When Can Experiments Be Reused? Certified Transport, Refusal, and Causal Bridging

Anonymous authors
Paper under double-blind review

Abstract

Past experiments can help answer new causal questions, but similar study reports do not guarantee that their results apply to a new setting. Differences in populations, control conditions, or exposure can change treatment effects even when past estimates are precise. We develop a framework for deciding when experimental results can be reused. It represents uncertainty about these differences as sets and combines this uncertainty with sampling error to bound the error of a target estimate. The bound remains valid even when studies and their weights are selected after observing their results. The framework releases an estimate only when its error bound meets a prespecified tolerance. Otherwise, it reports a range of plausible effects and considers additional experiments. We prove the bound's validity under stated assumptions, characterize what an archive cannot identify, and establish guarantees for choosing additional experiments under a restricted linear-Gaussian model. Controlled simulations show why both mechanism uncertainty and selection must be accounted for: after adaptive selection among 128 sources, simultaneous protection achieves 95.3% coverage, compared with 0.2% for unadjusted intervals. A Many Labs 2 benchmark using meta-analytic baselines illustrates how tighter precision requirements reduce the fraction of targets eligible for release. Together, these results support explicit reuse and refusal decisions, while independent calibration of mechanism uncertainty remains necessary for applications across different interventions.

1 Introduction

Randomized experiments are expensive, yet archives in product development, medicine, education, and development economics are growing. Reuse could make causal decisions faster and cheaper, but each target specifies an intervention, outcome, population, exposure rule, and horizon in a new setting. The practical question is whether the archive supports that target effect and what uncertainty to report when it does not.

Consider a policymaker deciding whether to expand microcredit in an eighth market after observing seven randomized experiments, a setting motivated by the cross-site archive analyzed by Meager (2019). The study reports may use similar language, but sites can differ in take-up, pre-existing credit access, control services, market equilibrium, and the timing of measured outcomes. A pooled estimate can therefore be precise about an average of completed studies while remaining unsupported for the new market. The same failure appears in digital experimentation: two notifications can be textually similar while acting through different recipients, exposure rules, or behavioral mechanisms. Compatibility depends on how these treatment contrasts carry across settings, as well as on sampling precision.

Figure 1 links the archive–target mismatch to our certificate and its transport, refusal, and bridge-design decisions.

Existing paradigms address important pieces of this problem. Meta-analysis quantifies heterogeneity after a collection of effects and an exchangeability model have been chosen, whereas causal transportability identifies target effects when cross-population mechanisms are encoded in a causal model (Bareinboim & Pearl, 2016; Correa et al., 2019). The decision layer remains open when compatibility is uncertain and studies or weights may be selected after outcomes are observed. We address it with a

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.

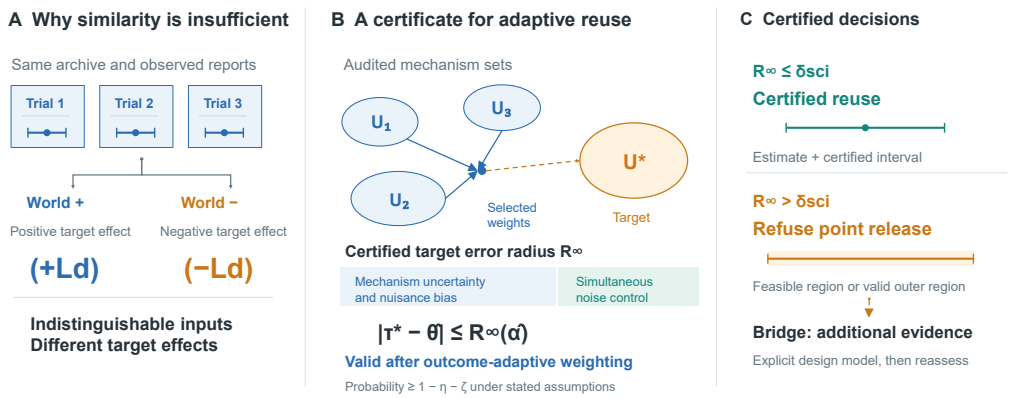

Figure 1: Why causal reuse needs a certificate. Observationally indistinguishable archive inputs can imply opposite target effects (left). Audited mechanism sets and simultaneous noise control yield a selection-safe target-error radius after outcome-adaptive weighting (middle). The radius determines certified reuse, refusal, or bridge design under the stated assumptions (right).

reuse rule that ties observable records to mechanism uncertainty, aligns optimization with the release certificate, and returns an identified uncertainty object after refusal.

We represent mechanism knowledge as sets. For study i, its report, design signature, and scientific audit produce a compact set $U_i \subseteq \mathcal{M}$ intended to contain its latent mechanism m_i. This representation retains uncertainty about omitted moderators even when study reports match. We bound the error of a convex archive composition uniformly over $U_\star \times \prod_i U_i$, choose weights with a small certified radius, and release only when that radius meets a prespecified tolerance. Otherwise the method reports a partial-identification region and considers bridge design.

Because wider mechanism sets can reduce release while protecting validity, release rates should be reported alongside coverage. Bridge value can also be complementary rather than submodular, so a greedy rule needs additional structure. In the stated linear-Gaussian regime, log-determinant information supplies that structure and controls target uncertainty when posterior anisotropy is bounded.

Our contributions are:

(i) **A selection-safe causal reuse certificate.** Calibrated mechanism sets, certified geometric bounds, and a simultaneous noise event give finite-sample validity for arbitrary outcome-adaptive simplex weights. Selection changes the weights, but not the event on which the certificate is valid.

(ii) **A refusal object with explicit identification limits.** Semantic observational equivalence shows why representation similarity cannot identify a transported effect; a finite-metric Lipschitz region gives a sharp refusal calculation, and a canonical Gaussian model isolates the support–noise minimax boundary.

(iii) **A restricted bridge-design guarantee.** Target-width reduction is not generally submodular. Under a nonadaptive linear-Gaussian model, log-determinant information is monotone submodular; its greedy guarantee controls target radius when a posterior condition-number bound holds.

(iv) **Evidence from controlled audits.** Outcome-blind joint-calibration audits use labelled errors; Many Labs 2 evaluates source holdouts under one standardized intervention, and NSW evaluates within-trial reconstruction. Applying the method across interventions requires separate mechanism-set calibration.

Section 2 compares adjacent paradigms; Sections 3–4 define the target, assumptions, and decision rule. Section 5 gives guarantees and limits, with refusal in Section 5.3 and bridge design in Section 5.5. Section 6 tests their empirical consequences, and Section 7 interprets the evidence boundaries. Proofs and implementation details appear in the appendix.

2 LITERATURE REVIEW

Relevant work falls into three connected strands. Meta-analysis and hierarchical synthesis quantify heterogeneity after a study collection and exchangeability model have been chosen (DerSimonian & Laird, 1986; Meager, 2019); causal transportability identifies target queries under explicit cross-population invariances (Pearl & Bareinboim, 2014; Bareinboim & Pearl, 2013; 2016). Orthogonal estimation supplies study-level summaries, while partial identification and post-selection inference address nuisance error and selection (Robins et al., 1994; Chernozhukov et al., 2018; Manski, 1990; Imbens & Manski, 2004; Berk et al., 2013). These approaches leave open the decision layer needed when compatibility is uncertain: a finite-archive rule must calibrate set-valued mechanism uncertainty, protect weights chosen after outcomes, return a sharp refusal object, and restrict bridge guarantees to a stated model. Appendix A gives the full comparison with trial generalization, learned representations, smooth optimal recovery, and active design.

3 DECISION PROBLEM

The target effect is the effect in the new setting. The archive's completed-study average is a distinct quantity. We define the archive, target estimand, and compatibility checks, then state the conditions linking study-level identification to archive-level reuse. These inputs determine whether the rule releases an estimate or returns a region in Section 4.

3.1 EXPERIMENTS, INFORMATION, AND TARGET

The archive contains K studies. Study i has units $W_{i\ell} = (X_{i\ell}, A_{i\ell}, Y_{i\ell})$, $\ell = 1, \dots, n_i$, with potential outcomes $Y_{i\ell}(1), Y_{i\ell}(0)$ and normalized average effect $\tau_i = \mathbb{E}_i\{Y_i(1) - Y_i(0)\}$. The target study $\star$ declares the treatment contrast, control, outcome transformation, population, exposure rule, and horizon before archive outcomes are inspected; its estimand is $\tau_\star$ on the same scale.

Each experiment has a latent mechanism $m_i \in \mathcal{M} \subset \mathbb{R}^p$ and an effect surface $\mu : \mathcal{M} \to \mathbb{R}$ such that $\tau_i = \mu(m_i)$. The algorithm never observes m_i. It observes $Z_i = (r_i, d_i, a_i)$, where r_i is a learned or hand-engineered representation, d_i is a versioned design signature, and a_i records assumptions and audit results. A scientifically specified map returns a compact set $U_i = \Gamma(Z_i) \subseteq \mathcal{M}$. Large U_i records unresolved mechanisms; $U_i = \{m_i\}$ recovers an observed-mechanism model.

The compatibility predicate $\kappa(i, \star)$ equals one only after treatment, control, outcome, exposure, population, horizon, and identification assumptions have either matched or undergone a recorded deterministic normalization. Let $C = \{i : \kappa(i, \star) = 1\}$. Incompatible studies may inform scientific design, but they do not enter the transported point estimate.

Study i supplies $\widehat{\tau}_i = \tau_i + \xi_i + b_i$ with centered noise ξ_i and declared nuisance envelope $|b_i| \le \beta_i$. AIPW construction and its product-error remainder are reviewed in Appendix G; summary-only archives can instead provide an externally audited $(\widehat{\tau}_i, s_i, \beta_i)$ certificate.

3.2 ASSUMPTIONS AS AUDITABLE SCOPE CONDITIONS

The assumptions have distinct roles: within-study identification validates each effect, joint coverage accounts for mechanism uncertainty, smoothness controls extrapolation, and noise certification protects selected weights. Each has a diagnostic and a fallback when certification is unavailable; Appendix Table 1 records those checks and their scope.

