T-learner (Ke et al., 2017) is frozen before experiment outcomes are used; the reference pool defines equal-mass bins. The Criteo certificate uses $\kappa = 0$: on every finite-SE row the recorded one-sided p-value is $1 - \Phi(\widehat{\tau}/\widehat{\mathrm{SE}})$. Full data roles, frozen geometry, and audit details are in Appendix F. This study tests the theory's certification-resolution proxy rather than a full HonestER path: each depth is selected and certified separately, so R_{80} summarizes depthwise certification rather than an HonestER traversal.

A **prospectively locked** scaling experiment used 100 repetitions for each $n = 2^{10}, \ldots, 2^{20}$ and depths 1–8. The locked endpoint required both exact recovery of the empirical reference-best bin and independent certification. A certification-only predictor was compared with a joint predictor that added the prespecified raw-cardinality localization term. The joint model failed: held-out log loss was 0.5244 versus 0.4008 for certification alone, and ordinary replicate-level Brier score was worse by 0.00330. We retain that negative result unchanged.

The failure decomposes sharply. At $n = 2^{20}$, exact reference-best recovery / certification / joint success were 96/100/96% at depth 6, 93/100/93% at depth 7, and **40/94/39%** at depth 8. Multiple-Comparisons-with-the-Best leaves 1, 3, and 8 plausible best bins at depths 6, 7, and 8, respectively. At depth 8 the best and runner-up are separated by only 0.35 standard errors and 2.9×10^{-5} population-value units. Thus the locked 40% exact-recovery rate is scored against an empirical argmax whose uniqueness is itself not established at the chosen family-wise confidence level.

We then ask the question directly suggested by the title: how much **certifiable resolution** does the frozen reference geometry predict? Using only the frozen best-bin effect/noise and the Normal power formula, with no fitted constants, we predict

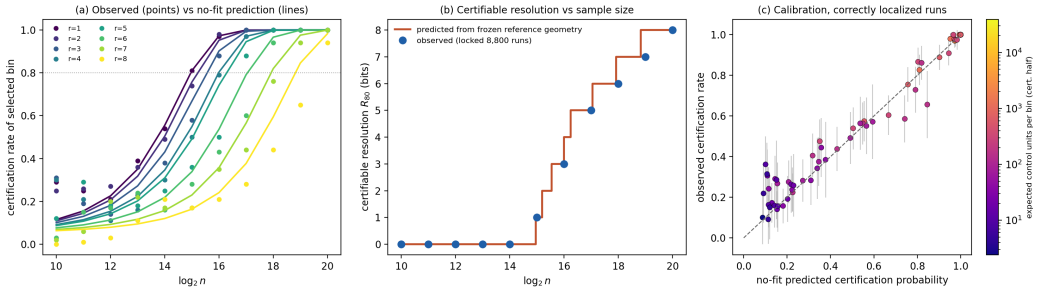

Figure 2: Criteo certification-resolution diagnostics, all post hoc and with no fitted calibration constants. (a) Observed selected-bin certification rates against $\log_2 n$ and the frozen-reference best-bin power curves. (b) R_{80}: observed versus the no-fit frozen-geometry prediction. (c) Calibration among correctly localized runs, colored by expected control units in the certification bin. The low-probability excess is confined to small-support cells.

$$P_{\mathrm{cert}}(n, r) = \Phi\left(d_{\mathrm{cert},r} \sqrt{\frac{n\, 2^{-r}}{2}} - z_{0.95} \right).$$

For a compact summary of the certification curves, define $R_{80}(n)$ as the finest depth whose certification frequency reaches 80%. The no-fit curve matches the observed R_{80} exactly at **8 of 11** sample sizes and is one bit too fine at the other three ($2^{16}, 2^{18}, 2^{19}$); every miss is optimistic, consistent with the prediction assuming selection of the reference-best bin. Observed resolution rises from 0 to 8 bits between 2^{15} and 2^{20}. It rises faster than one bit per doubling because the best-bin standardized certification gap increases with depth in this Criteo geometry; the symmetric log-minus-log-log result assumes a constant gap and should be read as a worst-case law, not as a quantitative prediction for this landscape.

The reference noise model itself is accurate where the resolution frontier lies: at $n \geq 131{,}072$, implied and observed median standard errors agree within 1.5%. More importantly, every nonzero observed R_{80} frontier point has at least about 307 expected control units per certification bin (and substantially more at the coarsest frontier points). Hence the title-level resolution curve is not driven by the rare-small-bin pathology described below.

Calibration is tight in adequately supported cells. Restricting to correctly localized runs, cells with at least 200 expected controls have row-weighted observed certification 0.928 versus 0.923 predicted, with weighted cell RMSE 0.027. For 50–200 expected controls the corresponding values are 0.472 versus 0.467 (RMSE 0.072). The excess is concentrated below 50 expected controls: 0.240 observed versus 0.179 predicted (RMSE 0.102).

A final robustness check removes the remaining reference-best selection assumption without touching certification outcomes. For each locked run, we condition on the **upstream-selected bin** and substitute that bin's frozen reference effect and noise into the same no-fit certification formula. This selection-conditioned curve matches the observed R_{80} at **10 of 11** sample sizes, compared with 8 of 11 for the deterministic reference-best curve. All three one-bit optimistic errors of the reference-best curve disappear; the only remaining discrepancy is a one-bit pessimistic error at $n = 2^{15}$. Mean absolute resolution error falls from 0.273 to 0.091 bits, while cell-level RMSE/MAE improve from 0.0785/0.0530 to 0.0719/0.0452. Because this diagnostic uses the realized upstream selection decision, it is not an ex-ante n, r-only prediction; rather, it isolates the residual error attributable to finite-sample localization while remaining independent of certification outcomes and fitted calibration constants. The robustness curve is shown in the supplement.

Together these results distinguish **certifiable resolution**, **exact best-region localization**, and **deployed value**. The locked endpoint was a valid test of exact-reference-best recovery plus certification, but it was not a pure test of certifiable resolution. The prospective failure is preserved; the deterministic no-fit frontier shows that certification follows frozen reference geometry, and the

selection-conditioned robustness check attributes most of the remaining frontier error to finite-sample localization rather than to the certification power model.

The earlier quota-only diagnostic is not used to infer population invalidity; registered Protocol B (Appendix G) instead prespecifies reference-supported valid/invalid/indeterminate labels to distinguish invalid proposals from valid-but-uncertified ones.

5.2 Representation can dominate certification

In the ACTG-175 real-covariate semi-synthetic cross-cut regime, using baseline covariates from the randomized ACTG 175 trial (Hammer et al., 1996; UCI Machine Learning Repository, 2023), unrestricted oracle value is 0.04363 while the best fixed hierarchy reaches only 0.01048. Representation loss is therefore 0.03315, or **76% of oracle value**, before downstream certification acts. A prospectively locked learned-path pilot raises the best valid representation value to 0.01880, reducing representation loss by 25.1%. Different representations dominate different geometries, motivating a portfolio rather than one universal hierarchy.

5.3 Portfolio benefit and development history

On a fresh 900-case prospective set,

$$V_{\mathrm{FixedECV}} - V_{\mathrm{SinglePath}} = 0.001187$$

with pointwise paired 95% CI $[0.000443, 0.001932]$. This is a **secondary/descriptive prospective comparison**, supporting representation diversity and failure isolation rather than serving as a primary confirmatory efficacy claim.

The remaining development history is deliberately secondary. Adaptive ECV family weighting did not demonstrate an advantage over the fixed 25/10/65 allocation. A prospectively locked stacking extension failed its value gate. Tested retrospective global schedulers lost value through cross-family blocking. Finally, an ECV-free simplification changed both entry ranking and deployment scoring and failed its locked noninferiority criterion: ValueFixed-FixedECV was -0.000256, with one-sided 95% lower bound -0.000553 versus margin -0.0005; its two-sided interval included zero. We therefore retain the existing fixed-allocation ECV procedure without claiming ECV superiority. The full experimental progression and post hoc 2×2 attribution diagnostic are in Appendix E.

5.4 Boundary calibration and rare binary outcomes

The certification theorem assumes valid component tests, while the implementation uses Welch standard errors with Normal critical values. A fresh 210,000-replicate exact-boundary study under Normal and heteroskedastic-Normal outcomes gives pooled entry-null false certification 0.05136 (95% CI [0.05012,0.05262]); the worst exact-boundary cell is 0.0564. We therefore describe the implemented Normal-critical certificate as approximately calibrated **only in those audited continuous-outcome regimes**.

Criteo exposes an additional limitation for rare binary outcomes at small support. The randomized design assigns about 15% of observations to control, while `visit` is the rare binary outcome. When a certification arm contains no events, its empirical Bernoulli variance is zero; the Welch standard error can collapse and the Normal test becomes anticonservative. At $n = 1024, r = 1$, 26 runs with standard error below 0.6 times the cell median certify in 96% of cases versus 5% for the remaining runs. Across cells with fewer than 50 expected controls per bin (3,600 of 8,800 rows), low-SE runs account for 47% of all certifications; the phenomenon is essentially absent once expected control support exceeds 200.

This pathology does **not** explain the large-n, fine-depth conclusions or the observed R_{80} frontier, whose nonzero frontier cells have hundreds of expected controls per bin. It does mean that the implemented Welch-Normal certificate is not approximately valid for rare binary outcomes at very small arm support. For such outcomes a deployment procedure should add a prespecified support/events requirement or use a rare-event-robust difference-of-proportions procedure, such as an Agresti–Caffo

add-two or Newcombe score-based interval/test. These are proposed fixes rather than retrospective changes to the locked Criteo endpoint.

6 Discussion and limitations

Evidence resolution reframes personalization as an information problem with distinct representation, localization, and certification bottlenecks. The Criteo results sharpen rather than simplify the theory: the symmetric $\log M$ frontier is a worst-case law, not a universal quantitative predictor of real candidate landscapes. Real fine-resolution difficulty can be dominated by a small set of near-tied alternatives. The instance-dependent localization complexity makes that distinction explicit while preserving the symmetric frontier as a corollary.

On Criteo, the no-fit frontier's 8/11 agreement (10/11 after conditioning on the upstream-selected candidate), together with unstable exact winner identity, empirically separates certifiable action granularity from localization identity and policy value.

Limitations remain. Theorem 1 matches the exact-localization complexity up to universal constants for $0 < \delta \leq 1/4$ over the stated balanced Gaussian local permutation class; it is **not** claimed pointwise instance-optimal outside that class, nor for correlated policy-value estimates, unequal masses, or nuisance-estimation error. The 25/10/65 portfolio allocation is empirically motivated and not claimed portable. The implemented Welch-Normal certificate is approximately calibrated in the audited continuous-outcome regimes but is anticonservative for rare binary outcomes at very small arm support; deployment requires a support guard or rare-event-robust binary certificate. Criteo provides randomized real outcomes but is an advertising incrementality benchmark, not clinical external validation. Historical clinical analyses mainly demonstrate restraint/backoff rather than positive external personalized benefit.

In summary, the paper asks a narrower question than generic policy learning but one that becomes increasingly important as flexible learners improve: **how much personalization can finite evidence actually defend?** The answer depends on what the representation can express, what competing refinements the data can distinguish, and what untouched evidence can certify.

Reproducibility statement

Prospective protocols, data-role definitions, analysis plans, and frozen experiment outputs are provided in the supplement. The Criteo scaling-law failure is preserved as a failed prospective test; the later instance-dependent theory is labeled as a theory refinement rather than a confirmation of that test. Appendix E labels the empirical analyses as prospective, conditional, or post hoc. Two follow-up analyses were registered before submission; protocol texts and SHA-256 hashes are in Appendix G. Protocol A, an independent-dataset test of the zero-fit predictor on Lenta, was run only through its prespecified reference-only informativeness gate and stopped as uninformative (none of its 15 prespecified sample sizes met the informativeness criterion), so no independent confirmation is claimed. Protocol B, a reference-labeled audit of standard learners on Criteo, remains pending.

Ethics statement

Clinical analyses use de-identified randomized-trial data for secondary methodological evaluation. Criteo is used as a randomized real-outcome methodological benchmark rather than as a clinical analogue. Statistical certification is not presented as a treatment recommendation or clinical validation.

AI use statement

Generative AI systems were used for brainstorming, code scaffolding and review, experiment planning, manuscript organization, and language editing. The authors remain responsible for verification.

REFERENCES

Angelopoulos, A. N., Bates, S., Candes, E. J., Jordan, M. I., & Lei, L. (2021). Learn then Test: Calibrating predictive algorithms to achieve risk control. arXiv:2110.01052.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *PNAS*, 113(27), 7353-7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2), 1148-1178.

Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1), 133-161.

Audibert, J.-Y., Bubeck, S., & Munos, R. (2010). Best arm identification in multi-armed bandits. *COLT 2010*, 41-53.

Ben-Michael, E., Greiner, D. J., Imai, K., & Jiang, Z. (2025). Safe policy learning through extrapolation: Application to pre-trial risk assessment. *JASA*, 120(551), 1386-1399.

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4), 283-319.

Bretz, F., Maurer, W., Brannath, W., & Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4), 586-604.

Carpentier, A., & Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. *COLT 2016*, PMLR 49, 590-604.

Cheng, N., Spector, A., & Janson, L. (2025). Chiseling: Powerful and valid subgroup selection via interactive machine learning. arXiv:2509.19490.

Chernozhukov, V., Lee, S., Rosen, A. M., & Sun, L. (2025). Policy learning with confidence. cemmap working paper CWP15/25; arXiv:2502.10653.

Diemert, E., Betlei, A., Renaudin, C., & Amini, M.-R. (2018). A Large Scale Benchmark for Uplift Modeling. *Proceedings of AdKDD & TargetAd (ADKDD'18).*

Garivier, A., & Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. *COLT 2016*, PMLR 49.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, 7(2), 225-245.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., & Merigan, T. C. (1996). 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.

Hsu, J. C. (1984). Constrained two-sided simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics*, 12(3), 1136-1144.

Kaufmann, E., Cappé, O., & Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *JMLR*, 17.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30*, 3146-3154.

Mbakop, E., & Tabord-Meehan, M. (2021). Model Selection for Treatment Choice: Penalized Welfare Maximization. *Econometrica*, 89(2), 825-848.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2), 265-278.

Muller, M. M., Reeve, H. W. J., Cannings, T. I., & Samworth, R. J. (2025). Isotonic subgroup selection. *JRSS-B*, 87(1), 132-156.

Poiani, R., Jourdan, M., Kaufmann, E., & Degenne, R. (2025). Best-Arm Identification in Unimodal Bandits. *AISTATS 2025*, PMLR 258, 2233-2241.

Sawarni, A., Jin, J., Whitehouse, J., & Syrgkanis, V. (2026). Policy Learning with Abstention. *AISTATS 2026*, PMLR 300, 4465-4473.

UCI Machine Learning Repository. (2023). *AIDS Clinical Trials Group Study 175* [Dataset]. https://doi.org/10.24432/C5ZG8F.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), Article 25.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *JASA*, 113(523), 1228-1242.

SUPPLEMENTARY MATERIAL

A COMPLETE ORACLE THEORY PACKAGE

The theory has three layers:

1. a population target: how fine the truly action-changing policy is under a fixed hierarchy;
2. a finite-sample minimax target: how fine a policy can be resolved with prescribed false-personalization and power guarantees;
3. a reported random output: the finest action-changing refinement actually certified from data.

The distinction is essential. Biological heterogeneity can exist beyond what a finite trial can resolve, and a learner can fit more structure than it can defend.

A.0.1 DECISION HIERARCHY AND POPULATION PERSONALIZATION RESOLUTION

Let `X ~ P_X` denote patient covariates, let `A` be a finite action set, and let `H` be a rooted hierarchy of measurable patient regions. Every node `v` has region `S_v`, proposed action `a_v`, and parent `pa(v)` with `S_v subset S_pa(v)`. Only edges for which `a_v != a_pa(v)` are action-changing and therefore count as personalization.

For a scalar favorable-oriented outcome, define the parent-relative gain

```
g_v(P) = E_P[Y(a_v) - Y(a_pa(v)) | X in S_v].
```

Let `kappa_v` be a prespecified minimum clinically meaningful gain. The local slack is

```
s_v(P) = g_v(P) - kappa_v.
```

An action-changing edge is clinically valid when `s_v(P) > 0`.

For multiple favorable-oriented clinical constraints `ell = 1,...,L`, define

```
s_v,ell(P) = g_v,ell(P) - kappa_v,ell,
```

where harm endpoints are sign-oriented so larger slack is always better. Edge `v` is locally valid exactly when every `s_v,ell(P) > 0`.

An edge is **deployably valid** when it is locally valid and every action-changing ancestor edge is also locally valid. This ancestry requirement formalizes backoff: a fine action cannot be deployed through an unsupported coarser action change.

Define node rarity/resolution

```
r_P(v) = -log2 P_X(S_v).
```

The **population decision resolution** is

```
R(P;H) = max { r_P(v) : v is a deployably valid action-changing edge },
```

with `R(P;H)=0` if no action-changing edge is deployably valid.

This is a property of the distribution and the prespecified hierarchy. It is not an estimator and not an information-theoretic quantity. It can be large even when the available sample is far too small to establish it.

A.0.2 REPORTED CERTIFIED RESOLUTION, OVER-RESOLUTION, AND UNDER-RESOLUTION

A data-dependent procedure returns a set `C_hat` of certified action-changing edges. After ancestry gating, let `D_hat` be the deployed set. The reported resolution is

```
R_hat = max { r_P(v) : v in D_hat },
```

with `R_hat=0` when `D_hat` is empty. In practice `P_X(S_v)` can be defined from a prespecified target population or estimated on an independent/pooled design sample; the inferential guarantee below concerns edge validity rather than error in the descriptive rarity estimate.

Define

```
over-resolution  = max(R_hat - R(P;H), 0),
under-resolution = max(R(P;H) - R_hat, 0).
```

Over-resolution is inferentially dangerous: it means the procedure deployed structure finer than any truly valid action change in the hierarchy. Under-resolution can be the correct consequence of finite evidence.

Proposition A.1 (strong edge-FWER implies over-resolution control) Suppose the certification procedure satisfies

```
P_P(there exists a certified clinically invalid edge) <= alpha
```

for every `P` in the model class. Ancestry gating can only remove certifications. Therefore

```
P_P(R_hat > R(P;H)) <= alpha.
```

Proof. If `R_hat > R(P;H)`, some deployed edge has resolution above the finest deployably valid action-changing edge. That deployed edge cannot be clinically valid under the definition of `R(P;H)`. Hence over-resolution is contained in the event of at least one false edge certification. QED.

The proposition deliberately does not bound under-resolution. Power and discovery coverage determine how much true personalization is recovered.

A.0.3 MINIMAX EVIDENCE RESOLUTION

Population decision resolution and finite-sample evidence resolution are different objects. Fix a statistical experiment class `P_r` in which the finest clinically valid action change occurs at resolution/depth `r` but its location is unknown.

For sample size `n`, false-certification target `alpha`, and miss probability `beta`, define the **minimax evidence resolution**

```
R*_{n,alpha,beta} = sup { r : there exists a procedure with
    strong false-certification probability <= alpha
    and true refinement certification probability >= 1-beta
    uniformly over P_r }.
```

This is the statistical resolution frontier. The actual procedure reports `R_hat`; `R*` characterizes what any procedure can hope to guarantee uniformly in the chosen experiment class.

A.0.4 ORACLE GAUSSIAN UNKNOWN-REGION EXPERIMENT

The core lower/upper bounds use an oracle model easier than causal learning. Candidate regions and the score law are known, so impossibility here applies a fortiori when nuisance functions or candidate structure must be estimated.

Let S_1, ..., S_J be **disjoint** candidate regions with common probability p, where J p <= 1. Under the global boundary distribution P_0,

```
G | X ~ Normal(kappa, sigma^2).
```

Under P_j, only region S_j has a clinically meaningful excess

```
G | X in S_j ~ Normal(kappa + Delta, sigma^2),
G | X not in S_j ~ Normal(kappa, sigma^2).
```

Write d = Delta/sigma and q = exp(d^2)-1.

A procedure returns a set J_hat of certified regions.

We distinguish two error notions:

- **global-null FWER:** P_0(J_hat != empty) <= alpha;
- **strong FWER on the one-signal family:** under each P_j, the probability of certifying any k != j is at most alpha, in addition to global-null control.

Uniform power requires

```
inf_j P_j(j in J_hat) >= 1-beta.
```

Strong FWER is the closer analogue of the clinical deployment guarantee. The older chi-square theorem needs only the weaker global-null condition and therefore remains a necessary condition under strong FWER.

A.0.5 Detection lower bound: exact null-versus-mixture condition

Theorem A.1 (exact unknown-region mixture lower bound) If global-null FWER is at most alpha, uniform power is at least 1-beta, and alpha+beta<1, then necessarily

```
([1 + p {exp(d^2)-1}]^n - 1) / J
    >= 4 (1-alpha-beta)^2.                          (T1)
```

Equivalently,

```
n log(1 + p {exp(d^2)-1})
    >= log(1 + 4 J (1-alpha-beta)^2).               (T1')
```

Proof. Let L_j=dP_j/dP_0 for the full sample and let Q=J^{-1} sum_j P_j. For one observation,

```
E_0 ell_j^2 = 1 + p {exp(d^2)-1}.
```

Because the regions are disjoint, E_0 ell_j ell_k=1 for j != k. Independence across n observations yields

```
E_0 L_j^2 = [1+p{exp(d^2)-1}]^n,
E_0 L_j L_k = 1, j != k.
```

Therefore

```
chi^2(Q||P_0)
 = J^{-1}([1+p{exp(d^2)-1}]^n - 1).
```

For A={J_hat != empty}, P_0(A)<=alpha while uniform power implies Q(A)>=1-beta. Hence TV(Q,P_0)>=1-alpha-beta. Since TV <= 0.5 sqrt(chi^2), (T1) follows. QED.

Interpretation. This is the cost of establishing that some unknown clinically meaningful refinement exists. It does not by itself force correct localization when candidate regions overlap or when the procedure is allowed to certify many regions under alternatives.

A.0.6 LOCALIZATION LOWER BOUND UNDER STRONG FALSE-PERSONALIZATION CONTROL

The clinical procedure must not react to a detected signal by declaring many unsupported subgroups. Strong FWER adds a distinct localization cost.

Theorem A.2 (Fano localization lower bound for disjoint candidates) Assume the disjoint equal-mass experiment above and strong FWER on `P_0,P_1,...,P_J`. If uniform true-region power is at least `1-beta`, then for `J>=2`,

```
(1 - 1/J) n p d^2
  >= [(1-alpha-beta) log J - log 2]_+.          (T2)
```

Proof. Under `P_j`, with probability at least `1-alpha-beta`, the procedure includes `j` and includes no false region. Therefore

```
P_j(J_hat = {j}) >= 1-alpha-beta.
```

Construct a decoder of the unknown index that outputs the unique certified index when `J_hat` is a singleton and outputs an arbitrary index otherwise. With a uniform prior on the true index, its average error is at most `alpha+beta`.

Fano's inequality gives

```
I(J;D) >= (1-alpha-beta) log J - log 2.
```

By convexity of KL,

```
I(J;D) <= J^{-2} sum_{j,k} KL(P_j^n || P_k^n).
```

For `j != k`, the alternatives differ by `Delta` on `S_j` and by `-Delta` on `S_k`. Because the two regions are disjoint,

```
KL(P_j || P_k) = p d^2
```

per observation, while the diagonal KL is zero. Averaging gives

```
I(J;D) <= (1-1/J) n p d^2.
```

Combining the bounds proves (T2). QED.

Interpretation. Detection and localization are distinct. The mixture bound asks whether the data can distinguish "no valid refinement" from "some valid refinement." The Fano bound asks whether a strongly valid procedure can determine which refinement is the one it may safely deploy.

A.0.7 LOCAL CLINICAL-EVIDENCE LOWER BOUND

Even if an oracle reveals the correct region, its action change still needs enough local evidence to cross the clinical margin.

Theorem A.3 (fixed-region local-support necessity) Fix region `S_j` with probability `p`. Compare a valid alternative with standardized excess `d` to an otherwise identical boundary distribution in which region `j` is exactly at the clinical margin. If certification of `j` has probability at least `1-beta` under the valid alternative and at most `alpha` under the boundary distribution, then

```
n p d^2 / 2 >= kl_Ber(1-beta || alpha).          (T3)
```

Proof. The full-sample KL divergence between the two Gaussian experiments is `n p d^2/2`. Map the data to the binary event `{j is certified}` and apply the data-processing inequality for KL. QED.

This is the pure rare-patient support cost. There is no `log J` term because the location is fixed.

A.0.8 EXACT CONSTRUCTIVE CERTIFICATE AND EXACT ORACLE POWER

Theorem A.4 (Bonferroni strong-FWER certificate) For each nonempty candidate region define

```
Z_j = sqrt(N_j) (mean_j(G)-kappa)/sigma,
```

where `N_j` is its observed support. Certify `j` when

```
Z_j > c_J,    c_J = Phi^{-1}(1-alpha/J).          (T4)
```

Then the procedure controls strong FWER at `alpha` over `P_0,P_1,...,P_J`.

Under the true alternative `P_j`, its **exact unconditional power** is

```
Power_B(n,p,d,J,alpha)
 = E_{N ~ Binomial(n,p)} [ Phi(d sqrt(N) - c_J) ]. (T5)
```

Empty-region support contributes zero automatically because `Phi(-c_J)` in (T5) must be replaced by zero when `N=0`; equivalently write

```
sum_{k=1}^n BinomPMF(k;n,p) Phi(d sqrt(k)-c_J).
```

Proof. Every invalid region has a standard-normal `Z_j` conditional on its nonzero support. Hence its marginal false-certification probability is at most `alpha/J`; a union bound over the invalid subset gives strong FWER at most `alpha`. Under `P_j`, conditional on `N_j=k>=1`, `Z_j ~ N(d sqrt(k),1)`, giving conditional power `Phi(d sqrt(k)-c_J)`. Average over the binomial support distribution. QED.

Equation (T5) is an exact achievable oracle curve, not a Monte Carlo approximation. It sharpens the older Chernoff-plus-Gaussian sufficient bound while keeping the same procedure.

A conservative closed-form sufficient condition follows by spending `beta/2` on support fluctuation and `beta/2` on score noise. For any `epsilon in (0,1)`, power is at least `1-beta` if

```
n p >= 2 epsilon^{-2} log(2/beta),

d sqrt((1-epsilon) n p)
  >= c_J + z_{1-beta/2}.                         (T6)
```

A.0.9 COMPLETE FINITE-SAMPLE RESOLUTION BRACKET

For the equal-disjoint oracle class define:

- `r_constructive`: the finest candidate rarity/depth at which exact Bonferroni power (T5) is at least `1-beta`;
- `r_detection`: the finest depth not ruled out by (T1);
- `r_localization`: the finest depth not ruled out by (T2);
- `r_local`: the finest depth not ruled out by (T3).

The **combined information ceiling** is the deepest level satisfying all three necessary conditions:

```
r_info = min-ceiling(T1, T2, T3).
```

Then

```
r_constructive <= R*_{n,alpha,beta} <= r_info.     (T7)
```

The left inequality follows because Theorem A.4 supplies an explicit strongly valid procedure. The right inequality follows because every strongly valid procedure must satisfy Theorems A.1–A.3.

This bracket is finite-sample and auditable. It does not claim sharp minimax constants.

A.0.10 RATE THEOREM AND THE PERSONALIZATION PHASE BOUNDARY

Theorem A.5 (search-support minimax rate) Fix `alpha,beta` with `alpha+beta<1`, and suppose the standardized gap is bounded, `0<d<=d_max`. For the equal-disjoint `J`-candidate experiment, any strongly valid procedure with uniform power `1-beta` requires

```
n p d^2 >= c_1 log J - C_1,                          (T8)
```

for constants depending only on `alpha,beta,d_max`. The Bonferroni procedure attains target power whenever

```
n p >= C_2
and
n p d^2 >= C_3 [log J + log(1/alpha)+log(1/beta)], (T9)
```

with universal/target-dependent constants. Therefore, for fixed `alpha,beta` and bounded nondegenerate `d`, the minimax local information requirement is

```
n p d^2 = Theta(log J).                               (T10)
```

Proof sketch. The lower rate follows directly from Theorem A.2, and also from Theorem A.1 after using `exp(d^2)-1 <= exp(d_max^2)d^2`. The upper rate follows from (T6) and `c_J <= sqrt(2 log(J/alpha))`, plus the constant occupancy requirement. QED.

This separates two quantities that are conflated in a balanced tree:

```
local evidence    = n p d^2,
search complexity = log J.
```

Personalization is resolvable only when local evidence dominates the logarithmic search cost.

Corollary A.5.1 (balanced hierarchy log-minus-log-log law) For a balanced depth-`r` partition, `J=2^r` and `p=2^{-r}`. Equation (T10) becomes

```
n d^2 2^{-r} = Theta(r).
```

Equivalently,

```
r = W(Theta(n d^2))/log 2
  = log2(n d^2) - log2 log(n d^2) + O(1).          (T11)
```

Thus the leading `log2(n d^2)` support law loses a `log2 log(n d^2)` search correction because the number of possible signal locations grows exponentially with resolution.