Assumption 3.1 (Within-study identification)**.** *Consistency and no interference hold within the declared exposure model. Treatment is randomized with propensity in $[\pi_0, 1 - \pi_0]$, or conditional exchangeability and overlap hold given recorded covariates. Cross-fitted nuisance estimates are trained outside their evaluation folds.*

Assumption 3.2 (Joint mechanism-set coverage)**.** *For a declared $\eta \in [0, 1)$, $\mathbb{P}(\mathcal{G}_U) \ge 1 - \eta$, where $\mathcal{G}_U = \{m_\star \in U_\star,\ m_i \in U_i \text{ for all } i \in C\}$. The probability may refer to an external calibration sample; deterministic scientifically justified sets correspond to $\eta = 0$.*

Assumption 3.3 (Smooth effect surface)**.** *$\mathcal{M}$ is compact and convex. The function μ is continuously differentiable, $\|\nabla\mu(m)\| \le L$, and $\|\nabla\mu(m) - \nabla\mu(m')\| \le H\,\|m - m'\|$ for all $m, m' \in \mathcal{M}$, where $L, H < \infty$ are declared or sensitivity-indexed constants.*

Assumption 3.4 (Certified summary noise)**.** *Conditional on designs, nuisance-training folds, and external set calibration, the ξ_i are independent and satisfy $\mathbb{E}\{\exp(t\xi_i) \mid \mathcal{D}\} \le \exp(t^2 s_i^2/2)$ for $t \in \mathbb{R}$. The proxies s_i are prespecified or covered by the conditioning information; a certified joint proxy matrix may replace independence for fixed weights.*

Appendix Table 1 records each assumption's role, diagnostic, and fallback. Assumption 3.2 is substantive: if no audit can rule out a mechanism that changes the target effect, the valid set must include it, and the certificate may correctly refuse.

The calibration unit must match the decision. Marginal coverage of one source set does not imply joint coverage of an archive and its target. An independent audit archive can calibrate the maximum standardized mechanism error over the full candidate menu, including the target; adding unaudited candidates requires a new argument. This concerns mechanism measurements and their sampling scheme, not agreement between a prediction and an outcome used later for evaluation.

4 Operational Transport, Refusal, and Bridging

Figure 1 becomes an operational rule in two steps. We first combine mechanism uncertainty and sampling error into a release radius, whose validity is established by Theorem 5.2. When this radius exceeds the scientific tolerance, we report the compatible target region and consider bridge design under an explicit planning model.

4.1 Worst-case radius and release

For $\alpha \in \Delta_C = \{\alpha \in \mathbb{R}_+^{|C|} : \sum_{i\in C} \alpha_i = 1\}$, write $\bar{m}_\alpha = \sum_i \alpha_i m_i$. The three quantities that drive the decision are

$$G(\alpha; \mathbf{m}) = L \left\| m_\star - \bar{m}_\alpha \right\| + \frac{H}{2} \sum_{i\in C} \alpha_i \left\| m_i - \bar{m}_\alpha \right\|^2,$$

$$B_U(\alpha) = \sup_{\mathbf{m} \in U_\star \times \prod_{i\in C} U_i} G(\alpha; \mathbf{m}) + \sum_{i\in C} \alpha_i \beta_i, \tag{4.1}$$

$$R_\infty(\alpha; \zeta) = B_U(\alpha) + \sqrt{2\log(2|C|/\zeta)} \sum_{i\in C} \alpha_i s_i. \tag{4.2}$$

Thus B_U combines support, curvature, and nuisance bias, while the final ℓ_1 term protects arbitrary outcome-adaptive weight selection. For design-measurable weights, the sharper radius is $R_2(\alpha; \zeta) = B_U(\alpha) + \{2\log(2/\zeta) \sum_i \alpha_i^2 s_i^2\}^{1/2}$. Appendix B gives the separated geometry, computational conditions, and threshold elicitation details.

The ideal rule computes $\widehat{\alpha} \in \arg\min_{\alpha\in\Delta_C} R_\infty(\alpha; \zeta)$ and $\widehat{\theta} = \sum_i \widehat{\alpha}_i \widehat{\tau}_i$. For a prespecified target-scale tolerance δ_{sci}, it releases $\widehat{\theta}$ with interval $[\widehat{\theta} \pm R_\infty(\widehat{\alpha}; \zeta)]$ only when $C \neq \varnothing$ and $R_\infty(\widehat{\alpha}; \zeta) \le \delta_{\text{sci}}$. The finite-budget implementation uses multi-start SLSQP with feasible fallbacks and evaluates a certified error radius at the returned weights; this radius determines its reported interval and release decision, without a global-optimality guarantee (Appendix L).

4.2 Refusal and bridging

On the simultaneous event, each source lies in $I_i = [\widehat{\tau}_i \pm \{\beta_i + \sqrt{2\log(2|C|/\zeta)} s_i\}]$. Let $\mathcal{F}_{L,H}$ be the surfaces satisfying Assumption 3.3. The model-relative feasible target set is

$$\Theta_U = \left\{ \mu(m_\star) : \begin{array}{l} \mu \in \mathcal{F}_{L,H},\ m_j \in U_j \ (j \in C \cup \{\star\}), \\ \mu(m_i) \in I_i \ (i \in C) \end{array} \right\}. \tag{4.3}$$

The refusal output is $[\inf \Theta_U, \sup \Theta_U]$. General nonlinear endpoints require global optimization or a labeled outer relaxation; known finite locations under Lipschitz smoothness admit the sharp linear program of Theorem 5.3. A bridge candidate declares a design and an expected mechanism measurement. Because direct interval-width reduction need not be submodular, unrestricted problems use exact or certified planning; the linear-Gaussian regime uses the proved log-determinant surrogate of Theorem 5.6.

Algorithm 1 states the ideal decision rule.

Algorithm 1 Ideal causal reuse with certified refusal and bridging

Require: Frozen target query; certified archive summaries and mechanism sets; tolerance δ_{sci}; bridge budget B.

1: Audit compatibility and set $C = \{i : \kappa(i, \star) = 1\}$; record Assumption 3.2's calibration protocol.
2: If $C \neq \varnothing$, minimize a certified upper bound on R_∞ over Δ_C.
3: If the optimized radius is at most δ_{sci}, **return** TRANSPORT: estimate, interval, weights, and audit.
4: Otherwise compute (4.3), or a labeled verified outer relaxation.
5: Choose an anchor if unbounded; else use proved logdet greedy only under its model, and certified planning otherwise.
6: **return** REFUSE: region, bridge design, and failed diagnostics.

5 MAIN RESULTS

The results below instantiate the decision rule in Section 4. Theorem 5.1 identifies what semantic retrieval leaves unresolved. Theorem 5.2 converts audited mechanism sets into a valid release certificate; Theorem 5.3 characterizes refusal in a finite metric model, and Theorem 5.4 separates support ambiguity from sampling noise in a canonical regime. Theorems 5.5–5.6 then delimit when greedy bridge design is justified. Appendix F contains the proofs.

5.1 IDENTIFICATION LIMITS OF SEMANTIC SIMILARITY

Let an r-only procedure be any randomized rule whose input is the archive data, archive representations, and target representation, but not target outcomes or a mechanism set that separates representation-equivalent mechanisms.

Theorem 5.1 (Semantic observational equivalence)**.** *Suppose two mechanisms $m_a, m_b \in \mathcal{M}$ satisfy $r(m_a) = r(m_b)$ and $\|m_a - m_b\| = d > 0$. There exist two L-Lipschitz effect surfaces and two data-generating processes that induce exactly the same distribution for every input of an r-only procedure, while their target effects differ by $2Ld$. Consequently, for every r-only point estimator T,*

$$\sup_{P \in \{P_+, P_-\}} \mathbb{E}_P \, |T - \tau_\star(P)| \geq Ld. \tag{5.1}$$

The construction places every archived study at m_a, assigns the same archive effect in both worlds, and places the target at m_b. The target report is identical because $r(m_a) = r(m_b)$; the two admissible surfaces take values $+Ld$ and $-Ld$ at m_b. No amount of archive sample size distinguishes the worlds.

Theorem 5.1 separates retrieval accuracy from identification. An encoder can help locate relevant studies, but mechanisms with the same observed record require an additional measurement or experiment to distinguish their effects. Until then, U_i retains that ambiguity.

5.2 FINITE-SAMPLE OPERATIONAL CERTIFICATE

To control the ambiguity in Theorem 5.1, the barycentric decomposition bounds causal approximation by G: target support plus nonlinear archive dispersion (Appendix F). Simultaneous noise control then permits outcome-adaptive composition.

Theorem 5.2 (Selection-safe operational certificate)**.** *Under Assumptions 3.2–3.4, suppose $C \neq \varnothing$ and let $\widehat{\alpha}$ be any random element of Δ_C, possibly selected after observing $(\widehat{\tau}_i)_{i \in C}$. Then, with probability at least $1 - \eta - \zeta$,*

$$\left| \tau_\star - \sum_{i \in C} \widehat{\alpha}_i \widehat{\tau}_i \right| \leq R_\infty(\widehat{\alpha}; \zeta). \tag{5.2}$$

Therefore Algorithm 1 may optimize over the entire continuous simplex and retain the same coverage statement; no post hoc Bonferroni argument over an uncountable weight family is required.

The joint selective-error bound in (F.7) is unconditional. Conditional-on-release coverage needs additional selective-inference structure. For design-measurable weights independent of evaluation noise, the same coverage holds with the sharper R_2 radius; a joint sub-Gaussian proxy replaces its

diagonal variance by $\alpha^\top \Sigma \alpha$. Appendix C states this corollary and its conditions. The contribution is archive-level accounting of mechanism ambiguity, curvature, nuisance bias, and selection, rather than experiment-level orthogonal estimation. Genuine s_i proxies and calibrated sets support the finite-sample claim; plug-in substitutes support asymptotic or approximate language.

Asymptotic counterpart. Under the cross-fitted linear expansion and Lindeberg conditions stated in Appendix C, $[\widehat{\theta}_n \pm \{z_{1-\zeta/2}\sqrt{\widehat{V}_n} + A_n\}]$ has asymptotic coverage at least $1-\zeta$ whenever the causal approximation obeys $|a_n| \le A_n$. More source precision shrinks variance, not A_n; unsupported transport can therefore become conventionally precise without becoming valid.

5.3 Sharp refusal in a finite metric model

When the certificate is too wide, the next question is which target effects remain compatible. In a finite metric model, suppose locations $x_1, \dots, x_K, x_\star$ are known and each source effect lies in a jointly valid interval $[\ell_i, u_i]$. Let $\mathcal{J} = ([\ell_i, u_i])_{i=1}^K$ and let $\mathcal{P}(\mathcal{J})$ enforce these intervals and all pairwise L-Lipschitz constraints (Appendix C).