A.0.11 MULTIPLE CLINICAL ENDPOINTS

Let the oracle favorable-oriented score vector have covariance `Sigma`, and let alternative `j` shift its mean by `delta_j` inside candidate region `S_j`.

For fixed possibly unequal/overlapping candidates define

```
p_jk = P(X in S_j intersection S_k),
a_jk = delta_j' Sigma^{-1} delta_k,
K_jk = [1 + p_jk {exp(a_jk)-1}]^n - 1.
```

For mixture weights `w`,

```
chi^2(Q_w || P_0) = w' K w.
```

Therefore global-null detection requires

```
min_{w in simplex} w' K w >= 4(1-alpha-beta)^2.   (T12)
```

This is the exact finite-class likelihood-ratio geometry already verified in Phase 10. Overlap and endpoint covariance can pool information for **locating** a jointly favorable candidate.

That pooled information does not replace local evidence for every clinical constraint. For constraint `ell`, compare a valid candidate distribution with an otherwise identical fixed-location distribution in which only that constraint is moved to its boundary. If the corresponding mean-vector difference is `h_{j,ell}`, then

```
n p_j h_{j,ell}' Sigma^{-1} h_{j,ell} / 2
  >= kl_Ber(1-beta || alpha)                    (T13)
```

is necessary for that constraint-specific certification problem.

Hence multi-endpoint evidence resolution is the intersection of:

1. joint search/detection information;
2. localization information when distinct candidates must be distinguished;
3. local evidence for each clinical constraint.

A strong benefit endpoint can help identify where to look; it cannot by itself establish a safety margin.

A.0.12 HONEST DATA-ADAPTIVE DISCOVERY

The fixed-class minimax theory does not license unrestricted reuse of the certification sample for candidate generation.

Let `D` be arbitrary discovery data and let it produce a finite random candidate class `A(D)` together with proposed actions, endpoints, margins, and a multiplicity rule. Let `C` be independent certification data.

Theorem A.6 (conditional strong validity after arbitrary discovery) If for every realized discovery output,

```
P_C(any false certificate in A(D) | D) <= alpha,
```

then

```
P(any false certificate) <= alpha,
P(R_hat > R(P;H)) <= alpha.
```

Proof. Take expectations over `D`; Proposition A.1 gives the second statement. QED.

Discovery can be arbitrarily complex without spending the final Type-I error **only because the certification sample is untouched**. Search performed in D can reduce the multiplicity faced in C, but it consumes sample size and can miss the true candidate. The fixed-class minimax lower bound applies to the full statistical problem before such information is available; the split-sample procedure is a valid construction, not claimed constant-optimal.

A.0.13 DISCOVERY COVERAGE AND EXACT POLICY RECOVERY

Let v* be the true finest deployably valid edge. Define the discovery event

```
E_D = {v* belongs to A(D) with the correct proposed action}.
```

Let E_C be the event that all required local/ancestor certificates for v* pass on certification data. Then

```
P(exactly recover the target path) <= P(E_D),
P(exactly recover the target path) >= 1 - P(E_D^c) - P(E_C^c)
```

when no competing false edge is certified. Strong FWER controls the last failure mode; it does not guarantee discovery coverage. This decomposition explains why an oracle known-region benchmark can substantially outperform a practical learner even when certification itself is calibrated.

A.0.14 WHAT THE THEORY DOES AND DOES NOT CLAIM

Supported claims

- Decision resolution is distinct from CATE estimation error and policy regret.
- Strong simultaneous certification controls false over-resolution.
- Unknown personalization has three separable information costs: existence/detection, location, and local clinical-margin evidence.
- In the equal-disjoint oracle class, the minimax information scale is `n p d^2 = Theta(log J)`.
- In a balanced hierarchy, this yields the `log - log log` resolution law.
- Exact Bonferroni power is available in closed finite-sample form as a binomial mixture.
- Multi-endpoint search information can pool through the Mahalanobis geometry, while every clinical constraint retains a local evidence requirement.
- Honest discovery plus untouched certification preserves strong validity for a random finite candidate class.

Deliberately not claimed

- sharp minimax constants;
- minimax optimality for infinite or data-adaptive function classes without sample splitting;
- that failure to certify implies biological homogeneity;
- that the finest true subgroup is necessarily discoverable by a chosen learner;
- that policy regret alone captures false personalization;
- that one favorable endpoint can certify another clinical constraint;
- clinical deployment validity from historical trials alone.

A.0.15 ICLR-LEVEL FALSIFIABLE PREDICTIONS

The theory predicts:

1. At fixed candidate search size J, certification power rises with local information `n p d^2`.

2. At fixed local information, increasing `J` lowers achievable power because the search/localization cost grows like `log J`.
3. In balanced hierarchies where `p=1/J=2^{-r}`, the maximum 80%-power depth follows `log2(n d^2)-log2 log(n d^2)` more closely than `log2(n d^2)` alone.
4. Monte Carlo power of the stated Bonferroni rule should match the exact binomial-mixture power (T5).
5. No observed 80%-power boundary of a strongly FWER-valid procedure may lie beyond the combined lower-bound ceiling (T1)–(T3), up to Monte Carlo classification uncertainty.

These are the locked targets of `ICLR_PHASE_DIAGRAM_PLAN.md`.

B Matching Instance-Dependent Localization and Price-of-Honesty Proof

B.1 Overview

This appendix studies exact localization in a balanced passive Gaussian candidate model. It combines:

- the empirical-mean upper bound
$$\sum_j \exp(-m\Delta_j^2/4);$$
- a nearest-competitor confidence lower bound; and
- a Fano multiplicity lower bound for clusters of near competitors.

Together these yield matching localization complexity up to universal constants over the stated local permutation class. The scope is deliberately limited: this is not a general active best-arm theorem and does not claim pointwise instance-optimality for arbitrary policy classes.

Main-text Theorem 1 corresponds to Theorem B.1 below. Main-text Proposition 1 corresponds to Proposition B.1 below.

B.2 Model

There are $M \geq 2$ candidate action regions/representations. In validation we observe exactly m independent Gaussian summaries per candidate:

$$X_{j,t} \sim N(\mu_j, 1), \qquad j = 1, \dots, M, \quad t = 1, \dots, m.$$

Assume a unique best candidate $\star$:

$$\mu_\star > \mu_j \qquad (j \neq \star).$$

Order the $M-1$ gaps:

$$0 < \Delta_{(1)} \leq \Delta_{(2)} \leq \cdots \leq \Delta_{(M-1)}, \qquad \Delta_j = \mu_\star - \mu_j.$$

B.2.1 Local permutation class

Let $\mu^{(0)} = \mu$. For each competitor $j \neq \star$, let $\mu^{(j)}$ be obtained by **swapping** the two coordinates $\mu_\star$ and μ_j, leaving every other mean unchanged.

Thus every instance in

$$\mathcal{P}(\mu) = \{\mu^{(0)}\} \cup \{\mu^{(j)} : j \neq \star\}$$

has the same multiset of means and the same ordered gap geometry; only the identity of the best candidate changes.

A selector $\widehat{J}$ is δ-correct on $\mathcal{P}(\mu)$ if

$$\max_{\nu \in \mathcal{P}(\mu)} \Pr_\nu(\widehat{J} \neq J^\star(\nu)) \leq \delta.$$

Define the local permutation minimax selection complexity

$$m^\star_{\mathrm{sel}}(\mu, \delta) = \inf \{m : \exists\ \delta\text{-correct selector on } \mathcal{P}(\mu)\}.$$

For $0 < \delta \leq 1/4$, define

$$\boxed{H_{\mathrm{sel}}(\mu, \delta) = \max\left\{ \frac{\log(1/\delta)}{\Delta_{(1)}^2},\ \max_{2 \leq k \leq M-1} \frac{\log k}{\Delta_{(k)}^2} \right\},}$$

with the second maximum omitted for $M = 2$.

The first term is the **confidence cost** of resolving the closest competitor. The second is the **local multiplicity cost**: k candidates lying within gap $\Delta_{(k)}$ cost $\log k$.

B.3 Main selection theorem

B.3.1 Theorem B.1 — matching instance-dependent localization complexity

For $0 < \delta \leq 1/4$,

$$\boxed{\frac{1}{4} H_{\mathrm{sel}}(\mu, \delta) \ \leq\ m^{\star}_{\mathrm{sel}}(\mu, \delta) \ \leq\ 14 H_{\mathrm{sel}}(\mu, \delta) + 1.}$$

Hence

$$\boxed{m^{\star}_{\mathrm{sel}}(\mu, \delta) = \Theta\big(H_{\mathrm{sel}}(\mu, \delta)\big)}$$

with universal constants.

This is the missing matching converse: the number of *nominal* candidates is not the right instance-dependent quantity. What matters is the ordered spectrum of near-best gaps.

B.4 Lower bound

B.4.1 Lemma B.1 — nearest-competitor confidence lower bound

Let j_1 index the closest competitor, with gap $\Delta_{(1)}$. Compare the base instance μ to the swapped instance $\mu^{(j_1)}$.

Let

$$A = \{\widehat{J} = \star\}.$$

By δ-correctness,

$$P_{\mu}(A) \geq 1 - \delta, \qquad P_{\mu^{(j_1)}}(A) \leq \delta.$$

Only two candidate coordinates change under the swap. Each of their m Gaussian observations shifts by magnitude $\Delta_{(1)}$. Therefore

$$D_{\mathrm{KL}}(P_{\mu} \| P_{\mu^{(j_1)}}) = m \Delta_{(1)}^2.$$

By data processing of KL through the binary event A,

$$m \Delta_{(1)}^2 \geq \mathrm{kl}_{\mathrm{Ber}}(1 - \delta \| \delta).$$

For $0 < \delta \le 1/4$,

$$\mathrm{kl}_{\mathrm{Ber}}(1-\delta\|\delta) \ge \frac{1}{4}\log(1/\delta).$$

Hence

$$\boxed{m \ge \frac{1}{4}\frac{\log(1/\delta)}{\Delta_{(1)}^2}.}$$

B.4.2 LEMMA B.2 — LOCAL MULTIPLICITY LOWER BOUND

Fix $k \ge 2$ and take the k closest competitors. For each such competitor j, consider the swapped instance $\mu^{(j)}$. Under these k hypotheses the correct best-candidate labels are all distinct.

Put a uniform prior on these k hypotheses and denote the random index by J. Use the base law P_μ as a reference distribution.

For the j-th swapped instance,

$$D_{\mathrm{KL}}(P_{\mu^{(j)}}\|P_\mu) = m\Delta_j^2 \le m\Delta_{(k)}^2.$$

The standard reference-distribution bound on mutual information gives

$$I(J;X) \le \frac{1}{k}\sum_{j=1}^{k} D_{\mathrm{KL}}(P_{\mu^{(j)}}\|P_\mu) \le m\Delta_{(k)}^2.$$

Fano's inequality therefore yields

$$\delta \ge 1 - \frac{m\Delta_{(k)}^2 + \log 2}{\log k}.$$

Thus

$$m \ge \frac{\max\{(1-\delta)\log k - \log 2, 0\}}{\Delta_{(k)}^2}.$$

For $\delta \le 1/4$ and $k \ge 4$,

$$(1-\delta)\log k - \log 2 \ge \frac{1}{4}\log k,$$

so

$$\boxed{m \ge \frac{1}{4}\frac{\log k}{\Delta_{(k)}^2} \qquad (k \ge 4).}$$

For $k = 2, 3$, since $\Delta_{(k)} \ge \Delta_{(1)}$ and $\log k \le \log 3 < \log 4 \le \log(1/\delta)$, the nearest-competitor lower bound already controls these terms.

Combining Lemmas 1 and 2 gives

$$\boxed{m \ge \frac{1}{4}H_{\mathrm{sel}}(\mu,\delta).}$$

B.5 UPPER BOUND

Use the empirical-mean selector

$$\widehat{J} = \arg\max_j \bar{X}_j.$$

Under any instance in the local permutation class, the ordered gap multiset remains

$$\Delta_{(1)}, \ldots, \Delta_{(M-1)}.$$

For a competitor with gap Δ,

$$\bar{X}_j - \bar{X}_\star \sim N\left(-\Delta, \frac{2}{m}\right),$$

and therefore

$$P(\bar{X}_j \geq \bar{X}_\star) \leq \exp\left(-\frac{m\Delta^2}{4}\right).$$

Ordering the competitors by gap and union-bounding,

$$P(\widehat{J} \neq \star) \leq \sum_{k=1}^{M-1} \exp\left(-\frac{m\Delta_{(k)}^2}{4}\right).$$

It is enough to require, for every k,

$$\exp\left(-\frac{m\Delta_{(k)}^2}{4}\right) \leq \frac{6\delta}{\pi^2 k^2}.$$

Equivalently,

$$m \geq 4 \max_{1 \leq k \leq M-1} \frac{\log(\pi^2 k^2/(6\delta))}{\Delta_{(k)}^2}.$$

Since $\delta \leq 1/4$,

$$\log \frac{\pi^2 k^2}{6\delta} = \log \frac{\pi^2}{6} + \log(1/\delta) + 2\log k.$$

Using $\Delta_{(k)} \geq \Delta_{(1)}$, the definition of H_{sel}, and $\log(\pi^2/6) \leq 0.36 \log(1/\delta)$,

$$4 \max_k \frac{\log(\pi^2 k^2/(6\delta))}{\Delta_{(k)}^2} \leq 14 H_{\mathrm{sel}}(\mu, \delta).$$

Hence $m \geq 14 H_{\mathrm{sel}}$ makes the error at most

$$\sum_{k=1}^{\infty} \frac{6\delta}{\pi^2 k^2} = \delta.$$

Thus

$$\boxed{m^{\star}_{\mathrm{sel}}(\mu,\delta) \le 14 H_{\mathrm{sel}}(\mu,\delta) + 1.}$$

This proves Theorem B.1.

B.6 CERTIFICATION COMPLEXITY

After selection, use an independent certification sample from the selected action region.

For the Gaussian boundary problem

$$H_0 : \theta \le 0 \qquad \text{versus} \qquad H_1 : \theta \ge d_c,$$

with m_c observations $N(\theta, 1)$, the uniformly most powerful one-sided test rejects when

$$\sqrt{m_c}\,\bar{X} > z_{1-\alpha}.$$

At the least-favorable alternative $\theta = d_c$, its power is

$$\Phi(\sqrt{m_c} d_c - z_{1-\alpha}).$$

Therefore type-I error at most α and power at least $1-\beta$ are achievable **iff**

$$\sqrt{m_c} d_c \ge z_{1-\alpha} + z_{1-\beta}.$$

Thus the exact Gaussian certification complexity is

$$\boxed{m^{\star}_{\mathrm{cert}} = \min\{m \in \mathbb{N} : m d_c^2 \ge (z_{1-\alpha} + z_{1-\beta})^2\}.}$$

B.7 SEARCH-VERSUS-MULTIPLICITY PHASE BOUNDARY

The disjoint-role additive budget is structurally immediate once separate validation and certification samples are required. The more informative comparison is with exhaustive simultaneous certification.

Define

$$B_1 = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{d_c^2},$$

and for Bonferroni certification of M fixed claims,

$$B_M = \frac{(z_{1-\alpha/M} + z_{1-\beta})^2}{d_c^2}.$$

The multiplicity premium is

$$\Delta_B(M) = B_M - B_1.$$

Let $m^{\star}_{\mathrm{sel}}$ be the exact selection complexity from Theorem B.1. Then the idealized local-evidence costs are

$$C_{\mathrm{hon}} = m^{\star}_{\mathrm{sel}} + B_1$$

and

$$C_{\mathrm{exh}} = \max\{m^{\star}_{\mathrm{sel}}, B_M\}.$$

The exhaustive architecture reuses the same full sample for localization and simultaneous certification, so its information cost is the larger of the two requirements rather than their sum.

B.7.1 PROPOSITION B.1 — SEARCH-VERSUS-MULTIPLICITY CROSSOVER

$$C_{\mathrm{hon}} < C_{\mathrm{exh}} \quad \Longleftrightarrow \quad m^{\star}_{\mathrm{sel}} < \Delta_B(M).$$

If $m^{\star}_{\mathrm{sel}} \geq B_M$, exhaustive reuse is cheaper by the single-candidate certification increment B_1, although the ratio tends to one as selection dominates both procedures.

Using Theorem B.1,

$$14H_{\mathrm{sel}} + 1 < \Delta_B(M)$$

is sufficient for the constructive honest cost to be lower, while

$$\frac{1}{4}H_{\mathrm{sel}} > \Delta_B(M)$$

is sufficient for the exact-localization honest cost to exceed the exhaustive baseline.

The unresolved constant-factor band between these conditions is explicit.

For fixed α, β,

$$\Delta_B(M) = \Theta\left(\frac{\log M}{d_c^2}\right).$$

Hence the phase boundary is governed by

$$H_{\mathrm{sel}} \quad \text{versus} \quad \frac{\log M}{d_c^2}.$$

This is an inferential-cost comparison, not a universal dominance theorem: exhaustive Bonferroni controls all fixed claims but does not itself solve exact best-candidate identification.

B.7.2 COROLLARY B.1.1 — ADDITIVE DISJOINT-ROLE BUDGET

For the stipulated select-then-certify architecture,

$$\frac{1}{4}H_{\mathrm{sel}} + m^{\star}_{\mathrm{cert}} \leq n^{\star}_h p \leq 14H_{\mathrm{sel}} + m^{\star}_{\mathrm{cert}} + 1.$$

The exact continuous split is proportional to the unknown exact selection and certification complexities,

$$\rho^{\star} = \frac{m^{\star}_{\mathrm{sel}}}{m^{\star}_{\mathrm{sel}} + m^{\star}_{\mathrm{cert}}},$$

while H_{sel} provides only a regime-level approximation.

B.8 Recovery of the symmetric frontier

If all $M-1$ competitors have a common gap d_s, then

$$\Delta_{(k)} = d_s$$

and, for fixed δ,

$$H_{\mathrm{sel}}(\mu, \delta) = \Theta\left(\frac{\log M}{d_s^2}\right).$$

Hence

$$n_h^\star\, p = \Theta\left(\frac{\log M}{d_s^2} + \frac{\log(1/\alpha)}{d_c^2}\right)$$

for fixed power and small α.

With

$$M = 2^r, \qquad p = 2^{-r}, \qquad d_s \asymp d_c \asymp d,$$

the localization term dominates at growing r:

$$n_h d^2\, 2^{-r} \asymp r.$$

Therefore

$$\boxed{r = \log_2(n_h d^2) - \log_2 \log(n_h d^2) + O(1).}$$

Thus the original log-minus-log-log law is recovered as the symmetric worst-case case of the instance-dependent theorem.

B.9 Instance-dependent interpretation

Define the empirical-mean sufficient localization complexity

$$A_{\mathrm{loc}} = \inf\left\{ m : \sum_j e^{-m\Delta_j^2/4} \le \delta \right\}.$$

Theorem B.1 shows that, on the finite local permutation class considered here,

$$A_{\mathrm{loc}}$$

has the same order, up to universal constants, as

$$H_{\mathrm{sel}} = \max\left\{ \frac{\log(1/\delta)}{\Delta_{(1)}^2}, \max_k \frac{\log k}{\Delta_{(k)}^2} \right\}.$$

The first term captures a single hardest competitor. The second captures multiplicity among near competitors: if k candidates lie inside gap $\Delta_{(k)}$, reliable localization requires order

$$\frac{\log k}{\Delta_{(k)}^2}.$$

Thus a single near tie produces a nearest-competitor cost, whereas a broad near-tie plateau recovers the symmetric $\log M$ worst-case behavior.

B.10 VALUE-TOLERANT LOCALIZATION

For $\epsilon \geq 0$, define

$$\mathcal{G}_\epsilon(\mu) = \{j : \mu_\star - \mu_j \leq \epsilon\}$$

and

$$\mathcal{B}_\epsilon = \{j \neq \star : \Delta_j > \epsilon\}.$$

Let the bad-candidate gaps be ordered

$$\Gamma^\epsilon_{(1)} \leq \cdots \leq \Gamma^\epsilon_{(L_\epsilon)}.$$

Define

$$H^{\mathrm{pass}}_\epsilon = \max\left\{\frac{\log(1/\delta)}{(\Gamma^\epsilon_{(1)})^2}, \max_{2\leq k\leq L_\epsilon} \frac{\log k}{(\Gamma^\epsilon_{(k)})^2}\right\},$$

with value zero when $L_\epsilon = 0$.

B.10.1 PROPOSITION B.2 — TOLERANCE-TRUNCATED UPPER BOUND

For empirical-mean selection,

$$\Pr\{\mu_\star - \mu_{\widehat{J}} > \epsilon\} \leq \sum_{k=1}^{L_\epsilon} \exp\left(-\frac{m(\Gamma^\epsilon_{(k)})^2}{4}\right).$$

Hence

$$m \geq 14 H^{\mathrm{pass}}_\epsilon + 1$$

is sufficient for (ϵ, δ)-correctness.

Proof An ϵ-error can occur only when a candidate $j \in \mathcal{B}_\epsilon$ beats the empirical mean of the true best candidate. For each such candidate,

$$\Pr(\bar{X}_j \geq \bar{X}_\star) \leq e^{-m\Delta_j^2/4}.$$

Union-bound only over $\mathcal{B}_\epsilon$, then allocate tail probability $6\delta/(\pi^2 k^2)$ to the k-th bad candidate. The constant audit is identical to the proof of Theorem B.1. □

B.10.2 COROLLARY B.2.1 — SEPARATED MATCHING REGIME

If

$$\epsilon < \Delta_{(1)},$$

then $\mathcal{G}_\epsilon = \{\star\}$, so ϵ-optimal selection is identical to exact selection and

$$H_\epsilon^{\mathrm{pass}} = H_{\mathrm{sel}}.$$

Therefore

$$\frac{1}{4} H_\epsilon^{\mathrm{pass}} \le m_\epsilon^\star \le 14 H_\epsilon^{\mathrm{pass}} + 1.$$

B.10.3 PROPOSITION B.3 — DEGENERACY OF THE ORIGINAL SINGLE-SWAP CLASS

Suppose there exists $g \neq \star$ with

$$\Delta_g \le \epsilon.$$

Then on the original single-swap class $\mathcal{P}(\mu)$, the deterministic zero-sample rule $\widehat{J} = g$ is ϵ-correct.

Proof Under the base instance, g is ϵ-optimal. Under the swap involving g, it becomes best. Under any other single swap its mean remains μ_g, while the best mean remains $\mu_\star$, so its gap remains $\Delta_g \le \epsilon$. □

This proves that a general multi-good-arm matching lower bound cannot be obtained by naively reusing the exact theorem's single-swap alternative class.

B.10.4 WORST-CASE PASSIVE SCALING

Since every bad candidate has gap $> \epsilon$,

$$m \ge \frac{4}{\epsilon^2} \log \frac{M-1}{\delta}$$

is sufficient in the worst case.

Conversely, for the symmetric instance

$$\mu_\star = 2\epsilon, \qquad \mu_j = 0 \quad (j \neq \star),$$

the best candidate is the only ϵ-optimal candidate, so Theorem B.1 gives

$$m = \Omega\left(\frac{\max\{\log M, \log(1/\delta)\}}{\epsilon^2}\right).$$

Thus the balanced passive worst-case rate is

$$m_{\epsilon,\mathrm{worst}} = \Theta\left(\frac{\log(M/\delta)}{\epsilon^2}\right)$$

up to universal constants.

B.11 TRANSLATION TO POPULATION POLICY VALUE

For equal-mass disjoint regions of mass p,

$$V_j - V_0 = p\mu_j$$

and therefore

$$V_\star - V_j = p\Delta_j.$$

For value tolerance η, put

$$\epsilon_\eta = \eta/p.$$

Then Proposition B.2 gives

$$\Pr\{V_\star - V_{\widehat{J}} > \eta\} \le \sum_{j:p\Delta_j>\eta} e^{-m\Delta_j^2/4}.$$

Define

$$H_{\mathrm{val}}(\mu, \eta, \delta) = H^{\mathrm{pass}}_{\eta/p}(\mu, \delta).$$

Then

$$m \ge 14H_{\mathrm{val}} + 1$$

is sufficient for value regret at most η with probability at least $1 - \delta$.

B.11.1 RESOLUTION-VALUE DECOUPLING

With $p = 1/M$, a local effect mistake of size Δ incurs population regret

$$\Delta/M.$$

Thus exact candidate identity can become harder with increasing M while the value cost of confusing two small regions decreases. This does not remove certification: a low-regret candidate can still be clinically invalid.

B.12 VALUE-TOLERANT PRICE OF HONESTY

Assume every η-optimal candidate has certification standardized margin at least $d_{c,\eta} > 0$, and define

$$m^\star_{\mathrm{cert},\eta} = \min\left\{m : md^2_{c,\eta} \ge (z_{1-\alpha} + z_{1-\beta})^2\right\}.$$

If

$$m_v \ge 14H_{\mathrm{val}} + 1$$

and

$$m_c \ge m^\star_{\mathrm{cert},\eta},$$

independence gives

$$\Pr\{\eta\text{-optimal selection and certification}\} \geq (1-\delta)(1-\beta).$$

Let $n^{\star}_{h,\eta}$ denote the minimum total sample size for this stipulated disjoint-role construction. For equal-mass regions,

$$n^{\star}_{h,\eta} p \leq 14 H_{\mathrm{val}} + m^{\star}_{\mathrm{cert},\eta} + 1.$$

Equivalently, taking the local validation and certification budgets equal to the two sufficient requirements is enough.

A matching general lower bound is not claimed when several candidates are η-optimal.

B.13 Simultaneous plausible-optimal confidence sets

Suppose simultaneous one-sided lower bounds satisfy

$$L_{kj} \leq \mu_k - \mu_j$$

for every ordered pair $k \neq j$ with joint probability at least $1-\alpha$.

Define

$$\widehat{\mathcal{G}}_{\epsilon} = \left\{ j : \max_{k \neq j} L_{kj} \leq \epsilon \right\}.$$

B.13.1 Proposition B.4 — outer confidence set

With probability at least $1-\alpha$,

$$\mathcal{G}_{\epsilon} \subseteq \widehat{\mathcal{G}}_{\epsilon}.$$

Proof If $j \in \mathcal{G}_{\epsilon}$, then for every k,

$$\mu_k - \mu_j \leq \mu_{\star} - \mu_j \leq \epsilon.$$

On simultaneous coverage,

$$L_{kj} \leq \mu_k - \mu_j \leq \epsilon,$$

so $j \in \widehat{\mathcal{G}}_{\epsilon}$. □

For independent Gaussian estimates with standard errors s_j, use all

$$M(M-1)$$

ordered comparisons and

$$L_{kj} = (\widehat{\mu}_k - \widehat{\mu}_j) - z_{1-\alpha/[M(M-1)]} \sqrt{s_k^2 + s_j^2}.$$

Equivalently, use all unordered pairs with a two-sided per-pair Bonferroni correction.

B.14 THEORY CLOSURE AND CRITEO INTERPRETATION

The theoretical claims are now intentionally stratified:

1. **Worst-case certifiable resolution:** $npd^2 = \Theta(\log M)$.
2. **Exact localization:** matching H_{sel} theorem on the balanced Gaussian local-permutation class.
3. **Value-tolerant localization:** tolerance-truncated passive upper bound, matching in the separated regime.
4. **Multi-good-arm caution:** the original single-swap class is provably degenerate once a common ϵ-optimal candidate exists.
5. **Independent certification:** additive certification information for the stipulated disjoint-role architecture.
6. **Simultaneous real-data interpretation:** outer confidence sets distinguish uncertain winner identity from demonstrated value loss.

The prospective Criteo raw-cardinality scaling test remains failed. The value-tolerant theory is a post-experiment refinement and is not used to relabel that endpoint.

The exact replay of the frozen Criteo geometry shows that the best-versus-runner local effect gap shrinks from 0.03745 at depth 6 to 0.00747 at depth 8, while the corresponding equal-mass population-value gap shrinks from approximately 5.85×10^{-4} to 2.92×10^{-5}. Thus exact winner identity and decision consequence demonstrably separate at fine depth.

For the simultaneous outer-set construction, use $M(M-1)$ ordered one-sided comparisons, or the equivalent two-sided correction over unordered pairs. The main Criteo analysis instead reports the less conservative Multiple-Comparisons-with-the-Best audit.

The remaining theoretical limitations are scope limitations rather than unresolved proof gaps: correlated policy-value estimates, unequal candidate masses, nuisance-estimation effects, and optimal adaptive allocation are outside the proved Gaussian oracle.

C Exact Final HonestER Procedure and Validity Proof

C.1 Exact final procedure

Let $D_{\mathrm{tr}}, D_{\mathrm{val}}, D_{\mathrm{cert}}$ be independent. Fit all candidate representation families, score functions, and thresholds using D_{tr} only. For each family b, compute the frozen validation statistics required by the ECV entry table on D_{val} and choose the best positive ECV entry with deterministic tie-breaking. Freeze that entry before opening D_{cert}.

Set $(w_{\mathrm{clin}}, w_{\mathrm{HGB}}, w_{\mathrm{Ridge}}) = (0.25, 0.10, 0.65)$ and $\alpha_b = \alpha w_b$. For each selected family path, test the entry benefit at α_b. If it passes, test successive shell exclusions at the same family-local level in fixed sequence, stopping at the first failure. Let $\widehat{d}_b$ be the deepest consecutive passing depth. A family with failed entry has depth zero.