Theorem 5.3 (Sharp Lipschitz target region)**.** *If $\mathcal{P}(\mathcal{J})$ is nonempty, the set of target values compatible with the source intervals and an L-Lipschitz surface on $\mathcal{M}$ is exactly*

$$\Theta_{\mathrm{Lip}}(\mathcal{J}) = [\underline{\theta}, \overline{\theta}], \qquad \underline{\theta} = \min_{t \in \mathcal{P}(\mathcal{J})} t_\star, \quad \overline{\theta} = \max_{t \in \mathcal{P}(\mathcal{J})} t_\star. \tag{5.3}$$

Both endpoints are linear programs after each absolute-value constraint is split into two inequalities. Every value in this interval is attained by an L-Lipschitz function on the full metric space. If the source intervals jointly cover the true source effects with probability at least $1-\zeta$, then (5.3) covers the true target effect with at least the same probability under the Lipschitz model.

The coverage statement is statistical; sharpness is relative to the stated Lipschitz model. Appendix C gives the exact-source closed form and the extension argument.

5.4 A matching minimax regime, and its boundary

The refusal region in Theorem 5.3 reflects both unsupported transport and sampling uncertainty. To separate their unavoidable costs, consider one archive anchor m_0, a target $m_\star$ at distance d, and independent $Z_i \sim \mathsf{N}(\mu(m_0), s_i^2)$ with information $\mathcal{I} = \sum_i s_i^{-2}$. The target is $\mu(m_\star)$ over L-Lipschitz functions on this two-point subspace.

Theorem 5.4 (Canonical support–noise minimax characterization)**.** *For absolute loss, let*

$$\mathcal{R}^* = \inf_T \sup_{\mu:\, \mathrm{Lip}(\mu) \le L} \mathbb{E}_\mu |T(Z_1, \dots, Z_K) - \mu(m_\star)|. \tag{5.4}$$

Then

$$\max\left\{Ld, \frac{1}{8\sqrt{\mathcal{I}}}\right\} \le \mathcal{R}^* \le Ld + \sqrt{\frac{2}{\pi \mathcal{I}}}. \tag{5.5}$$

Hence $\mathcal{R}^* \asymp Ld + \mathcal{I}^{-1/2}$ with universal constants.

The inverse-variance archive mean attains the upper bound. The lower bound combines two experiments: one keeps the archive law fixed while moving the target by $\pm Ld$; the other uses a two-point Gaussian location test. Thus neither infinite within-study precision nor perfect semantic retrieval removes unsupported target ambiguity.

The matching result characterizes estimation risk in the equal-anchor setting. With K equal-noise studies at the same known anchor, $s_i = s$ and zero nuisance bias, the simultaneous radius is $Ld + s\sqrt{2\log(2K/\zeta)}$, whereas equal design weights give $R_2 = Ld + s\sqrt{2\log(2/\zeta)/K}$. At fixed confidence, the latter has the canonical sampling rate. The former pays for arbitrary outcome-dependent weight selection and retains a larger radius as K grows. Validity and statistical efficiency are separate properties.

The convex-hull condition alone is insufficient: zero target-to-hull distance can coexist with unidentified interior effects. Appendix C gives the counterexample and the resulting need for curvature, fill distance, or a function-class modulus.

5.5 BRIDGE DESIGN: A NEGATIVE RESULT AND A STRUCTURED REPAIR

Refusal motivates new measurements. Greedy bridge selection requires additional structure: Theorem 5.5 shows the obstruction, while Theorem 5.6 gives a structured alternative. For exact bridge outcomes, let $w(S)$ be the target-projection width after bridge set S, and $F_{\mathrm{PI}}(S) = w(\varnothing) - w(S)$.

Theorem 5.5 (Target-width reduction is not generally submodular)**.** *There is a two-dimensional linear problem with two measurements such that* $F_{\mathrm{PI}}(\{1\}) - F_{\mathrm{PI}}(\varnothing) = 0 < 2 = F_{\mathrm{PI}}(\{1,2\}) - F_{\mathrm{PI}}(\{2\})$. *Thus direct PI-width reduction need not have diminishing returns.*

The witness uses two measurements that are individually uninformative but jointly identify the target. A valid surrogate applies under the nonadaptive linear-Gaussian model in Appendix C: with posterior covariance V_S, define $F_{\log\det}(S) = \frac{1}{2}\log\{\det(V_0)/\det(V_S)\}$.

Theorem 5.6 (Submodular information bridge)**.** *The set function* $F_{\log\det}$ *is normalized, monotone, and submodular. If* S_g *is the size-*B *greedy set and* S^* *is optimal, then* $F_{\log\det}(S_g) \geq (1 - e^{-1})F_{\log\det}(S^*)$.

Log-determinant information is global, so it controls the target radius only with the testable posterior condition-number bound stated in Appendix C. If that diagnostic fails, the fallback is target-aware exact design.

6 EXPERIMENTS

Each audit targets a distinct claim: hidden mechanism shifts examine semantic nonidentification and refusal; adaptive selection tests the finite-sample certificate in Theorem 5.2; joint calibration probes the joint-coverage assumption; and Many Labs 2 evaluates source and effect-family extrapolation against model-based baselines. We use two controlled synthetic audits and two leakage-free holdouts, unsealing held-out outcomes only after fitting and release decisions are frozen. The main metrics are coverage or noisy-reference inclusion, interval width, release, and prediction error, as appropriate to each audit. Figures 2–3 show the results; Appendix Table 2 maps each audit to its claim and assumptions, and the appendix reports the full protocols, ablations, and perturbation checks.

6.1 THEOREM-LINKED SYNTHETIC AUDIT

Panel A uses four zero-effect sources with aggregate standard error 0.1 and 20,000 draws per hidden shift $h \in \{0, 0.02, \ldots, 0.50\}$; recorded source and target descriptors remain identical while the declared-set and semantic-only intervals are compared. Panel B uses 50,000 draws per archive size, selects the largest of K independent standard-normal absolute summaries, and compares reused pointwise with simultaneous Gaussian intervals. Panel C uses 40,000 common Gaussian draws at each point of a 27×27 grid in mechanism ambiguity Ld and sampling scale $\mathcal{I}^{-1/2}$. Panel D uses 200 six-dimensional linear-Gaussian instances, 12 candidate bridges, noise standard deviation 0.5, and budget four; greedy and random selection are compared with all 495 subsets.

Empirical declared-set coverage is at least 0.9742, whereas the semantic-only interval falls to 0.1510 at the boundary. At $K = 128$, pointwise and simultaneous coverage are 0.00158 and 0.9526. Normalized risk is 0.796–1.171 times the predicted support–noise scale; greedy's minimum exhaustive-search ratio is 0.947, versus 0.704 on average for random subsets. These Monte Carlo checks assess coverage, risk, and design efficiency in the stated Gaussian models.

6.2 OUTCOME-BLIND MECHANISM-SET CALIBRATION AUDIT

We generate independent bivariate-Gaussian proxy errors for four archive sources, six candidate bridges, and one target, giving 11 entities in a frozen decision universe. A row-wise maximum radius is calibrated on 199 independent archives and evaluated on 2,000 new archives; no treatment, target, or future bridge outcomes enter calibration. We compare joint-max with entity-wise calibration, then multiply only test errors by 1.5 to create distribution shift. Joint coverage is 0.951 under the matched distribution, versus 0.544 for marginal calibration, and falls to 0.348 after the shift. This illustrates the exchangeability boundary under the tested synthetic error model.

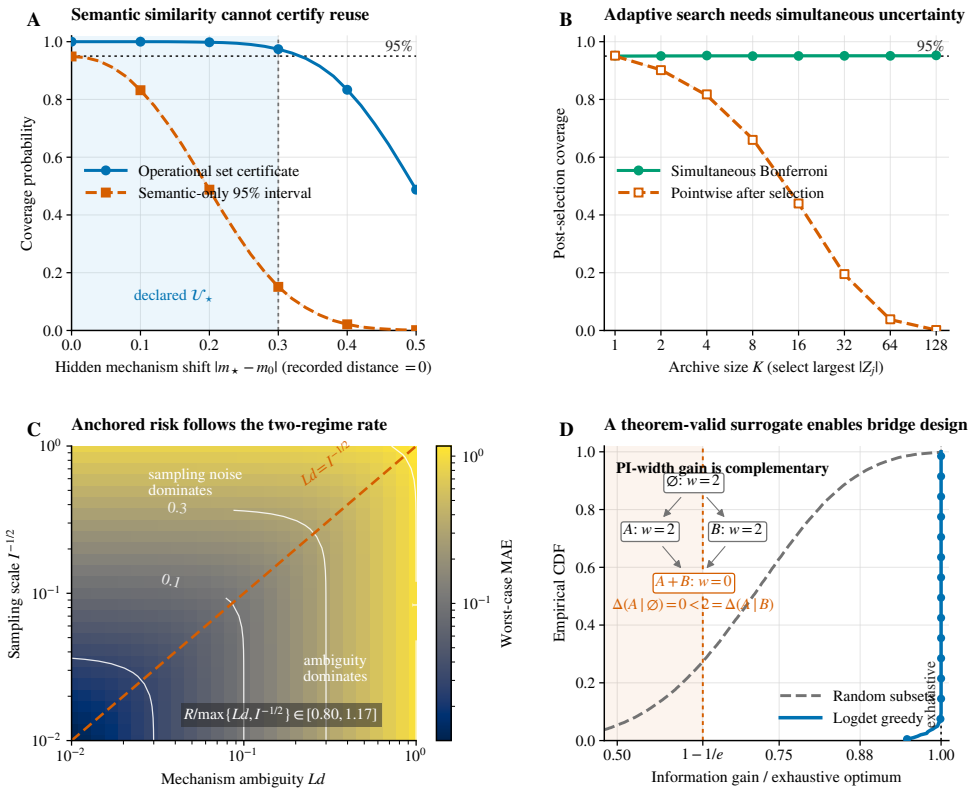

Figure 2: Theorem-linked synthetic audit. **A**: semantic proximity fails under hidden mechanism shift, while the declared-set interval remains valid on its calibrated region. **B**: pointwise coverage collapses after adaptive search; Gaussian Bonferroni protection remains valid. **C**: risk follows the canonical support–noise rate. **D**: log-determinant greedy nearly matches exhaustive information search; the inset gives an exact PI-width complementarity witness. Ribbons are 95% Monte Carlo intervals; protocols appear in Appendix H.