For each family, associate the certified depth with the **empirical-Bayes shrunk validation value-by-depth score** produced by the frozen ECV table before certification. Among families with positive score and $\widehat{d}_b > 0$, deploy the highest-scoring certified output with deterministic family tie-breaking. If none exists, deploy baseline/abstain.

C.2 Proof of conditional fixed-portfolio validity

Condition on $D_{\mathrm{tr}}, D_{\mathrm{val}}$. All selected entries, fixed sequences, family weights, margins, endpoints, and deployment scores are now deterministic. Let E_b be the event that family b certifies beyond its first population-invalid condition. By the first-true-null argument for fixed-sequence testing, $\Pr(E_b \mid D_{\mathrm{tr}}, D_{\mathrm{val}}) \le \alpha_b$. If the final portfolio output is population-invalid, at least one E_b must occur because final selection is restricted to certified family prefixes. Therefore

$$\Pr(\text{invalid final deployment} \mid D_{\mathrm{tr}}, D_{\mathrm{val}}) \le \Pr\left(\bigcup_b E_b\right) \le \sum_b \alpha_b = \alpha \sum_b w_b \le \alpha.$$

No independence is needed. Integrating over upstream data gives the corresponding unconditional statement when component-test validity holds conditionally for every frozen upstream realization.

C.3 Why the direct value-selection theorem is auxiliary

For a fixed finite candidate class, honest point-value maximization enjoys the usual uniform-deviation regret bound. This theorem remains useful for understanding selection error and for the ValueFixed ablation, but the final method ranks entries by ECV. We therefore do not present the finite-class value bound as a theorem about ECV optimality.

D ECV ENTRY SCORE AND THE ADAPTIVE-ALLOCATION AUDIT

For an entry or signed shell condition with validation estimate m_ℓ and standard error s_ℓ, the implementation estimates a zero-centered Normal prior variance

Define

$$q_\ell = \max\{m_\ell^2 - s_\ell^2, 0\},$$

and set

$$\widehat{\tau}^2 = \min\{\max\{\mathrm{mean}_\ell\, q_\ell, 0.02^2\}, 0.50^2\}.$$

The Normal-Normal posterior uses shrinkage

$$\lambda_\ell = \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + s_\ell^2}, \qquad \mu_\ell = \lambda_\ell m_\ell,$$

with posterior variance

$$v_\ell = \frac{\widehat{\tau}^2 s_\ell^2}{\widehat{\tau}^2 + s_\ell^2}.$$

If projected certification standard error is $s_{c,\ell} = s_\ell\sqrt{n_{\mathrm{val}}/n_{\mathrm{cert}}}$, the approximate probability of passing a one-sided condition at local level α_b is computed from the posterior predictive Normal distribution. Entry and shell pass probabilities form an approximate fixed-sequence deployment-depth distribution; value at each depth is region mass times posterior mean net benefit, truncated at zero for the ECV objective.

The historical adaptive portfolio additionally enumerated joint path-depth outcomes and optimized weights on a 0.05 simplex grid. Honest evaluation found no demonstrated advantage for this **adaptive across-family weighting** over the fixed 25/10/65 allocation in the tested settings. The final method therefore uses fixed family weights while retaining two outputs of the ECV table: the selected within-family entry and the empirical-Bayes shrunk value-by-depth score used for deployment ranking after certification. ECV remains a planning surrogate, not a calibrated certificate or an unbiased deployment-value estimator.

D.1 OBSERVABLE ECV SCORE-SELECTION AMBIGUITY

Let the complete validation parameter vector be θ, and let the deterministic ECV score for candidate j be

$$S_j(\theta) = F_j(\theta).$$

Let $C(D_{\mathrm{val}})$ be any joint $1-\gamma$ confidence region for θ, and define

$$L_j = \inf_{\theta\in C} F_j(\theta), \qquad U_j = \sup_{\theta\in C} F_j(\theta).$$

If $\widehat{j}$ is the observed ECV maximizer, then on the coverage event,

$$\boxed{\max_j S_j(\theta_0) - S_{\widehat{j}}(\theta_0) \le \max_j U_j - L_{\widehat{j}}.}$$

This gives an observable **ECV-score ambiguity bound**. It is not a policy-value regret bound. A value guarantee would additionally require a validated bridge between population ECV score and policy value, which we do not assume.

D.2 WHY PROJECTED ECV PASS PROBABILITIES NEED NOT CALIBRATE

Consider a simplified Gaussian claim with validation estimate

$$\widehat{\tau}_v = \tau + \frac{\sigma}{\sqrt{n_v}} Z, \qquad Z \sim N(0,1).$$

True certification power at certification size n_c is

$$q(\tau) = \Phi\left(\frac{\sqrt{n_c}\tau}{\sigma} - z_{1-\alpha}\right).$$

The raw plug-in projection is

$$\widehat{q} = \Phi\left(\frac{\sqrt{n_c}\widehat{\tau}_v}{\sigma} - z_{1-\alpha}\right).$$

Writing

$$a = \frac{\sqrt{n_c}\tau}{\sigma} - z_{1-\alpha}, \qquad b = \sqrt{\frac{n_c}{n_v}},$$

gives

$$\widehat{q} = \Phi(a + bZ).$$

Therefore

$$\mathbb{E}\left[\left(\frac{\sqrt{n_c}(\widehat{\tau}_v - \tau)}{\sigma}\right)^2\right] = \frac{n_c}{n_v},$$

and the Normal identity

$$\mathbb{E}[\Phi(a + bZ)] = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

yields

$$\boxed{\mathbb{E}[\widehat{q}] = \Phi\left(\frac{a}{\sqrt{1 + n_c/n_v}}\right).}$$

Near the certification boundary, if $n_c/n_v \to c > 0$, the raw projected probability does not converge to the true power $\Phi(a)$. This explains why ECV pass probabilities are planning surrogates rather than calibrated probabilities. Empirical-Bayes shrinkage changes the projection but does not by itself create a calibration theorem.

E EXPERIMENTAL PROGRESSION AND PRESPECIFICATION

The empirical program is sequential, not one prospectively frozen mega-study.

E.1 ORIGINAL FIXED-HIERARCHY AND FIXED-SEQUENCE DIAGNOSTICS

The original 600-repetition aligned boundary benchmark gave 92.17% target recovery for the partition-aware fixed-hierarchy rule versus 70.83% ForestDRLearner and 66.50% GRF under a common held-out certificate. A full-sample Bonferroni reference reached 89.5%, making the method-specific increment only 2.67 points (95% CI 0.33-5.00). Later hierarchy-misspecification and calibration studies motivated the broader representation and certifier audits rather than a claim of universal algorithmic superiority.

E.2 ACTG SEMI-SYNTHETIC REPRESENTATION STUDY AND LEARNED-PATH PILOT

The UCI ACTG-175 table (Hammer et al., 1996; UCI Machine Learning Repository, 2023) is used only for baseline covariate geometry. Sixteen baseline variables remain after pre-outcome QC; age, CD4, CD8, and weight define eight fixed upper/lower-tail paths. The v1.2 geometry was locked before powered method results. Cross-cut unrestricted oracle value is 0.04363 and best fixed-hierarchy value 0.01048. The learned-path pilot uses cross-fitted DR scores on the representation-learning split and fixed percentile masses. Its positive result is attributed to representation improvement.

E.3 HISTORICAL PORTFOLIO AND HONEST-SELECTION AUDIT

The historical Stage-2 grid used a fresh seed namespace after a gated Stage 1. Mean clinician/HGB/Ridge alpha shares at $n = 8000$ were (0.2273, 0.0523, 0.7203) for smooth, (0.4778, 0.0510, 0.4712) for rare nonlinear, and (0.0282, 0.0508, 0.9210) for cross-cut. The later honest-selection audit used a separate locked protocol and found no demonstrated advantage for adaptive weighting over the fixed allocation in the tested settings.

E.4 PROSPECTIVELY STOPPED STACKING EXTENSION

CertiStack used 25% representation training, 12.5% candidate selection, 12.5% independent safe-improvement gating, and 50% certification. Stage 1 contained 800 fresh cases over axis-tail, oblique-linear, XOR cross-cut, and mixed linear-interaction geometries at $n = 8000$. It passed the uniform-portfolio and safety gates but failed the predeclared value gate versus honest base selection and failed to improve adaptation regret. Stage 2 was not run and no rescue retuning was performed.

E.5 RETROSPECTIVE GLOBAL ORDERED-CERTIFICATION DIAGNOSTIC

After the honest-selection audit, a retrospective claim-level replay asked whether one validation-fixed global testing order could remove family-wise alpha splitting. The reachable-policy oracle had mean headroom 0.00410 over the fixed portfolio. Yet GlobalMaxZ and EntryFirstMaxZ were lower by 0.001054 pooled, and GlobalValue by 0.000751. Cross-cut blocking loss was substantial. These results retire the tested schedulers from the final procedure but do not imply that every possible global ordering must fail. Because this analysis was designed after observing the honest-selection results, it is reported as a mechanism diagnostic only.

E.6 FRESH PROSPECTIVE ECV-FREE SIMPLIFICATION

Development and retrospective method selection used replicate indices 0-299. Before confirmation outcomes, ValueFixed and the analysis were locked on replicate indices 300-599, with no seed overlap, three DGPs, $n = 8000$, 300 replicates per DGP, primary endpoint supported deployed value, comparator FixedWeights_25_10_65_FS, and noninferiority margin 0.0005. ValueFixed changed two validation-side components: ECV entry ranking was replaced by positive point-value entry ranking, and empirical-Bayes shrunk deployment scores were replaced by unshrunk validation values. Family allocation and independent certification were unchanged. Replay controls reproduced historical SinglePath, RidgeOnly, and HGBOnly outputs exactly in all 900 cases.

The mean ValueFixed-minus-FixedECV difference was -0.000256. The primary one-sided 95% LCB was -0.000553, so the prespecified noninferiority criterion was not met. The pooled two-sided 95% interval [-0.000610, 0.000099] includes zero; the experiment does not establish superiority or inferiority. Cross-cut supplied the clearest descriptive deficit. No post-hoc margin change or selector rescue is used.

E.7 Post hoc two-by-two attribution diagnostic

The saved 61,200 prospective claim rows contain no duplicate case-claim keys, certification p-values reconstruct correctly, and both reported procedures reproduce exactly on all 900 cases. A post hoc 2x2 replay crossed ECV versus point-value entry selection with empirical-Bayes shrunk versus unshrunk deployment scoring. Mean supported values were 0.030366 (ECV/shrunk), 0.030393 (ECV/unshrunk), 0.030073 (Value/shrunk), and 0.030110 (Value/unshrunk). Holding deployment scoring fixed, the entry-rule differences were -0.000293 (pointwise 95% CI [-0.000636, +0.000050]) and -0.000283 ([-0.000619, +0.000052]). These diagnostics suggest that entry ranking explains most of the observed mean difference, but both intervals include zero and the analysis is explicitly post hoc.

The same fresh paired rows give FixedECV minus validation single-path = +0.001187 (pointwise 95% CI [+0.000443, +0.001932]) and FixedECV minus adaptive honest ECV = +0.000508 ([-0.000045, +0.001061]). These are secondary/descriptive comparisons.

E.8 Fresh boundary confirmation

The boundary protocol fixes $\alpha = .05$, $\kappa = .10$, propensity .5, path masses (.5, .25, .125, .0625), $n \in \{512, 2048, 4096\}$, and homoskedastic $(\sigma_0, \sigma_1) = (.3, .3)$ or heteroskedastic $(.2, .45)$ Normal outcomes. The 42 exact-boundary cells contain 5,000 repetitions each. Pooled entry-boundary false certification is 6163/120000=0.05136; the worst cell is entry level 4, $n = 512$, heteroskedastic, at 0.0564. In all 210,000 exact-boundary rows, false-certificate and invalid-final-deployment indicators agree.

F CRITEO REAL-OUTCOME PROTOCOL, RESULTS, AND AUDITS

F.1 DATASET AND PROSPECTIVE DATA ROLES

We use the unbiased CRITEO-UPLIFT release derived from randomized advertising incrementality tests (Diemert et al., 2018). The official release contains 13,979,592 rows. The protocol uses `visit` as the primary outcome and excludes `exposure`, which is post-treatment. A deterministic SplitMix64 hash of immutable row ID with seed 20260924 created three disjoint pools:

Role	Rows	Fraction
Representation	2,795,392	20.0%
Reference geometry	2,796,306	20.0%
Experiment	8,387,894	60.0%

A LightGBM T-learner score (Ke et al., 2017) is fit only in the representation pool (maximum one million training rows) and frozen before the experiment outcomes are used. Reference-pool score quantiles define 2^r equal-mass bins for depths $r = 1, \ldots, 8$.

F.2 FROZEN REFERENCE GEOMETRY

The reference geometry shows increasing certification signal with depth but a collapse in best-versus-competitor separation after depth 5.

r	J	Best bin	Mass	Best effect	d_{cert}	Hardest competitor	Competitor effect	d_{loc}
1	2	1	0.500159	0.017414	0.027876	0	0.003085	0.018576
2	4	3	0.250018	0.031238	0.036585	0	0.005954	0.023928
3	8	7	0.125000	0.050438	0.045931	0	0.011859	0.027811
4	16	15	0.062500	0.071723	0.055222	0	0.021262	0.029590
5	32	31	0.031250	0.102156	0.071743	30	0.032499	0.037991
6	64	63	0.015625	0.116751	0.076780	62	0.079297	0.018586
7	128	127	0.007813	0.126103	0.079997	126	0.103376	0.010585
8	256	255	0.003907	0.131009	0.083263	254	0.123541	0.003363

The v1.1 amendment was made before the powered run because the initial reference statistic measured best-bin effect relative to the clinical boundary, whereas exact localization is governed by best-versus-competitor separation. A second pre-power amendment replaced an n_{80}-only endpoint with held-out prediction of the full success curves to avoid discarding right-censored high-resolution depths.

F.3 LOCKED SCALING STUDY: CERTIFICATION HELD, EXACT LOCALIZATION FAILED

The powered study used 100 repetitions at each of 11 sample sizes $n = 2^{10}, \ldots, 2^{20}$ and eight depths, yielding exactly 8,800 rows. For each (n, r), half of the sampled observations selected the bin with the largest estimated treatment effect and the other half applied the one-sided Welch-Normal certificate.

The locked endpoint was

$$\{\text{selected bin equals empirical reference-best bin}\} \cap \{\text{selected bin certifies}\}.$$

Depths 1–4 fit the prespecified scaling constants and depths 5–8 were held out. The certification-only predictor was compared with a joint predictor that added the raw-cardinality localization threshold.