6.3 Independent-source transport: Many Labs 2 framing

The real-data holdout uses 57 complete Many Labs 2 source samples and 7,228 randomized responses under the "Cheap" and "Expensive" price contexts. Each leave-one-source-out split seals the held-out outcomes before fitting or release and scores the frozen prediction against that source's noisy risk-difference estimate. We compare fixed-effect and DerSimonian–Laird random-effects pooling, including release and refusal at the illustrative outcome-blind half-width threshold 0.15. Figure 3 displays the source-risk variation, leave-one-source-out predictions, and release frontier.

The methods have MAE 0.0800 and 0.0801; their predictive intervals include 56 of 57 noisy held-out estimates and have mean widths 0.4041 and 0.4149. At threshold 0.15, fixed and random effects release 10 and 6 sources. The audit evaluates leakage-free source-level prediction under one standardized intervention.

6.4 Cross-effect-family holdout and baseline comparisons

The second holdout uses author-level Many Labs 2 original-effects summaries: 23 eligible families and 28 study-level effects. We hold out all rows of five complete families (ten targets), then each eligible family (28 scored targets). Target effects remain sealed until scoring. Baselines are transport meta-regression, hierarchical meta-analysis, robust partial identification, and study- and family-level split conformal intervals.

Noisy-reference inclusion ranges from 0.800–1.000 in the five-family panel and 0.929–1.000 across all families. Transport and hierarchical intervals release at most $2/10$ and $2/28$ targets; robust and

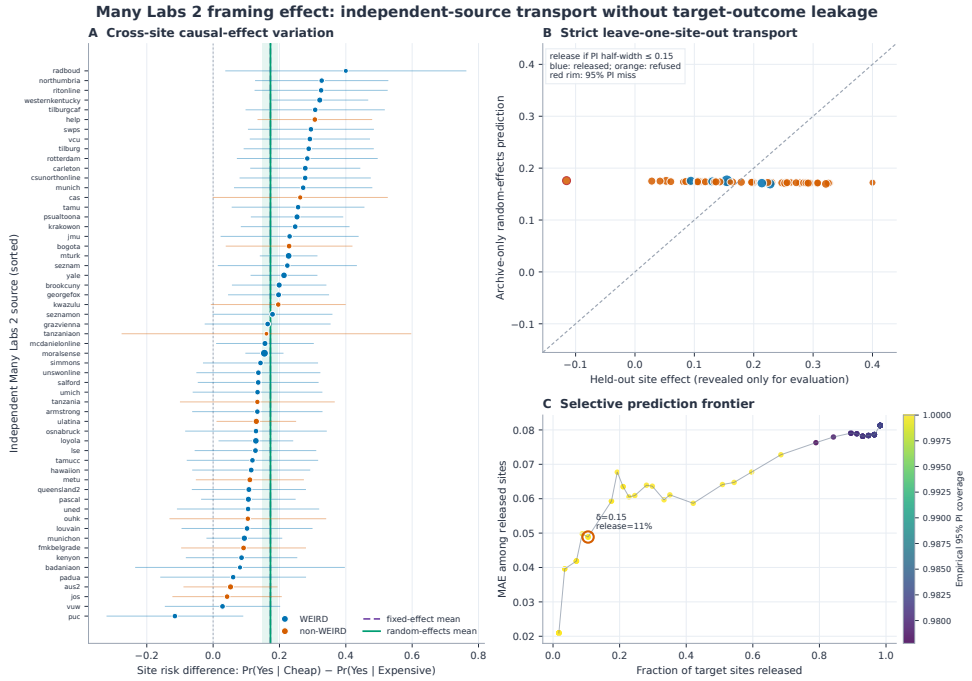

Figure 3: Leakage-free Many Labs 2 benchmark. **Left**: source risk differences and global meta-analytic means. **Upper right**: strict leave-one-source-out predictions under an illustrative outcome-blind half-width threshold. **Lower right**: the complete risk–release frontier. These baselines serve as model-based comparators.

conformal intervals release none. Flipping held-out effects leaves predictions and radii unchanged. The checks show that cross-family extrapolation often requires wide intervals or refusal. These figures quantify inclusion of noisy references; a causal coverage interpretation requires calibrated mechanism sets, an independent calibration archive, and noise-free target truth.

7 Discussion

The release rule returns a point estimate when the certified radius meets the tolerance and a model-relative region otherwise. We report both all-target and selective performance so that refusal remains part of the evaluation. The published-summary microcredit pilot refuses at all four frozen tolerances (0.10, 0.20, 0.40, and 0.80); unresolved design-coordinate provenance precludes a calibrated transport claim.

The results leave three practical questions. Mechanism-set calibration for new interventions needs an outcome-blind archive with independent labels, and the amount of cross-intervention data needed for that calibration is unknown. The smoothness pair (L, H) calls for sensitivity analysis that links weaker conditions to usable refusal regions. The bridge guarantee is nonadaptive and linear-Gaussian; adaptive acquisition calls for sequential inference and conditions supporting greedy guarantees (Howard et al., 2021; Golovin & Krause, 2011). An external audit should freeze tolerances and release rules before scoring unseen targets against direct experimentation.

In Figure 3, the source-risk panel describes heterogeneity within one standardized intervention, the leave-one-source-out panel evaluates prediction against a noisy held-out reference, and the release frontier shows how eligibility changes as the half-width requirement tightens. The panels connect the holdout to Theorem 5.2: prediction may be informative even when the certified radius is too wide for release. After refusal, the next step is to evaluate a bridge experiment under the declared design model. The practical sequence is therefore explicit: audit mechanism sets, compute the certified radius, refuse when the tolerance is missed, and reserve new experiments for targets whose uncertainty remains decision-relevant.

AI USE STATEMENT

Generative AI tools assisted with theoretical framing and draft proof text, synthetic-data and experiment design, implementation and debugging of analysis code, figure scripts, literature search, data organization, result interpretation, translation, and manuscript drafting and editing.

Validation included executable tests, independent arithmetic checks of the microcredit pilot, LaTeX compilation, and comparisons between formalized statements and manuscript claims, as recorded in the supplementary material. The authors manually reviewed the manuscript against the recorded analysis outputs, checked the reported values and displayed figures and tables, and inspected the final compiled PDF for layout and cross-reference errors. The authors take responsibility for the AI-assisted material and for the final scientific decisions, claims, proofs, and presentation.

ETHICS STATEMENT

This paper does not recruit participants, collect new human-subject data, contact individuals, or administer interventions. It uses publicly accessible or deposited secondary data from the Many Labs 2 and NSW studies, published study-level summaries, and synthetic simulations. Some supplementary analyses process publicly released individual-level records for outcome-blind splitting and aggregate estimation; we do not report direct identifiers, attempt re-identification, or link these records to external personal information. The reported microcredit pilot uses published aggregate effect estimates and sampling uncertainties; the OpenICPSR raw microcredit archive was inventoried but is not redistributed in the supplementary materials, and we do not assume that the package-level license resolves the trial-specific data-use restrictions.

Any deployment of the reuse rule should undergo domain-specific review of privacy, measurement validity, data-use permissions, legal requirements, and potential effects on affected populations. The framework's refusal decision is a statistical safeguard against unsupported transport; it does not establish policy eligibility, replace human-subjects or data-governance review, or replace the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

This manuscript is accompanied by an anonymous, downloadable supplementary source archive. The main paper states the decision problem, estimands, release and refusal rule, information boundary, and the distinction between a valid certificate and a model-relative diagnostic in Sections 3 through 6. The appendix contents provide a direct map from each claim to its supporting material: Appendix B states the assumptions, admissible information, optimization problem, and computational fallbacks; Appendix C gives the extended theoretical statements and their scope conditions; Appendix F contains the proofs of the main results; and Appendix G records the experiment-level expansion, nuisance-bias accounting, and the conditions under which an asymptotic substitute may be used. The exact deterministic algebraic and interval statements that are machine checked are separately provided in `LeanProofs/CausalAtlasBridge.lean`, with the axiom audit in `LeanProofs/AxiomAudit.lean` and the formalization boundary in the accompanying Lean README. Results outside that deterministic core remain supported by the conventional mathematical arguments in the appendices; the Lean files are not presented as a formalization of the statistical theorems.

The experimental protocols are documented independently of the reported numbers. Appendix H specifies the data-generating processes, randomization devices, Monte Carlo units, and audit checks for the theorem-linked synthetic panels. Appendix D maps those panels to the claims and assumptions they are intended to probe. Appendix L describes the complete calibration-to-transport-to- bridge workflow, while Appendices N and P document the outcome-blind calibration audit, sensitivity surfaces, matched-certificate weight comparison, and geometric-bound comparison. The source file `paper_original/run_operational_experiments.py` implements the theorem-linked synthetic panels and writes their plotted quantities and unrounded summaries to the recorded artifact directory. The appendix and the corresponding manifest identify the panel-specific seeds, replication counts, and source files rather than relying on values copied from the manuscript.

The data-processing provenance is recorded separately for each empirical component. Appendix J and Appendix M describe the Many Labs 2 source-level holdout, its effect-family extension, the estimand and eligibility rules, and the separation of fitting from scoring. The source file `paper_original/run_manylabs2_framing.py` verifies the commit-pinned public input before constructing the cleaned source summaries; the supplementary files `DATA_SOURCES.md`, `current_method/data/provenance.json`, and `current_method/data/manylabs2_original_effects_provenance.json` record the upstream locations, revisions, byte counts, and checksums. The NSW processing and within-trial reconstruction are kept in the separately labelled legacy audit described in Appendix I; its participant-disjoint covariate proxy and semisynthetic response-surface audit are described in the same appendix and are not treated as independent cross-intervention calibration. The microcredit material is indexed in `MICROCREDIT_OPENICPSR_INVENTORY_20260925.md`, while the estimand, design-compatibility, and descriptive leave-one-study-out procedures are specified in `MICROCREDIT_INVITATION_ITT_PROTOCOL_20260925.md`, `MICROCREDIT_LOSO_PILOT_PROTOCOL_20260925.md`, and `MICROCREDIT_PUBLISHED_V1_REPORT_20260925.md`. These documents give the source cells, unit conversions, eligibility gates, and qualification rules for the published-summary pilot without redistributing the raw OpenICPSR trial files.