The joint predictor failed prospectively. Held-out log loss was 0.5244208 versus 0.4008157 for certification only; ordinary replicate-level Brier score was 0.1347364 versus 0.1314364. The negative endpoint is retained unchanged.

The two components behave very differently. At $n = 2^{20}$, exact reference-best recovery / candidate certification / joint success were 96/100/96% at depth 6, 93/100/93% at depth 7, and 40/94/39% at depth 8. Thus the **certification side remains strong while exact winner identity fails at fine depth**.

This distinction matters because the locked endpoint tests exact reference-best recovery **plus** certification, not pure certifiable resolution. The prospective failure therefore does not falsify the certifiable-resolution theorem; it falsifies the specified quantitative extrapolation that used raw candidate count to model exact localization jointly with certification.

F.4 POST HOC MECHANISM CLOSURE: CERTIFICATION CALIBRATION AND MCB WINNER UNCERTAINTY

The original reference-preparation pipeline was rerun and reproduced the frozen `reference_geometry_v1_1.csv` byte-for-byte against the locked SHA256 manifest. The reference geometry is therefore an exact replay rather than a refit.

At depth 8, selected-bin frequencies in the powered experiment are 47/100 for bin 254, 40/100 for empirical reference winner 255, and 13/100 for the remaining bins. The frozen reference effects of bins 255 and 254 are 0.131009 and 0.123541, respectively, a local effect gap of 0.007468 and a descriptive best-runner separation of only 0.351 standard errors.

Across depths 6, 7, and 8:

Depth	Best-runner local effect gap	Descriptive pairwise z	Equal-mass population-value gap
6	0.037454	3.885	5.85×10^{-4}
7	0.022726	1.565	1.78×10^{-4}
8	0.007468	0.351	2.92×10^{-5}

Thus exact winner identity becomes harder at fine depth while the policy-value cost of confusing the two leading equal-mass bins becomes much smaller.

We next run a post hoc **no-fit certification calibration** over all 8,800 locked rows. For each run i, let j_i be the bin selected by the locked selection split. The primary diagnostic predicts

$$\widehat{q}_i = \Phi\left(\frac{\tau^{\mathrm{ref}}_{j_i}}{s^{\mathrm{cert}}_i} - z_{1-\alpha}\right),$$

where $\tau^{\mathrm{ref}}_{j_i}$ is the frozen reference effect and s^{cert}_i is the run's certification standard error. No constants are fitted and the run's certification effect/pass indicator is not used to construct $\widehat{q}_i$.

Across all 8,800 runs, predicted mean certification probability is 0.4903 versus observed 0.5227. Row-level Brier score is 0.1076, log loss 0.3286, and mean absolute error 0.1977. Across the 88 (n, r) cells, RMSE is 0.0667 and MAE is 0.0412. On fine depths $r = 5, \ldots, 8$, observed mean certification is 0.4123 versus predicted 0.3822, with cell RMSE 0.0610 and MAE 0.0423.

Because the primary diagnostic uses a realized certification standard error estimated inside the certification split, we also report a stricter sensitivity using only frozen reference effect/noise and nominal bin support implied by (n, r). It remains close: row-level Brier 0.1201, log loss 0.3760, MAE 0.2126, overall predicted mean 0.4916 versus observed 0.5227, and cell RMSE 0.0719. The agreement is therefore not driven solely by the realized-SE projection.

Winner identity is audited using a Bonferroni-within-candidate Multiple-Comparisons-with-the-Best construction. The MCB plausible-best set is:

Depth	Critical value	Plausible-best count	Plausible-best bins
6	3.158	1	63
7	3.357	3	127, 126, 124
8	3.545	8	255, 254, 253, 252, 249, 251, 248, 250

The candidate-wise 0.01-tolerance audit expands those sets to 2, 4, and 9 candidates, respectively, but we treat that quantity only as a descriptive candidate-wise practically-best analysis rather than a simultaneous outer confidence set for every ϵ-optimal candidate.

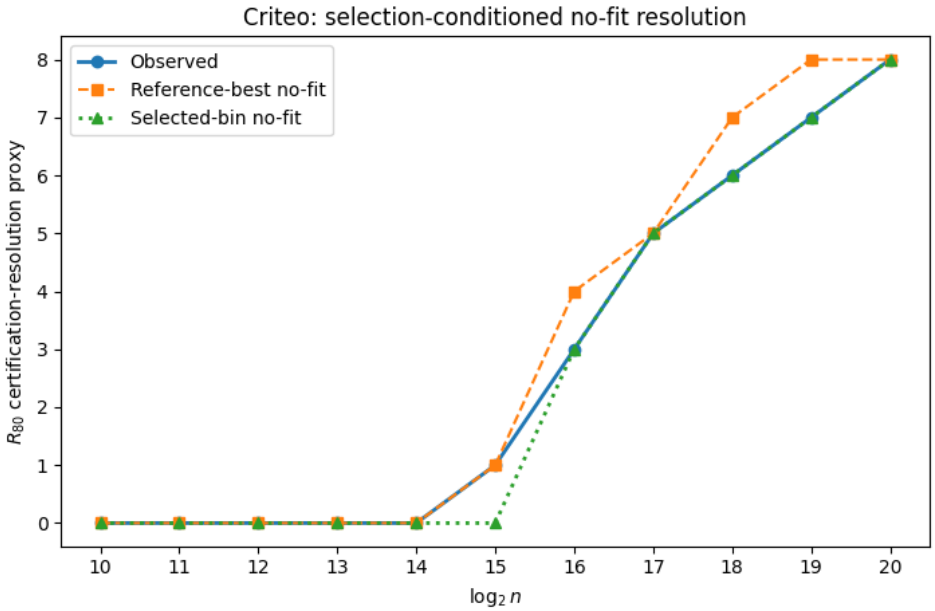

Figure 3: Observed R_{80}, deterministic reference-best no-fit prediction, and selection-conditioned no-fit prediction. Conditioning on the upstream-selected region raises exact agreement from 8/11 to 10/11 sample sizes and removes all optimistic frontier errors.

The mechanism conclusion is therefore two-part. First, certification probability is reasonably predictable from frozen reference geometry without fitting a new scaling constant. Second, fine-depth winner identity is genuinely ambiguous under an MCB analysis, while the value loss attached to confusing the top candidates is tiny. The locked exact-bin-plus-certification endpoint remains failed; these post hoc analyses explain **where** it failed and why that failure is not equivalent to a collapse in certifiable resolution or deployed value.

F.4.1 Selection-conditioned no-fit robustness

The deterministic reference-best frontier assumes that the selection split identifies the frozen reference-best region. To isolate this assumption, we recompute certification probabilities using the frozen reference effect and noise of the region actually selected upstream in each locked run. This prediction uses the realized upstream selection decision but **no certification effect, certification pass/fail outcome, or fitted calibration constant**.

The resulting R_{80} curve matches the observed certification-resolution proxy at 10 of 11 sample sizes, compared with 8 of 11 for the reference-best no-fit curve. All three one-bit optimistic reference-best errors at $2^{16}, 2^{18}, 2^{19}$ disappear; the only remaining discrepancy is one bit pessimistic at 2^{15}. Mean absolute resolution error falls from 0.273 to 0.091 bits, and cell-level RMSE/MAE improve from 0.0785/0.0530 to 0.0719/0.0452.

This is a decomposition/robustness diagnostic rather than an ex-ante n, r-only prediction: it shows that most remaining deterministic-frontier error is attributable to finite-sample localization rather than failure of the certification power model.

F.5 Proposed versus independently certified quota resolution

A separate prospectively specified diagnostic evaluates ordinary T- and DR-learner score pipelines. Each repetition uses 50% training, 25% selection, and 25% certification. On the selection split, the learned score induces nested top-mass quota policies at masses $1/2, 1/4, \ldots, 1/256$; the positive point-value depth is selected. The certification split audits the corresponding nominal quota path. Because each split forms its own top-mass set using the common learned scoring function, this is

a **quota-policy resolution audit at common nominal mass**, not strict replication of an identical numerical score threshold.

The six-row aggregate summary contains 30 repetitions per method/sample-size cell:

Method	n	Proposed bits	Certified bits	Mean unsupported gap	Unsupported runs
DR-Learner	65,536	1.233	0.633	0.600	15/30 = 50%
T-Learner	65,536	1.467	0.633	0.833	16/30 = 53%
DR-Learner	262,144	1.200	1.000	0.200	6/30 = 20%
T-Learner	262,144	1.233	0.967	0.267	6/30 = 20%
DR-Learner	1,048,576	1.000	1.000	0.000	0/30
T-Learner	1,048,576	1.000	1.000	0.000	0/30

Wilson 95% intervals for unsupported-run rates are approximately 33%-67% for 15/30, 36%-70% for 16/30, 10%-37% for 6/30, and 0%-11% for 0/30. These intervals are computed from the retained aggregate counts.

The supported conclusion is narrow: **validation-selected quota policies can propose finer nominal action granularity than the specified independent certifier supports at smaller sample sizes in this benchmark**. The disappearance of the gap at $n = 2^{20}$ coincides with proposed resolution falling to one bit; it is not evidence that increasing sample size monotonically produces finer supported personalization. This experiment also does not evaluate external policy value of the final clinician/HGB/Ridge HonestER portfolio.

G REGISTERED FOLLOW-UP PROTOCOLS

G.1 STATUS AT SUBMISSION

Protocol A — Lenta independent confirmation. The frozen Protocol A v3.1 was executed only through its prespecified reference-only informativeness gate. The gate returned `UNINFORMATIVE_STOP`: 0 of 15 sample sizes were informative, with `Rmax = 6`. No Lenta experiment-pool outcome analysis was opened. The terminal stop record SHA-256 is `8620a51fc097944d3e725c478b387643c07ebe1b38d763bc7b96c4caba5514ab`. The protocol text below is reproduced verbatim and remains the record of the prespecified analysis.

Protocol B — Criteo invalid versus unprovable personalization. Protocol B v3 remains frozen for a prospectively specified analysis on previously inspected randomized Criteo data. It is not described as independent external confirmation. The protocol text below is reproduced verbatim.

Frozen protocol hashes:

- Protocol A v3.1: `13b21c1aa44f90f324d7c2b3fb50eace1048cbc514ec5226fe8c57c674747be2`
- Protocol B v3: `c33f93265813309bd7b7a18bc6611293bdc43b3458d2ecc321744bff800460ed`

G.2 PROTOCOL A: PROSPECTIVE CONFIRMATION OF THE ZERO-FIT CERTIFICATION-RESOLUTION PREDICTOR

Version 3.1. This pre-Stage-A amendment supersedes `PROTOCOL_A_Lenta_NoFit_Confirmation_v3.md`.

The scientific analysis is frozen by the SHA-256 hash recorded in `LOCK_MANIFEST_v3_1.json`.

Pre-Stage-A schema amendment Before model fitting or any Lenta treatment-effect, power, or experiment-outcome statistic was computed, direct inspection of the downloaded raw CSV header established that this public release contains no `CardHolder` or other customer identifier. The attempted Stage-A preparation therefore stopped.

Version 3.1 changes only deterministic unit identity. Each unit is keyed by the exact raw-dataset SHA-256 and its zero-based row number in that byte-identical CSV release. No endpoint, outcome, representation model, threshold, sample-size grid, R1/R2 role, certificate, or success criterion is changed. Protocol-A results may be described as prospective confirmation only if the attestation in A0 is true and the second-stage lock in A9 is completed before any outcome from the Lenta experiment pool is examined.

G.2.1 A0. PROSPECTIVE-STATUS ATTESTATION

Prospective-confirmation status requires all of the following at the time this protocol is submitted/frozen:

1. no author has computed a treatment-effect statistic on the Lenta dataset;
2. no author has inspected any outcome from the eventual Lenta experiment pool;
3. no Lenta analysis result has been used to alter this protocol.

If any item is false, the analysis may still be run, but it must be labelled post hoc rather than prospective. Any later deviation is recorded in a dated deviations file with reason and code diff. A deviation made after experiment-pool outcomes are accessed is reported as a protocol violation.

G.2.2 A1. HYPOTHESIS

A certification-resolution curve computed from frozen reference geometry, with no fitted calibration constants, predicts independent certification frequencies and the certification-resolution proxy R_{80} observed on independent subsamples of an independent randomized dataset.

The primary prediction P1 depends only on the frozen reference pool, n, and depth r. The selection-conditioned prediction P2 is secondary because it uses the realized upstream selection decision, although it still uses no certification outcome.

G.2.3 A2. DATASET, ROW IDENTITY, FEATURES, AND DETERMINISTIC POOLS

- Dataset: Lenta uplift data distributed by scikit-uplift using `sklift.datasets.fetch_lenta(return_X_y_t=True)`.
- Treatment: `group`, mapped deterministically to treatment $A = 1$ for `test`/treated and $A = 0$ for control. The exact observed labels and mapping are written to the dataset manifest before any treatment-effect statistic is computed.
- Outcome: `response_att` (binary).
- Immutable unit identity: the pair `(raw_dataset_sha256, zero_based_raw_csv_row_number)`.
- The row number is the row position in the verified raw `lenta_dataset.csv.gz` after CSV parsing and before any filtering or reordering.
- The raw-file SHA-256 is recorded before Stage-A analysis; any byte-level change to the raw file invalidates the lock.
- The row key is excluded from every model feature set.
- Rows with missing treatment may be excluded using treatment availability only. Outcome missingness is not used to define experiment-pool membership before the Stage-A lock.
- Feature set: every raw column except `group` and `response_att`; the row key is never added as a feature.
- Non-numeric features are converted to pandas categorical dtype with categories sorted lexicographically. Numeric missing values and categorical missing values are left to LightGBM's native missing-value handling.
- Treatment share π is the treatment mean in the outcome-independent design table and is computed without using the outcome.

Deterministic pool assignment For raw row number `i`, construct

```
unit_key = raw_dataset_sha256 + ":" + str(i)
h = SHA256( UTF8(unit_key) || UTF8("|2027") )
u = unsigned_big_endian_integer(h[0:8]) / 2^64
```

Assign:

- representation pool if $0 \le u < 0.20$;
- reference pool if $0.20 \le u < 0.60$;
- experiment pool if $0.60 \le u < 1$.

Independent reference split The 40% reference pool is deterministically split into two outcome-independent halves using different hash bytes from the same immutable row key. For every row already assigned to the reference pool, define

```
v = unsigned_big_endian_integer(h[8:16]) / 2^64
```

and assign:

- `R1` if $0 \le v < 0.5$;
- `R2` if $0.5 \le v < 1$.

`R1` is used **only to identify the reference-best bin at each depth**. `R2` is used **only to estimate the effect, variance/noise, and mass used by P1/P2 and by eligibility calculations**. No R2 outcome enters the R1 winner choice. This split is specifically intended to remove winner's-curse bias from plugging the maximum of noisy reference-bin effect estimates into the certification-power prediction.