The archive-level organization is documented in `README.md`, `SUPPLEMENTARY_README.md`, and `CURRENT_ARTIFACTS.md`. The last file links paper figures, tables, appendix objects, code modules, and saved outputs; `provenance/release_manifest.json` pins the delivered bytes with SHA-256 hashes, and the run manifests retain the corresponding software versions, configurations, seeds, and output hashes. This makes it possible to distinguish a byte-level integrity check from a fresh rerun and to trace a displayed result back to its source file. The archive excludes local caches, scratch directories, private raw microcredit files, and machine-specific paths. Public upstream data remain subject to their original terms, and the package-level license information is not treated as resolving every trial-specific data-use or consent restriction.

Finally, the supplementary reports mark the evidentiary boundary for every component. Many Labs 2 is a source-holdout benchmark under one standardized intervention; NSW is a within-trial reconstruction with a semisynthetic known-truth extension; and the microcredit analysis is a qualified descriptive feasibility pilot based on published summaries. The labelled synthetic calibration audit does not establish real cross-intervention mechanism-set calibration, and noisy held-out randomized estimates are not treated as exact causal truth. These labels, together with the refusal and qualification gates in the appendices and supplementary reports, are part of the reproducibility record and prevent a successful rerun of a descriptive artifact from being mistaken for evidence of a stronger transport guarantee.

REFERENCES

T. B. Armstrong and M. Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

E. Bareinboim and J. Pearl. Meta-transportability of causal effects: a formal approach. In *Proceedings of AISTATS*, volume 31 of *Proceedings of Machine Learning Research*, pp. 135–143, 2013.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

S. R. Cole and E. A. Stuart. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115, 2010.

J. D. Correa, J. Tian, and E. Bareinboim. Adjustment criteria for generalizing experimental findings. In *Proceedings of ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1361–1369, 2019.

I. J. Dahabreh, S. E. Robertson, E. J. Tchetgen Tchetgen, E. A. Stuart, and M. A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

I. J. Dahabreh, S. E. Robertson, L. C. Petito, M. A. Hernán, and J. A. Steingrimsson. Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072, 2023.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188, 1986.

D. Golovin and A. Krause. Adaptive submodularity: theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

W. Guo, M. Yin, Y. Wang, and M. Jordan. Partial identification with noisy covariates: a robust optimization approach. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 318–335, 2022.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

G. W. Imbens and C. F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

R. A. Klein, M. Vianello, F. Hasselman, et al. Many Labs 2: investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4): 443–490, 2018.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

Q. Lanners, C. Rudin, A. Volfovsky, and H. Parikh. Data fusion for partial identification of causal effects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 150638–150702, 2025.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

C. F. Manski. Nonparametric bounds on treatment effects. *American Economic Review: Papers and Proceedings*, 80(2):319–323, 1990.

E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40 (12):837–842, 1934.

R. Meager. Understanding the average impact of microcredit expansions: a Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11 (1):57–91, 2019.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978.

J. Pearl and E. Bareinboim. External validity: from do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

M. Phan, D. Arbour, D. Dimmery, and A. Rao. Designing transportable experiments under S-admissability. In *Proceedings of AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2539–2547, 2021.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. Zhang, H. Bao, P. Wang, A. Yan, X. Liu, and J. A. Evans. Building an Atlas of Social Experiments to Link Studies, Reconcile Conflicts, and Bridge Gaps. arXiv preprint, 2026.

APPENDIX CONTENTS

Appendix A: Extended Related Work and Positioning 15
Appendix B: Assumption Audit and Extended Operational Method 16
Appendix C: Extended Theoretical Statements 17
Appendix D: Extended Empirical Audits 19
Appendix E: Notation, Geometry, and Claim Scope 21
Appendix F: Proofs of the Main Results 21
Appendix G: Experiment-Level Estimation and Bias Accounting 26
Appendix H: Synthetic Protocols and Reproducibility 27
Appendix I: Legacy Synthetic and NSW Audits 28
Appendix J: Independent-Source Benchmark and Extension Protocol 28
Appendix K: Machine-Checked Deterministic Core 32
Appendix L: Complete Calibration-to-Transport-to-Bridge Workflow 32
Appendix M: Real-Data Supplement: Country Holdout and Competing Estimators 36
Appendix N: Supplementary Selection and Calibration Audits 38
Appendix O: Source-Level Effect Estimates 39
Appendix P: Matched-Certificate Weight and Geometric-Bound Ablations 39

A Extended Related Work and Positioning

Evidence synthesis and external validity. Classical meta-analysis estimates a common or heterogeneous distribution of completed-study effects (DerSimonian & Laird, 1986). Hierarchical models can separate apparent cross-site variation from sampling noise, as in the seven-experiment microcredit analysis of Meager (2019). Our estimand is instead the effect in a declared new setting, and our first output can be refusal. Causal transportability and data fusion formalize when experimental and observational distributions identify a target causal query (Pearl & Bareinboim, 2014; Bareinboim & Pearl, 2013; 2016). Trial-to-population methods and causally interpretable meta-analysis give point estimators under measured effect-modifier exchangeability and positivity (Cole & Stuart, 2010; Dahabreh et al., 2019; 2023). Under S-admissibility and covariate shift, transportable experimental design uses the target covariate distribution to choose treatment allocation and balance in a source-population experiment (Phan et al., 2021). Our setting complements these exact identification results by studying a finite-archive problem in which the invariances needed for point transport are not established and the remaining mechanism knowledge is represented by calibrated sets.

Learned representations and mechanism uncertainty. Learned representations organize treatment–outcome descriptions and propose candidate matches. Our framework uses them through the set map $Z \mapsto U(Z)$ and its calibration diagnostics, which carry the mechanism uncertainty into the certificate. The resulting contribution is calibrated set-valued mechanism uncertainty, simultaneous source-noise control after weight selection, sharp finite-metric refusal regions, and a stated nonadaptive bridge-design guarantee.

Orthogonal estimation and selective uncertainty. Within each study, semiparametric orthogonalization has earlier roots in influence-function estimation (Robins et al., 1994), while modern orthogonal scores with cross-fitting reduce first-order nuisance sensitivity (Chernozhukov et al., 2018). We use these tools as experiment-level inputs and carry their residual bias envelopes into the archive certificate. The new issue is selection across studies. A pointwise sub-Gaussian bound does not automatically survive outcome-dependent weight search. Our simultaneous certificate is finite-sample under declared noise proxies; plug-in proxies support asymptotic, rather than exact, language.

Partial identification and active design. Partial identification treats a set as the conclusion when assumptions do not point-identify a parameter (Manski, 1990; Imbens & Manski, 2004). Recent data-fusion work combines sensitivity analysis and partial identification (Lanners et al., 2025). Likewise, honest inference over smooth function classes has a substantial decision-theoretic literature (Armstrong & Kolesár, 2018). We connect refusal to a finite multi-source Lipschitz feasibility problem, propagate source uncertainty, and prove sharpness through the classical extension principle of McShane (1934). Evidence selection creates a separate inferential problem (Berk et al., 2013); Theorem 5.2 addresses it with a simultaneous event. Under a cardinality constraint, the classical greedy guarantee applies to normalized monotone submodular objectives (Nemhauser et al., 1978). Gaussian information gain is a canonical example in sensor placement (Krause et al., 2008); outcome-adaptive bridge policies may call for time-uniform inference, while greedy approximation guarantees would require adaptive submodularity or another suitable sequential argument (Howard et al., 2021; Golovin & Krause, 2011). We identify why target PI width lacks even static submodularity in general, then state precisely the nonadaptive model in which the log-determinant surrogate is justified.

Uncertain covariates and calibration. Guo et al. (2022) formulate partial identification with noisy unit-level covariates as robust optimization under a specified noise level. Here uncertainty concerns study-level mechanisms and a target-setting effect; the inputs and identification objects differ. Rank-based calibration also draws on the exchangeability principle used in conformal inference (Lei et al., 2018). Our synthetic audit applies that principle to archive-wise maximum scores, and requires exchangeability across audit and test archives.

B Assumption Audit and Extended Operational Method

Table 1: Assumption audit. A failed diagnostic changes the output; it is not repaired by semantic similarity.

Condition	Role	Diagnostic or falsification attempt	Fallback when doubtful
Within-study identification	Gives each archive effect a causal meaning	Randomization audit, overlap plot, balance, nuisance residual checks	Enlarge β_i, use sensitivity bounds, or exclude the study
Design compatibility	Places effects on one estimand and scale	Versioned treatment, control, outcome, and horizon checklist	Apply an audited map or set $\kappa = 0$
Mechanism-set coverage	Connects observed records to latent mechanisms	Leave-one-study-out calibration, negative controls, moderator audit	Enlarge U_i; at the limit use $U_i = \mathcal{M}$ and refuse
Smoothness (L, H)	Converts mechanism geometry to effect error	Multiscale residual plots and a sensitivity path over (L, H)	Report a family of intervals or a weaker shape-restricted region
Summary noise	Provides finite-sample stochastic coverage	Randomization-based checks, tail plots, dependence audit	Use exact design inference, a joint proxy, or asymptotic-only intervals

B.1 Separated geometry and computation

For interpretation, the joint envelope in (4.1) obeys

$$B_U(\alpha) \le L\rho_U(\alpha) + \frac{H}{2} Q_U(\alpha) + \sum_i \alpha_i \beta_i, \tag{B.1}$$

where, on $\mathcal{W}_U = U_\star \times \prod_{i \in C} U_i$,

$$\rho_U(\alpha) = \sup_{\mathbf{m} \in \mathcal{W}_U} \|m_\star - \bar{m}_\alpha\|, \qquad Q_U(\alpha) = \sup_{\mathbf{m} \in \mathcal{W}_U} \sum_i \alpha_i \|m_i - \bar{m}_\alpha\|^2. \tag{B.2}$$

The ideal optimization targets the joint supremum. The finite-budget implementation in Appendix L instead minimizes a valid causal-error envelope using local feasible solutions; neither exact evaluation of the joint supremum nor global optimality is claimed for that implementation. The simultaneous term in R_∞ is deliberately an ℓ_1 aggregation because it remains valid after outcome-adaptive weight selection. If the weights are measurable with respect to design information that excludes evaluation noise, the sharper radius is

$$R_2(\alpha; \zeta) = B_U(\alpha) + \sqrt{2 \log(2/\zeta) \sum_{i \in C} \alpha_i^2 s_i^2}. \tag{B.3}$$

Compactness ensures a minimizer when B_U is continuous. Box and ellipsoidal mechanism sets admit standard robust-optimization bounds; nonconvex scientific sets require a global solver or a verified upper relaxation. A relaxation can make the method more conservative but cannot license an invalid release. The release tolerance δ_{sci} is the largest target-scale absolute error acceptable for the decision and is fixed before target outcomes are observed. For a deploy/no-deploy threshold $\tau_{\min}$, one defensible choice is the largest error that cannot cross a predeclared indifference region around $\tau_{\min}$. If no downstream loss can be elicited, the method reports the full radius–release frontier instead of selecting δ_{sci} from target outcomes.