The raw row count, outcome-independent design row count, pool counts, `R1`/`R2` counts, treatment shares, feature list, row-key rule, and SHA-256 of the cached raw dataset file are written to `LENTA_DATA_MANIFEST_v3_1.json` before the experiment run.

G.2.4 A3. FROZEN REPRESENTATION

Fit a two-model LightGBM T-learner on the representation pool only.

For both treatment arms use `lightgbm.LGBMClassifier` with:

```
objective="binary"
n_estimators=500
```

```
learning_rate=0.05
num_leaves=31
min_child_samples=200
subsample=0.8
subsample_freq=1
colsample_bytree=0.8
reg_alpha=0.0
reg_lambda=1.0
random_state=2027
n_jobs=-1
verbosity=-1
```

No early stopping, hyperparameter search, or outcome-dependent feature selection is allowed. The uplift score is

$$s(x) = \widehat{P}(Y = 1 \mid A = 1, X = x) - \widehat{P}(Y = 1 \mid A = 0, X = x).$$

The fitted representation model is serialized and hashed before any experiment-pool outcome is read.

Frozen score bins For each depth r, $J = 2^r$. Bin boundaries are computed from **all reference-pool scores** because score quantiles use no outcome information. Compute quantiles at $0, 1/J, \ldots, 1$ using `numpy.quantile(..., method="linear")`. Replace the first and last boundaries by $-\infty, +\infty$. Bin assignment uses `numpy.searchsorted(edges, score, side="right") - 1`, clipped to `[0,J-1]`.

If any two interior quantile boundaries are identical, that depth and all finer depths are excluded before experiment outcomes are examined.

The maximum depth $R_{\max}$ is the largest remaining $r \leq 8$ for which the expected control support is at least 500 per equal-mass bin in **both** independent reference halves:

$$\min(N_{R1}, N_{R2})\, 2^{-r}(1 - \pi) \geq 500.$$

This preserves approximately the same total reference-support criterion as the earlier unsplit 1,000-control rule while allowing winner identification and effect estimation to remain independent.

G.2.5 A4. Independent frozen reference geometry and predictions

For every depth $r \leq R_{\max}$ and bin b, compute separate reference summaries.

R1 — winner-identification table Using `R1` outcomes only, record treated/control counts and event counts and the effect

$$\widehat{\tau}_{rb}^{R1} = \widehat{p}_{1rb}^{R1} - \widehat{p}_{0rb}^{R1}.$$

At every depth, define the frozen reference-best identity

$$b_r^\star = \arg\max_b \widehat{\tau}_{rb}^{R1},$$

with ties broken by the lowest bin index.

`R1` is used only for this identity choice. Its effect magnitude is **not** plugged into the power prediction.

R2 — independent effect/noise table Using `R2` outcomes only, record for every bin:

- treated/control counts $n_{1rb}^{R2}, n_{0rb}^{R2}$;
- treated/control event counts $x_{1rb}^{R2}, x_{0rb}^{R2}$;
- treated/control rates $p_{1rb}^{R2}, p_{0rb}^{R2}$;
- effect
$$\tau_{rb}^{R2} = p_{1rb}^{R2} - p_{0rb}^{R2};$$
- empirical R2 bin mass p_{rb}^{R2}.

The actionability margin is fixed at

$$\kappa = 0.$$

Define the per-observation randomized-design variance from `R2`

$$(\sigma_{rb}^{R2})^2 = \frac{p_{1rb}^{R2}(1-p_{1rb}^{R2})}{\pi} + \frac{p_{0rb}^{R2}(1-p_{0rb}^{R2})}{1-\pi}.$$

With $z_{.95} = \Phi^{-1}(0.95)$:

P1 — primary deterministic no-fit prediction P1 uses the bin identity selected on `R1` but the independent effect/noise/mass estimate from `R2`:

$$P_1(n,r) = \Phi\left[\frac{\tau_{rb_r^\star}^{R2} - \kappa}{\sigma_{rb_r^\star}^{R2}}\sqrt{\frac{n\,p_{rb_r^\star}^{R2}}{2}} - z_{.95}\right].$$

No calibration constant is fit to `R1`, `R2`, or experiment outcomes. The role separation is:

$$R1:\ \text{winner identity}, \qquad R2:\ \text{independent effect/noise/mass}.$$

P2 — secondary selection-conditioned no-fit prediction For each experiment replicate after upstream selection chooses bin $\widehat{b}$,

$$P_2(n,r,\widehat{b}) = \Phi\left[\frac{\tau_{r\widehat{b}}^{R2} - \kappa}{\sigma_{r\widehat{b}}^{R2}}\sqrt{\frac{n\,p_{r\widehat{b}}^{R2}}{2}} - z_{.95}\right].$$

P2 uses the realized upstream selection identity and is therefore a decomposition diagnostic, not an ex-ante n, r-only prediction; it still uses no certification outcome.

Before any experiment-pool outcome is accessed, write and hash:

- fitted representation model;
- bin-edge file;
- complete `R1` winner-identification table;
- complete `R2` effect/noise/mass table;
- P1 prediction table over the final n, r grid;
- $R_{\max}$;
- dataset/pool/R1-R2 manifest;
- executable analysis code;
- synthetic-smoke-test log;
- full environment lock.

These hashes constitute Protocol A's Stage-A second-stage lock.

G.2.6 A5. Frozen primary eligibility and informativeness gates

The primary analysis excludes cells in which the Normal approximation is predictably stressed by rare-event support.

For the `R1`-selected reference-best identity at each candidate (n, r), use its independent `R2` mass and let the certification-half expected bin size be

$$m_{nr} = (n/2)\,p_{rb_r^\star}.$$

Using the independent frozen `R2` rates, compute expected counts

$$m_{nr}\pi p_1, \quad m_{nr}\pi(1-p_1), \quad m_{nr}(1-\pi)p_0, \quad m_{nr}(1-\pi)(1-p_0).$$

A cell is **primary-eligible** only if the minimum of these four expected counts is at least 20. Eligibility is frozen from the reference table before experiment outcomes are examined.

For the primary E1 endpoint, define an **informative sample size** as an n for which:

1. at least one depth is primary-eligible; and
2. the P1-predicted $R_{80}(n) \geq 1$.

The dataset is declared informative only if at least three sample sizes are informative. If fewer than three are informative, the Lenta confirmation is declared uninformative before the experiment run; no experiment-pool outcome is examined and no substitute dataset is selected after seeing results.

G.2.7 A6. Experiment sampling and upstream selection

Dense, outcome-independent sample-size grid Use half-power spacing from 2^{10} through 2^{17}, truncated only if required by experiment-pool size.

For integer grid index

$$j = 0, 1, \ldots, 14,$$

define

$$n_j = \text{round}\left(2^{10+j/2}\right).$$

Retain every n_j satisfying

$$n_j \leq \min\left\{2^{17}, \lfloor N_{\text{exp}}/2 \rfloor\right\}.$$

Thus, when the experiment pool is large enough, the exact 15-size grid is approximately

```
1024, 1448, 2048, 2896, 4096, 5793, 8192, 11585,
16384, 23170, 32768, 46341, 65536, 92682, 131072
```

where each value is generated by the formula above rather than copied from this display. The denser grid is fixed before any Lenta outcome statistic is computed; it is not adapted to observed effects.

Use 200 replicates per retained sample size.

For each (j, rep), set

```
seed_sample = 2027 + 1000*j + rep
```

and draw n_j experiment-pool rows without replacement with NumPy `default_rng(seed_sample)`.

Split the sampled rows into equal selection and certification halves using a permutation from

```
seed_split = seed_sample + 500000
```

with the first `floor(n_j/2)` permuted rows assigned to selection and the remainder to certification.

For every depth $r \leq R_{\max}$:

1. assign bins using the frozen representation and frozen outcome-free reference edges;
2. on the selection half, compute treated-minus-control difference in means in each bin;
3. bins with fewer than 2 treated or 2 control units receive score $-\infty$;
4. select the bin with the largest estimated effect, ties broken by the lowest bin index;
5. apply that frozen bin rule to the certification half.

No model is retrained inside a replicate.

G.2.8 A7. PRIMARY AND SENSITIVITY CERTIFICATES

Primary Welch-Normal certificate For the selected certification bin, let Y_1, Y_0 be binary outcomes in treated and control arms.

The claim is ineligible to certify unless **each arm contains at least 5 events and at least 5 non-events**. If the guard passes, compute

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0,$$

$$\widehat{\mathrm{SE}} = \sqrt{s_1^2/n_1 + s_0^2/n_0},$$

where s_a^2 is the sample variance with `ddof=1`. Pass the one-sided certificate iff

$$1 - \Phi\left(\frac{\widehat{\tau} - \kappa}{\widehat{\mathrm{SE}}}\right) \leq 0.05.$$

A zero or non-finite standard error fails the certificate.

Newcombe sensitivity Use the 90% two-sided Newcombe hybrid-score interval for the difference $p_1 - p_0$, so its lower endpoint is a one-sided 95% lower confidence bound. The implementation is frozen as:

```
statsmodels.stats.proportion.confint_proportions_2indep(
    count1=x1, nobs1=n1,
    count2=x0, nobs2=n0,
    method="newcomb",
    compare="diff",
    alpha=0.10
)
```

The sensitivity certificate passes iff the lower endpoint exceeds $\kappa = 0$.

G.2.9 A8. ENDPOINTS AND SUCCESS CRITERIA

E1 — certification-resolution proxy For every informative sample size, define

$$R_{80}(n) = \max\{r : \text{cell is primary-eligible and observed certification frequency} \geq 0.80\},$$

with value 0 if no eligible depth passes.

The P1-predicted R_{80} uses the same eligible grid and threshold.

E1 success: exact observed/P1 agreement at at least two thirds of informative sample sizes, and no informative sample size differs by more than one bit.

E2 — calibration conditional on correct localization Among runs whose selected bin equals the frozen reference-best bin, compare observed cell certification frequency with P1 over primary-eligible cells containing at least 20 such runs.

E2 success: run-weighted RMSE ≤ 0.08 and run-weighted absolute mean bias ≤ 0.05.

The confirmation succeeds only if E1 and E2 both succeed. All outcomes are reported regardless of success.

Secondary analyses

1. P2 selection-conditioned calibration and R_{80}.
2. All primary-ineligible cells, reported separately.
3. Newcombe sensitivity.
4. Exact reference-best recovery.
5. Bonferroni-within-candidate Multiple-Comparisons-with-the-Best (MCB) plausible-best set.

For the MCB audit at depth r, with $J = 2^r$, use the independent `R2` bin estimates. Candidate j is retained as plausible best iff

$$\max_{k \neq j} \frac{\widehat{\tau}_k - \widehat{\tau}_j}{\sqrt{\widehat{\mathrm{SE}}_k^2 + \widehat{\mathrm{SE}}_j^2}} \leq z_{1-0.05/(J-1)}.$$

No exact Hsu/Dunnett critical value is substituted after results are seen.

G.2.10 A9. Analysis discipline, smoke test, and immutable second-stage lock

There is one outcome-bearing experiment run and no interim look.

Mandatory synthetic smoke test Before the Stage-A lock, execute `synthetic_outcome_smoke_test_v3.py`.

The smoke test:

- must not read or contain `response_att`;
- runs the representation, hashing/splitting, score-binning, independent R1/R2 geometry, prediction, upstream selection, and certification code paths on synthetic outcomes;
- imports the pinned scikit-uplift / LightGBM / EconML / statsmodels APIs so incompatibilities are detected before any outcome-bearing execution;
- may optionally consume an **outcome-blind** Lenta feature cache, but it refuses any input containing a `response_att` column;
- writes a JSON smoke-test report whose SHA-256 is included in the Stage-A lock.

A smoke-test failure may be repaired before outcome-bearing execution. The protocol, estimands, thresholds, and scientific decision rules may not be changed merely to improve expected results.

Immutable Stage-A lock `LOCK_MANIFEST_v3_1.json` is the immutable protocol-text manifest whose published hash must remain verifiable. It is **never modified in place**.

Before any experiment-pool outcome is accessed, create a new chained manifest, for example `LOCK_MANIFEST_v3_1_STAGE_A.json`, containing SHA-256 hashes of:

- cached raw Lenta dataset;
- `LENTA_DATA_MANIFEST_v3_1.json`;
- fitted T-learner model;
- bin edges;
- `R1` winner-identification table;
- `R2` effect/noise/mass table;
- P1 prediction table;
- analysis code;
- synthetic smoke-test report;
- full `pip freeze` / environment lock.

Use:

```
python finalize_second_stage_lock_v3_1.py \
  --base-manifest LOCK_MANIFEST_v3_1.json \
  --output-manifest LOCK_MANIFEST_v3_1_STAGE_A.json \
  --stage A_SECOND_STAGE \
  --artifact /path/to/raw_lenta_file \
  --artifact /path/to/LENTA_DATA_MANIFEST_v3_1.json \
  --artifact /path/to/frozen_tlearner.joblib \
  --artifact /path/to/bin_edges.json \
  --artifact /path/to/R1_winner_table.csv \
  --artifact /path/to/R2_reference_bin_stats.csv \
  --artifact /path/to/P1_predictions.csv \
  --artifact /path/to/analysis_code.py \
  --artifact /path/to/synthetic_smoke_report.json \
  --artifact /path/to/pip_freeze.txt
```

The finalizer refuses to overwrite the base manifest or an existing output manifest.

A code defect found after outcome-bearing execution is fixed and documented in a dated deviations file; the complete experiment is rerun, and both runs are reported. No endpoint, threshold, sample size, proposer, certificate, reference split, or eligibility rule is altered after experiment outcomes are inspected.

G.3 Protocol B: Reference-supported invalid versus evidence-limited personalization by standard learners on Criteo

Version 3. This protocol supersedes `PROTOCOL_B_Criteo_Invalid_vs_Unprovable_v2.md`.

The analysis is frozen by the SHA-256 hash recorded in `LOCK_MANIFEST_v3.json`. It is prospective with respect to this analysis, not prospective with respect to the Criteo data.

G.3.1 B0. Status of the data

This analysis reuses the already-frozen Criteo representation, reference, and experiment pools from the locked scaling study. Aggregate outcomes from these pools have already been examined in earlier studies.

Therefore:

- this protocol does **not** describe a fresh-data prospective test;
- it is a prospectively specified secondary analysis of previously inspected data;
- none of the three proposers below has previously been fit under this protocol to produce the endpoints below.

No proposer, endpoint, margin, threshold, or label definition may be changed after Protocol-B results are computed.

G.3.2 B1. Question

When standard granularity-seeking learners propose treatment regions, how often are proposals:

1. **reference-supported invalid**;
2. **reference-supported valid but not certifiable at the replicate sample size**; or
3. indeterminate under the high-powered reference pool?