On the simultaneous coordinate event, the source interval is $I_i = [\widehat{\tau}_i - \beta_i - \sqrt{2 \log(2|C|/\zeta)} s_i, \widehat{\tau}_i + \beta_i + \sqrt{2 \log(2|C|/\zeta)} s_i]$. In the general set-valued nonlinear model, exact endpoints of (4.3) require global optimization; a verified outer relaxation remains coverage-valid and must be labeled as such.

B.2 Expanded ideal decision rule

Algorithm 2 Expanded ideal causal reuse with certified refusal and bridging

Require: Target query; archive summaries $(Z_i, \widehat{\tau}_i, s_i, \beta_i)$; set map Γ; (L, H, η, ζ); tolerance δ_{sci}; bridge menu $\mathcal{B}$ and budget B.

Freeze the estimand and audit design signatures; set $C = \{i : \kappa(i, \star) = 1\}$.
Construct $U_i = \Gamma(Z_i)$ and record the calibration protocol for Assumption 3.2.
if $C \neq \varnothing$ **then**
    Globally minimize a certified upper bound on $R_\infty(\alpha; \zeta)$ over Δ_C.
    if $R_\infty(\widehat{\alpha}; \zeta) \leq \delta_{\mathrm{sci}}$ **then**
        return TRANSPORT: $(\widehat{\theta}, [\widehat{\theta} \pm R_\infty(\widehat{\alpha}; \zeta)], \widehat{\alpha})$ and the assumption audit.
    end if
end if
Compute the exact envelope (4.3) or a labeled verified outer relaxation.
if the target projection is unbounded **then**
    return REFUSE: unbounded region and an outcome-scale anchor experiment.
end if
if the linear-Gaussian bridge model and its diagnostics hold **then**
    Greedily select B candidates by marginal log-determinant information.
else
    Use exact/certified planning; attach no submodular guarantee.
end if
return REFUSE: the identified envelope, selected bridge design, and failed diagnostics.

C Extended Theoretical Statements

C.1 Barycentric and fixed-weight refinements

Lemma C.1 (Barycentric smoothness decomposition)**.** *Under Assumption 3.3, for every $\alpha \in \Delta_C$ and every realized mechanism tuple $\mathbf{m}$,*

$$\left| \mu(m_\star) - \sum_{i \in C} \alpha_i \mu(m_i) \right| \leq G(\alpha; \mathbf{m}). \tag{C.1}$$

The first term in G measures target support, while the second is a curvature penalty. Even when $m_\star = \bar{m}_\alpha$, a nonlinear surface need not equal the average of its endpoint values.

Corollary C.2 (Sharper certificate for design-measurable weights)**.** *If $\alpha \in \Delta_C$ is measurable with respect to the conditioning information in Assumption 3.4 and is independent of the evaluation noises, then with probability at least $1 - \eta - \zeta$,*

$$\left| \tau_\star - \sum_i \alpha_i \widehat{\tau}_i \right| \leq R_2(\alpha; \zeta). \tag{C.2}$$

If the noise vector obeys $\mathbb{E}\{\exp(t^\top \xi) \mid \mathcal{D}\} \leq \exp(t^\top \Sigma t / 2)$, the stochastic term in (B.3) can be replaced by $\sqrt{2 \log(2/\zeta) \alpha^\top \Sigma \alpha}$.

The ℓ_1 simultaneous radius may be wider than the fixed-weight ℓ_2 radius. This is the price of letting the same archive outcomes choose weights and certify the result without further structural restrictions. Sample splitting, design-only weights, or a valid uniform empirical-process bound can recover sharper inference.

C.2 Bias-aware asymptotic inference

For growing archives, write α_n for design-measurable weights and assume

$$\widehat{\tau}_{i,n} - \tau_{i,n} = \frac{1}{n_i} \sum_{\ell=1}^{n_i} \psi_{i\ell} + r_{i,n}, \qquad \mathbb{E}\psi_{i\ell} = 0, \quad \mathrm{Var}(\psi_{i\ell}) = v_i. \tag{C.3}$$

Let $V_n = \sum_i \alpha_{n,i}^2 v_i / n_i$ and $a_n = \tau_{\star,n} - \sum_i \alpha_{n,i}\tau_{i,n}$.

Theorem C.3 (Bias-aware asymptotic inference)**.** *Suppose the influence-function summands are independent across (i,ℓ), $\sum_i \alpha_{n,i} r_{i,n} = o_{\mathbb{P}}(\sqrt{V_n})$, and, for every $\varepsilon > 0$,*

$$\frac{1}{V_n}\sum_i \sum_{\ell=1}^{n_i} \mathbb{E}\left[\left(\frac{\alpha_{n,i}\psi_{i\ell}}{n_i}\right)^2 \mathbf{1}\left\{\frac{|\alpha_{n,i}\psi_{i\ell}|}{n_i} > \varepsilon\sqrt{V_n}\right\}\right] \to 0. \tag{C.4}$$

Then

$$\frac{\widehat{\theta}_n - \sum_i \alpha_{n,i}\tau_{i,n}}{\sqrt{V_n}} \rightsquigarrow \mathsf{N}(0,1). \tag{C.5}$$

If $\widehat{V}_n / V_n \to_{\mathbb{P}} 1$ and an externally calibrated allowance A_n satisfies $|a_n| \le A_n$, then $[\widehat{\theta}_n \pm \{z_{1-\zeta/2}\sqrt{\widehat{V}_n} + A_n\}]$ has asymptotic coverage at least $1-\zeta$. Centering the normal limit at $\tau_{\star,n}$ additionally requires $A_n = o(\sqrt{V_n})$.

C.3 Explicit sharp-refusal program and minimax boundary

For known locations, define

$$\begin{aligned}\mathcal{P}(\mathcal{J}) = \big\{(t_1,\dots,t_K,t_\star) \in \mathbb{R}^{K+1} : \ell_i \le t_i \le u_i,\ |t_i - t_j| \le Ld(x_i,x_j),\\ |t_\star - t_i| \le Ld(x_\star,x_i), \quad i,j \le K\big\}.\end{aligned} \tag{C.6}$$

With exact mutually compatible source values t_i, the sharp region is

$$\left[\max_i\{t_i - Ld(x_\star,x_i)\}, \min_i\{t_i + Ld(x_\star,x_i)\}\right]. \tag{C.7}$$

The minimax error depends on more than a convex-hull distance. Take $\mathcal{M} = [-1,1]$, noiseless archive points -1 and 1 with effects zero, and target 0. The target-to-hull distance is zero, but both $\mu_0(x) = 0$ and $\mu_1(x) = a(1-|x|)$ are Lipschitz-compatible and disagree at the target. General matching theory must include curvature, fill distance, or the modulus of the chosen function class.

C.4 Bridge complementarity and the structured repair

For the partial-identification witness,

$$F_{\mathrm{PI}}(\{1\}) - F_{\mathrm{PI}}(\varnothing) = 0 < 2 = F_{\mathrm{PI}}(\{1,2\}) - F_{\mathrm{PI}}(\{2\}). \tag{C.8}$$

The construction takes $\beta \in [-1,1]^2$, target $\theta = \beta_1$, and bridge equations $\beta_1 + \beta_2 = 0$ and $\beta_2 = 0$. Neither equation alone narrows θ, while together they identify it.

For the structured repair, let $\beta \sim \mathsf{N}(0, V_0)$ with $V_0 \succ 0$, and let candidate b observe

$$Y_b = x_b^\top \beta + \epsilon_b, \qquad \epsilon_b \sim \mathsf{N}(0,\sigma_b^2), \tag{C.9}$$

independently. After selecting $S \subseteq \mathcal{B}$,

$$V_S = \left(V_0^{-1} + \sum_{b\in S} \sigma_b^{-2} x_b x_b^\top\right)^{-1}, \tag{C.10}$$

$$F_{\log\det}(S) = \frac{1}{2}\log\frac{\det V_0}{\det V_S}. \tag{C.11}$$

The greedy marginal and guarantee used in Theorem 5.6 are

$$\Delta(b \mid S) = \frac{1}{2}\log\{1 + \sigma_b^{-2} x_b^\top V_S x_b\}, \tag{C.12}$$

$$F_{\log\det}(S_g) \ge (1 - e^{-1}) F_{\log\det}(S^*). \tag{C.13}$$

Corollary C.4 (From global information to target radius)**.** *Suppose $\mathrm{cond}(V_S) \le \varkappa$ for every feasible size-B posterior. For $r_S(q) = q\sqrt{x_\star^\top V_S x_\star}$,*

$$r_S(q) \le q\,\|x_\star\|\,\sqrt{\varkappa}\{\det(V_S)\}^{1/(2p)}. \tag{C.14}$$

Moreover,

$$r_{S_g}(q) \le q\,\|x_\star\|\,\sqrt{\varkappa}\left\{(\det V_0)^{1/e}(\det V_{S^*})^{1-1/e}\right\}^{1/(2p)}. \tag{C.15}$$

The condition-number requirement is testable from the design matrix and noise levels. If it fails, D-optimality may spend information on directions irrelevant to $x_\star$, so the fallback is a target-aware exact design or another proved objective.

D Extended Empirical Audits

Table 2: Theorem–experiment correspondence. A panel supports its assigned diagnostic only; it cannot validate assumptions in a different application.