Only category 1 is called over-personalization.

G.3.3 B2. Frozen data, features, margins, and reference-supported labels

Data roles Use the exact frozen Criteo pools from the locked scaling study:

- representation pool;
- reference pool;
- experiment pool.

Outcome: `visit`.

Pre-treatment features: exactly `f0,...,f11`. Exclude `treatment`, `visit`, and post-treatment `exposure` from all model features.

Treatment share is computed from the relevant training split when needed for the propensity nuisance.

Primary and secondary margins The primary margin is frozen numerically at

$$\boxed{\kappa = 0.010252}$$

in absolute `visit` risk-difference units.

This scalar is recorded in `PROTOCOL_B_CONSTANTS_v3.json` and is **not recomputed** during Protocol B. It is the six-decimal operational reference-ATE scalar derived before this protocol from the already-frozen depth-1 Criteo reference summary.

Secondary analysis uses

$$\kappa = 0.$$

Reference-supported validity label For each proposal rule S, apply the frozen rule to the entire reference pool and record:

- region mass $P_{\text{ref}}(S)$;
- treated/control counts n_1, n_0;
- treated/control event counts x_1, x_0;
- effect $\widehat{\tau}_S = x_1/n_1 - x_0/n_0$.

If either arm has fewer than 20 reference units, label the proposal indeterminate.

Otherwise compute a two-sided 95% Newcombe interval for the risk difference:

```
statsmodels.stats.proportion.confint_proportions_2indep(
    count1=x1, nobs1=n1,
    count2=x0, nobs2=n0,
    method="newcomb",
    compare="diff",
    alpha=0.05
)
```

Using the frozen primary margin κ:

- **reference-supported valid** if the lower endpoint exceeds κ;
- **reference-supported invalid** if the upper endpoint is below κ;
- **indeterminate** otherwise.

These are high-powered reference labels, not population ground truth.

Proposal resolution is

$$r(S) = -\log_2 P_{\text{ref}}(S).$$

An empty reference region is labelled indeterminate and assigned no finite resolution.

G.3.4 B3. Replicate grid and train/certification split

Sample sizes:

$$n \in \{2^{14}, 2^{16}, 2^{18}, 2^{20}\}.$$

Replicates:

- 100 for $n \leq 2^{18}$;
- 50 for $n = 2^{20}$.

For each (n, rep),

```
seed = 3027 + 1000*log2(n) + rep
```

Draw n rows without replacement from the frozen experiment pool using NumPy `default_rng(seed)`.

Permute the sample using `default_rng(seed + 500000)` and split it into equal training and certification halves, with the first `floor(n/2)` rows used for training and the remainder for certification.

All proposer fitting and rule selection use the training half only. Certification outcomes are untouched until the proposal rule is frozen.

G.3.5 B4. COMMON IMPLEMENTATION SPECIFICATION

Package/API lock Protocol B uses:

- `econml==0.16.0`;
- `lightgbm==4.7.0`;
- `statsmodels==0.15.0`.

The full resolved environment is captured by `pip freeze` and hashed before the run.

Common LightGBM configuration Where a LightGBM regressor/classifier is used:

```
n_estimators=500
learning_rate=0.05
num_leaves=31
min_child_samples=200
subsample=0.8
subsample_freq=1
colsample_bytree=0.8
reg_alpha=0.0
reg_lambda=1.0
n_jobs=-1
verbosity=-1
```

For binary classifiers use `objective="binary"`. For regressors use `objective="regression"`.

Every model's `random_state` is fixed deterministically from the replicate seed and proposer-specific offset.

Constant propensity nuisance For DR estimators, propensity is a custom sklearn-compatible constant classifier. On `.fit`, it stores the treatment mean π_{train} of its training data. For every input row, `.predict_proba` returns $(1 - \pi_{\text{train}}, \pi_{\text{train}})$. It uses no covariates.

G.3.6 B5. PROPOSERS

B5.1 DR policy tree with the same treatment margin The primary policy tree must optimize the **same utility margin used to judge validity**.

For the primary margin $\kappa = 0.010252$, construct the observed utility outcome on the training half

$$Y_i^{(\kappa)} = Y_i - \kappa A_i.$$

Because

$$E\{Y^{(\kappa)}(1) - Y^{(\kappa)}(0) \mid X = x\} = \tau(x) - \kappa,$$

a treatment recommendation from a policy learner fit to $Y^{(\kappa)}$ targets positive margin-adjusted utility rather than merely $\tau(x) > 0$.

Use `econml.policy.DRPolicyTree` with:

```
model_regression = LGBMRegressor(common parameters,
    random_state=seed+101)
model_propensity = ConstantPropensityClassifier()
cv = 2
mc_iters = None
max_depth = 3
min_samples_split = 0.02
min_samples_leaf = 0.01
max_features = None
min_impurity_decrease = 0.0
min_balancedness_tol = 0.45
honest = True
random_state = seed+102
```

No controls `W` are supplied; `X=f0,...,f11`. Fit with treatment `A` and outcome $Y^{(\kappa)}$.

The proposal region S is the union of leaves for which `.predict(X)` recommends treatment 1. The fitted tree is then applied without refitting to the certification and reference pools.

If the tree recommends treatment to no training rows, record "no proposal". If it recommends treatment to all training rows, retain the all-treated proposal and evaluate it normally.

For the labelled secondary sensitivity $\kappa = 0$, **refit the tree from scratch** on the original outcome Y using the same hyperparameters and seeds. Do not evaluate the primary $\kappa > 0$ tree against the secondary margin or vice versa.

Reference-supported labels and certification always use the original binary `visit` outcome and compare τ_S against the corresponding frozen κ; the utility transformation is used only to align the tree's proposal objective with that same margin.

B5.2 DR-learner threshold rule Use `econml.dr.DRLearner` with:

```
model_propensity = ConstantPropensityClassifier()
model_regression = LGBMRegressor(common parameters,
    random_state=seed+201)
model_final      = LGBMRegressor(common parameters,
    random_state=seed+202)
discrete_outcome = False
multitask_model_final = False
featurizer = None
min_propensity = 1e-6
categories = [0, 1]
cv = 2
mc_iters = None
random_state = seed+203
allow_missing = True
use_ray = False
```

Fit on the training half only.

The frozen proposal is

$$S = \{x : \widehat{\tau}(x) > \kappa\}.$$

Ties at exactly $\widehat{\tau}(x) = \kappa$ are excluded from treatment. The fitted rule is applied without refitting to certification and reference pools.

If no training row is proposed for treatment, record "no proposal".

B5.3 T-learner finest top-q rule Candidate masses:

$$q \in \{2^{-1}, 2^{-2}, \ldots, 2^{-8}\}.$$

Use a two-arm `LGBMClassifier` with the common LightGBM parameters.

To choose q without using certification outcomes:

1. split the training half into two folds with `StratifiedKFold(n_splits=2, shuffle=True, random_state=seed+301)`, stratified by treatment;
2. on each fold, fit treated/control T-learner models on the opposite fold;
3. generate out-of-fold uplift scores for every training row;
4. for each candidate q, define the top-q OOF region by stable descending score rank, ties broken by original frozen Criteo row ID;
5. compute the training-half difference in means within that OOF region;
6. choose the **finest** q whose estimated effect exceeds κ, ties in depth resolved toward finer q;
7. if no q passes, record "no proposal".

After q is chosen, fit a final two-arm T-learner on the full training half using `random_state=seed+302`. Set the frozen threshold to the empirical $(1 - q)$-quantile of the final training-half scores using

`numpy.quantile(..., method="linear")`. The proposal applied to certification/reference pools is

$$S = \{x : s_{\text{final}}(x) \geq t_q\}.$$

No quota re-ranking is performed separately in certification or reference data.

G.3.7 B6. INDEPENDENT CERTIFICATION ON THE CERTIFICATION HALF

For the frozen proposal rule S, use the same primary certificate for all proposers.

Let Y_1, Y_0 be binary `visit` outcomes in treated/control certification units that fall in S.

The claim cannot certify unless each arm contains at least 5 events and at least 5 non-events.

If the guard passes, compute

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0$$

and

$$\widehat{\text{SE}} = \sqrt{s_1^2/n_1 + s_0^2/n_0},$$

where s_a^2 uses `ddof=1`.

Pass iff

$$1 - \Phi\left(\frac{\widehat{\tau} - \kappa}{\widehat{\text{SE}}}\right) \leq 0.05.$$

A zero/non-finite standard error fails.

Newcombe sensitivity Use the 90% two-sided Newcombe interval (`method="newcomb", compare="diff", alpha=0.10`) as a one-sided 95% lower bound. The sensitivity certificate passes iff that lower endpoint exceeds κ.

G.3.8 B7. ESTIMANDS

Report separately by proposer and n, with denominators explicit:

1. proposal rate;
2. reference-supported invalid-proposal rate;
3. reference-supported valid-proposal rate;
4. indeterminate-proposal rate;
5. **evidence-limited rate**:

$$P(\text{reference-supported valid and not certified});$$

6. certified rate;
7. proposal resolution $r(S)$;
8. certified resolution $r(S)\mathbf{1}\{\text{certified}\}$.

For proposer-specific binary proportions, report Wilson 95% intervals.

False certification among reference-supported invalid proposals For each proposer and n, report Wilson 95% intervals for

$$P(\text{certified} \mid \text{reference-supported invalid}).$$

Also report a pooled-across-proposer estimate within each n, but account for shared replicate data by a **replicate-cluster bootstrap**:

- resample replicate IDs with replacement;

- retain all proposer rows belonging to each sampled replicate;
- compute the pooled invalid-certified numerator divided by pooled invalid-proposal denominator;
- 10,000 bootstrap draws;
- seed `9027 + log2(n)`;
- percentile 95% interval.

A descriptive certifier-failure flag is raised only if the **lower** endpoint of this cluster-bootstrap 95% interval exceeds 0.05. This flag is not a primary success criterion.

G.3.9 B8. INTERPRETATION RULES

- "Over-personalization" is used only for reference-supported invalid proposals.
- Reference-supported valid but uncertified proposals are called **evidence-limited**, never errors.
- Indeterminate proposals are reported separately and counted in neither category.
- Reference-supported labels are not called population ground truth.
- No proposer is added, removed, or retuned after results are computed.
- No proposer-specific threshold is altered after seeing reference or certification labels.
- Primary results use $\kappa = 0.010252$; $\kappa = 0$ is a labelled secondary sensitivity.

G.3.10 B9. ANALYSIS DISCIPLINE AND IMMUTABLE EXECUTION LOCK

Protocol B has one planned analysis run.

`LOCK_MANIFEST_v3.json` is immutable and is never overwritten. Before execution, create a new chained manifest, for example `LOCK_MANIFEST_v3_STAGE_B.json`, containing hashes of:

- this protocol;
- `PROTOCOL_B_CONSTANTS_v3.json`;
- analysis code;
- synthetic smoke-test report;
- full environment lock;
- exact frozen Criteo data-manifest / pool-lock files reused from the earlier study.

Use `finalize_second_stage_lock_v3.py` with `--stage B_EXECUTION_LOCK`. The finalizer records the SHA-256 of the immutable base manifest and refuses in-place overwrite.

The common synthetic smoke test must complete successfully before this execution lock is written. In its Protocol-B branch it fits the margin-adjusted DRPolicyTree on synthetic data to verify the pinned EconML API and the $Y - \kappa A$ objective path without using Criteo outcomes.

A code defect found after execution is fixed and logged with a code diff; the full analysis is rerun and both runs are retained.

The status label is:

> prospectively specified analysis on previously inspected randomized data.

It is not described as independent external confirmation.

H HISTORICAL CLINICAL ANALYSES

The clinical analyses are methodological decision illustrations rather than efficacy claims. ACTG-175 retains a population action when a finer selected refinement fails certification. IST supports no subgroup withholding rule. In the panitumumab development release the frozen policy deploys no subgroup action at the prespecified PFS margin; an external age-≥ 75 diagnostic is not multiplicity-robust under the frozen candidate family. PACCE, SCLC, and SPECTRUM are negative or unresolved actionability examples. No real-data abstention is presented as proof of false-selection calibration or personalized clinical benefit.

H.1 ACTG 175 AND IST

The ACTG 175 two-arm contrast contains 1,054 participants (532 ZDV, 522 ZDV+ddI). The recorded population 96-week event-free gain is 0.1477 (95% CI [0.0992, 0.1962]). Discovery proposes a 3.11-bit region with 49 certification subjects, but the simultaneous lower confidence bound for its parent-relative gain is -0.3734 against a 0.03 margin; the final policy therefore retains the population action.

The International Stroke Trial analysis uses the corrected 19,435-person release and a locked 11,654/7,781 discovery/certification split. Aspirin's recorded population effect on six-month alive-and-independent status is 0.01254 ([0.00021, 0.02487]), with a favorable-oriented bleeding difference of -0.00678 ([-0.01035, -0.00321]). The joint benefit-harm certificate supports no subgroup withholding rule. These are historical analysis outputs, not treatment recommendations.

H.2 PANITUMUMAB DEVELOPMENT AND EXTERNAL EVALUATION

Study 20050203 contributes 935 released analytic participants, split into 467 discovery and 468 certification subjects. The frozen candidate registry has 78 entries, of which 65 satisfy support eligibility. The development root estimate is 0.0192 and the discovery-selected age-≥ 75 subgroup has certification estimate -0.0922 with LCB -0.3519; the frozen method and full-search Bonferroni reference therefore deploy inactive policies.

External study 20050181 contains 946 released participants (471 treatment, 475 control). No refitting or new subgroup search is allowed after the development lock. The frozen age-≥ 75 diagnostic has $n = 64$, estimate 0.325703, SE 0.126340, and one-sided 95% LCB 0.117892. Correcting across four frozen external candidate hypotheses lowers this LCB to 0.042524, below the 0.10 clinical margin. Thus the single-primary diagnostic does not establish a multiplicity-robust external treatment-policy benefit. The frozen inactive policy treats zero subjects; its incremental value relative to the same inactive baseline is exactly zero.

H.3 ESTIMATOR ROBUSTNESS ACROSS FIXED ONCOLOGY TARGETS

The post-primary robustness audit uses 5,000 valid arm-stratified bootstrap draws per target for both direct Kaplan-Meier and IPCW estimators, re-estimating censoring weights in every bootstrap replicate. No invalid draws occur and no target changes its certified/unresolved/evidence-against-actionability classification. This supports estimator stability for the fixed targets; it does not establish clinical transportability or independence of overlapping targets.

Statuses use one-sided 95% bounds relative to margin 0.10: certified when the LCB is greater than 0.10, evidence against actionability when the UCB is less than 0.10, and unresolved otherwise. These are per-target statuses, not simultaneous certification of all targets. Evidence against actionability concerns the prespecified margin and is not proof of treatment harm or biological homogeneity.