Claim	Controlled intervention	Baseline or ablation	Primary diagnostic
Semantic impossibility	Increase a hidden shift while recorded distance remains zero	Sampling-only semantic interval	Coverage over the declared mechanism set
Selection safety	Select the largest $\lvert Z_i \rvert$ as archive size grows	Reused pointwise 95% interval	Post-selection coverage
Canonical minimax rate	Cross a 27×27 grid of Ld and $\mathcal{I}^{-1/2}$	Rate surface $\max\{Ld, \mathcal{I}^{-1/2}\}$	Worst-case MAE and normalized risk
Bridge boundary	Random linear-Gaussian instances plus exact PI counterexample	Random subsets and exhaustive optimum	Information ratio and PI widths

D.1 Full synthetic-audit synopsis

Panel A uses four source summaries with aggregate standard error 0.1, $L = 1$, and declared target mechanism radius 0.3; each grid point has $20{,}000$ replications. Panel B draws K independent standard Gaussian summaries, selects the coordinate with largest absolute value, and uses $50{,}000$ replications per K. Panel C evaluates the inverse-information anchor estimator on a 27×27 grid of $(Ld, \mathcal{I}^{-1/2})$ with $40{,}000$ common Gaussian draws per cell. Panel D draws 200 six-dimensional Gaussian design instances with 12 candidate bridges, noise standard deviation 0.5, and budget four; it enumerates all 495 size-four subsets and compares greedy with 250 random subsets per instance. Appendix H gives the full data-generating processes and artifact checks.

The hidden-shift panel makes the calibration boundary visible: robust coverage deteriorates once the truth leaves the declared set. The selection panel isolates post-search failure independently of causal smoothness. The phase diagram checks only the canonical anchored model, not general nonlinear convex-hull composition. The bridge panel compares the implemented surrogate with exhaustive and random selection, while the exact inset checks the boundary theorem; none licenses greedy optimization of direct PI width.

D.2 Outcome-blind mechanism-set calibration details

The controlled audit freezes four archive sources, six candidate bridges, and one target. It uses 199 independent labelled audit archives to calibrate the row-wise maximum of 11 two-dimensional proxy errors at $\eta = .05$; no treatment outcome, target effect, or future bridge outcome enters calibration. Across 2,000 independent nominal test archives, joint coverage is 0.951 (95% Wilson interval $[0.9406, 0.9596]$). Marginal 95th-percentile calibration gives 0.544 joint coverage (interval $[0.5221, 0.5657]$), despite 0.9465 per-entity coverage. Under a 1.5 scale shift applied only to test proxy errors, joint coverage falls to 0.348.

D.3 Many Labs 2 diagnostics and evidence boundary

The all-source fixed-effect estimate is 0.1728 risk-difference units (SE 0.0108). The DerSimonian–Laird estimate is 0.1729, the estimated between-source standard deviation is 0.0051, and $I^2 = 0.4\%$. Fixed- and random-effects predictive intervals have mean widths 0.4041 and 0.4149. At half-width threshold 0.15, six releases are too few for a confirmatory conditional-coverage claim; the complete frontier is reported to prevent threshold selection from target outcomes.

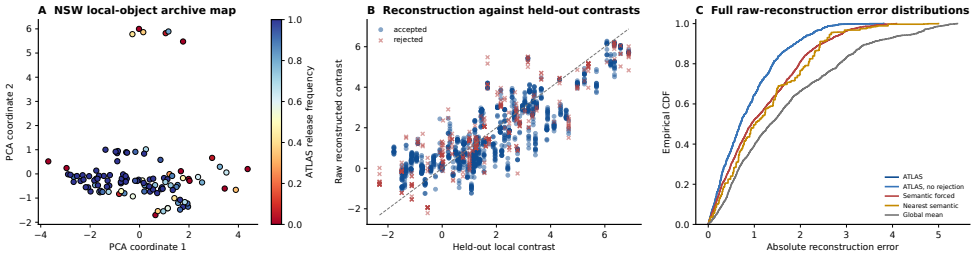

Figure 4: NSW within-trial reconstruction audit. **Left**: PCA map of 112 overlapping local archive objects. **Middle**: reconstruction against noisy held-out local contrasts, with released and refused targets distinguished. **Right**: absolute-error distributions. Inclusion refers to a noisy held-out reference, not coverage of a latent subgroup effect or a new study.

The Many Labs 2 split excludes the independently collected target source from fitting and release decisions. The benchmark evaluates source holdout under one standardized intervention; mechanism-set calibration requires separate latent-mechanism measurements. A confirmatory extension must calibrate U_i without target outcomes and test across treatment versions or effect families.

A future seven-site microcredit benchmark would first harmonize treatment versions and outcomes among the trials studied by Meager (2019), then predict a held-out eligible trial from the remaining eligible summaries. The executed Many Labs 2 effect-family holdout is reported below. Primary endpoints are target coverage, interval width conditional on coverage, selective risk–coverage, sign error among releases, and bridge value under a fixed budget. Support for the framework requires mechanism sets and smoothness constants calibrated without held-out target outcomes.

D.4 NSW WITHIN-TRIAL RECONSTRUCTION AUDIT

We retain a stress test based on the National Supported Work experiment (LaLonde, 1986; Dehejia & Wahba, 1999). It forms 112 local archive objects and evaluates 1,680 held-out local-contrast reconstructions. These objects overlap in units and arise from one trial; their held-out contrasts are noisy references rather than latent subgroup causal truths. Mechanism sets are not externally calibrated, so the audit can diagnose representation, reconstruction, and interval conservatism, but not independent-study transport or nominal target-effect coverage.

The operational prototype has all-target MAE 0.862 thousand dollars, sign agreement 0.854, reference inclusion 0.974, mean width 5.630, and refusal rate 0.232. Its released-target MAE is 0.864, so refusal does not reduce reconstruction error here. Nearest-semantic reconstruction has inclusion 0.992 with width 8.311; semantic forced composition has MAE 1.169, inclusion 0.605, and width 2.854. These comparisons report different interval constructions. Full committed tables and stronger baselines appear in Appendix I.

D.5 REPRODUCIBILITY AND ETHICS

The theorem-linked simulation is generated by `run_operational_experiments.py`; it writes every plotted quantity to CSV and records seeds and replication counts. The Many Labs 2 analysis is generated by `run_manylabs2_framing.py`, which verifies a commit-pinned checksum before reproducing all holdouts and the selective frontier. NSW artifacts remain separated from both new audits. No held-out target outcome enters fitting or release in the stated protocols.

Unsupported causal transfer can affect policy allocation and individual welfare. A certificate does not replace domain expertise, measurement validation, privacy review, or a causal graph when available. Refusal reduces one failure mode, but harm remains possible if users understate mechanism sets, smoothness constants, dependence, or nuisance bias.

E NOTATION, GEOMETRY, AND CLAIM SCOPE

Meaning of the mechanism geometry. The norm on $\mathcal{M}$ is not invariant to arbitrary rescaling. Before analysis, each coordinate must therefore have a declared scientific unit or be normalized by a versioned reference scale. Coordinates may represent effect-modifier distributions, exposure intensity, treatment components, control services, or measurement horizon. A convex barycenter $\bar{m}_\alpha$ is a mathematical interpolation point used by the smoothness argument; it need not be a physically deployable experiment. The method requires μ to be defined on the containing convex domain, not every point in that domain to be realizable. If interpolation lacks scientific meaning, use a metric-space Lipschitz model and the sharp region of Section 5.3 instead of the curvature certificate.

Table 3: Core notation. All target quantities refer to the estimand frozen before target outcomes are inspected.

Symbol	Meaning
$i \in \{1, \dots, K\}$, $\star$	Archive-study index and target-study index
$Z_i = (r_i, d_i, a_i)$	Observed representation, design signature, and assumption audit
$m_i \in \mathcal{M}$	Latent causal mechanism; never an algorithm input
$U_i = \Gamma(Z_i)$	Compact set of mechanisms compatible with the observed record
η	Joint mechanism-set miscoverage budget
$\tau_i = \mu(m_i)$	Normalized study average effect
C	Studies passing the deterministic compatibility audit
$\widehat{\tau}_i = \tau_i + \xi_i + b_i$	Archive effect summary, centered error, and residual bias
s_i, β_i	Sub-Gaussian proxy and deterministic/high-probability bias envelope
L, H	Gradient norm and gradient-Lipschitz bounds
$B_U(\alpha)$	Worst-case causal approximation plus nuisance envelope
R_∞, R_2	Selection-safe and design-measurable total radii
δ_{sci}	Prespecified target-scale loss tolerance for release
Θ_U	Target effects compatible with mechanism sets, source intervals, and the function class
$V_S, F_{\log\det}(S)$	Posterior covariance and Gaussian bridge information

Empty anchors and unbounded effects. If $C = \varnothing$ and no external effect-range restriction is supplied, constant shifts of μ make $\Theta_U = \mathbb{R}$. Algorithm 1 then returns an explicitly unbounded region; it does not form an undefined $\infty - \infty$ bridge value. The first bridge must provide an outcome-scale anchor, or the user must supply a scientifically justified range $[M_-, M_+]$. Subsequent width-based planning begins only after the target projection is bounded.

Optimization error. Validity uses the radius evaluated at the returned weights. An approximate optimizer $\widetilde{\alpha}$ is therefore safe whenever $R_\infty(\widetilde{\alpha}; \zeta)$ is itself upper-certified. Optimization error can reduce release power, but cannot invalidate coverage. Claims of near-optimal release require a separate bound $R_\infty(\widetilde{\alpha}) \le \inf_{\alpha \in \Delta_C} R_\infty(\alpha) + \epsilon_{\mathrm{opt}}$.

F PROOFS OF THE MAIN RESULTS

F.1 SEMANTIC OBSERVATIONAL EQUIVALENCE

Proof of Theorem 5.1. It is enough to consider the two-point metric subspace $\{m_a, m_b\}$; the functions constructed below extend to a larger metric space by the McShane extension theorem. Put every archive mechanism at m_a, choose any archive data law whose normalized causal effect is zero, and use that same law in both worlds. Let the target mechanism be m_b, but provide no target outcome. Because $r(m_a) = r(m_b)$ and the target design signature is the same in both worlds, every observed input to an r-only rule has the same distribution.

Define

$$\mu_+(m_a) = \mu_-(m_a) = 0, \qquad \mu_+(m_b) = Ld, \qquad \mu_-(m_b) = -Ld.$$

Both functions are L-Lipschitz on their respective two-point models. Let Q denote the common distribution of the estimator T. The triangle inequality gives, pointwise in t,

$$|t - Ld| + |t + Ld| \geq 2Ld.$$

Integrating with respect to Q yields

$$\mathbb{E}_{P_+}|T - Ld| + \mathbb{E}_{P_-}|T + Ld| \geq 2Ld.$$

At least one term is at least Ld, proving (5.1). □

F.2 Smooth barycentric composition

Lemma F.1 (First-order remainder). *If $\nabla\mu$ is H-Lipschitz on a convex set, then for all x, y in that set,*

$$\left|\mu(x) - \mu(y) - \nabla\mu(y)^\top(x - y)\right| \leq \frac{H}{2}\|x - y\|^2. \tag{F.1}$$

Proof. Let $h = x - y$. By the fundamental theorem of calculus,

$$\mu(y + h) - \mu(y) - \nabla\mu(y)^\top h = \int_0^1 \{\nabla\mu(y + th) - \nabla\mu(y)\}^\top h\, dt.$$

Cauchy–Schwarz and gradient Lipschitzness bound the integrand in absolute value by $Ht\|h\|^2$. Integration over $t \in [0, 1]$ gives (F.1). □

Proof of Lemma C.1. Write $\bar{m} = \sum_i \alpha_i m_i$. The gradient bound and the line-integral formula imply

$$|\mu(m_\star) - \mu(\bar{m})| \leq L\|m_\star - \bar{m}\|. \tag{F.2}$$

Apply Lemma F.1 to $(m_i, \bar{m})$:

$$\mu(m_i) = \mu(\bar{m}) + \nabla\mu(\bar{m})^\top(m_i - \bar{m}) + R_i, \qquad |R_i| \leq \frac{H}{2}\|m_i - \bar{m}\|^2.$$

Multiplying by α_i and summing cancels the linear term because

$$\sum_i \alpha_i(m_i - \bar{m}) = 0.$$

Hence

$$\left|\mu(\bar{m}) - \sum_i \alpha_i\mu(m_i)\right| \leq \frac{H}{2}\sum_i \alpha_i\|m_i - \bar{m}\|^2. \tag{F.3}$$

Combining (F.2) and (F.3) by the triangle inequality proves (C.1). □

F.3 Finite-sample selection-safe certificates

Proof of Theorem 5.2. For each i, the conditional sub-Gaussian bound implies

$$\mathbb{P}(|\xi_i| > q_{K,\zeta}s_i \mid \mathcal{D}) \leq 2\exp(-q_{K,\zeta}^2/2) = \frac{\zeta}{|C|}.$$

The union bound gives

$$\mathbb{P}(\mathcal{E}_\xi \mid \mathcal{D}) \geq 1 - \zeta, \qquad \mathcal{E}_\xi = \bigcap_{i \in C}\{|\xi_i| \leq q_{K,\zeta}s_i\}. \tag{F.4}$$

This step uses no finite discretization of Δ_C.

On $\mathcal{G}_U$, Lemma C.1 and the definition of the supremum in (4.1) imply, simultaneously for all $\alpha \in \Delta_C$,

$$\left|\tau_\star - \sum_i \alpha_i\tau_i\right| \leq B_U(\alpha) - \sum_i \alpha_i\beta_i. \tag{F.5}$$

On $\mathcal{E}_\xi$, nonnegativity of simplex weights gives, again simultaneously for all α,

$$\left|\sum_i \alpha_i \xi_i\right| \le \sum_i \alpha_i |\xi_i| \le q_{K,\zeta} \sum_i \alpha_i s_i. \tag{F.6}$$

Also $|\sum_i \alpha_i b_i| \le \sum_i \alpha_i \beta_i$. Substituting the random value $\widehat{\alpha}$ into the simultaneous inequalities and applying the triangle inequality proves (5.2) on $\mathcal{G}_U \cap \mathcal{E}_\xi$.

Finally,

$$\mathbb{P}(\mathcal{G}_U \cap \mathcal{E}_\xi) \ge 1 - \mathbb{P}(\mathcal{G}_U^c) - \mathbb{P}(\mathcal{E}_\xi^c) \ge 1 - \eta - \zeta,$$

where the conditional guarantee in (F.4) remains valid after averaging over $\mathcal{D}$. This completes the proof. □

Proof of Corollary C.2. Conditional on $\mathcal{D}$, Assumption 3.4 gives

$$\mathbb{E}\left[\exp\left\{t\sum_i \alpha_i \xi_i\right\} \middle| \mathcal{D}\right] \le \exp\left\{\frac{t^2}{2}\sum_i \alpha_i^2 s_i^2\right\}.$$

The Chernoff bound applied to both tails yields

$$\mathbb{P}\left(\left|\sum_i \alpha_i \xi_i\right| > \sqrt{2\log(2/\zeta)\sum_i \alpha_i^2 s_i^2} \,\middle|\, \mathcal{D}\right) \le \zeta.$$

Combine this event with (F.5), the bias envelope, and the set-coverage event as in the preceding proof. Under the joint mgf condition, substitute $t\alpha$ for the vector argument to obtain proxy $\alpha^\top \Sigma \alpha$. □

Selective error control. Let Rel be the event that Algorithm 1 releases and let δ_{sci} be its threshold. Theorem 5.2 immediately gives

$$\mathbb{P}\{\mathsf{Rel} \text{ and } |\widehat{\theta} - \tau_\star| > \delta_{\mathrm{sci}}\} \le \eta + \zeta. \tag{F.7}$$

The guarantee controls joint selective error. Conditional-on-release coverage requires dividing by the release probability or imposing additional selective-inference structure.

F.4 Asymptotic archive composition

Proof of Theorem C.3. Define the triangular-array variables

$$X_{n,i\ell} = \frac{\alpha_{n,i}}{n_i}\psi_{i\ell}.$$

They are independent, centered, and their total variance is

$$\sum_i \sum_{\ell=1}^{n_i} \mathrm{Var}(X_{n,i\ell}) = \sum_i \frac{\alpha_{n,i}^2 v_i}{n_i} = V_n.$$

Condition (C.4) is exactly the Lindeberg condition for the standardized array $X_{n,i\ell}/\sqrt{V_n}$. The Lindeberg–Feller central limit theorem therefore yields

$$V_n^{-1/2} \sum_i \sum_{\ell=1}^{n_i} X_{n,i\ell} \rightsquigarrow \mathsf{N}(0,1).$$

By (C.3),

$$\widehat{\theta}_n - \sum_i \alpha_{n,i}\tau_{i,n} = \sum_i \sum_{\ell=1}^{n_i} X_{n,i\ell} + \sum_i \alpha_{n,i} r_{i,n}.$$

The last term is $o_{\mathbb{P}}(\sqrt{V_n})$, so Slutsky's theorem proves (C.5). Ratio consistency of $\widehat{V}_n$ permits studentization.

For the interval claim, $|a_n| \le A_n$ implies

$$|\widehat{\theta}_n - \tau_{\star,n}| \le \left|\widehat{\theta}_n - \sum_i \alpha_{n,i}\tau_{i,n}\right| + A_n.$$

The studentized central limit theorem makes the first term at most $z_{1-\zeta/2}\sqrt{\widehat{V}_n}$ with limiting probability $1-\zeta$, which proves the coverage lower bound. If $A_n = o(\sqrt{V_n})$, then subtracting $a_n/\sqrt{V_n} = o(1)$ does not change the limiting law; without that condition, a target-centered normal limit does not follow. □

F.5 Sharp Lipschitz refusal

Proof of Theorem 5.3. Let $X_0 = \{x_1, \dots, x_K, x_\star\}$. If an L-Lipschitz function $f : \mathcal{M} \to \mathbb{R}$ satisfies $f(x_i) \in [\ell_i, u_i]$, then its restriction $t = (f(x_1), \dots, f(x_K), f(x_\star))$ satisfies every inequality in (C.6). Thus every compatible target value belongs to the projection of $\mathcal{P}(\mathcal{J})$.

Conversely, fix any $t \in \mathcal{P}(\mathcal{J})$. All pairwise inequalities say that the finite assignment $f_0(x_j) = t_j$ on X_0 is L-Lipschitz. Define, for $x \in \mathcal{M}$,

$$F_t(x) = \inf_{z \in X_0} \{f_0(z) + Ld(x, z)\}. \tag{F.8}$$

For $x = z_0 \in X_0$, Lipschitz compatibility gives $f_0(z) + Ld(z_0, z) \ge f_0(z_0)$ for every z, while choosing $z = z_0$ gives equality. Hence $F_t(z_0) = f_0(z_0)$. For arbitrary x, y and each z,

$$f_0(z) + Ld(x, z) \le f_0(z) + Ld(y, z) + Ld(x, y).$$

Taking infima gives $F_t(x) \le F_t(y) + Ld(x, y)$; exchanging x, y proves that F_t is L-Lipschitz. Therefore every feasible $t_\star$ is attained by a full-space function.

The polyhedron is convex. Its projection on the $t_\star$ coordinate is therefore a convex subset of $\mathbb{R}$, hence an interval. Because $K \ge 1$, the constraints $|t_\star - t_i| \le Ld(x_\star, x_i)$ and the bounded source intervals make that projection compact. Linear objectives attain their minimum and maximum, giving (5.3). If the source intervals jointly contain the true source effects and the true surface is L-Lipschitz, its finite restriction lies in $\mathcal{P}(\mathcal{J})$, so the target lies in the projection. Taking probabilities proves the final statement. □

Derivation of (C.7). When t_i are fixed and mutually Lipschitz-compatible, the only constraints containing $t_\star$ are $t_i - Ld(x_\star, x_i) \le t_\star \le t_i + Ld(x_\star, x_i)$. Their intersection is (C.7). Nonemptiness follows from the triangle inequality and source compatibility, or directly from the extension constructed above. □

F.6 Canonical minimax characterization

Proof of Theorem 5.4. Let

$$\overline{Z} = \mathcal{I}^{-1} \sum_{i=1}^{K} s_i^{-2} Z_i.$$

Then $\overline{Z} \sim \mathsf{N}(\mu(m_0), \mathcal{I}^{-1})$. The estimator $T_0 = \overline{Z}$ obeys, for every admissible μ,

$$\mathbb{E}_\mu|T_0 - \mu(m_\star)| \le |\mu(m_0) - \mu(m_\star)| + \mathbb{E}|\overline{Z} - \mu(m_0)| \le Ld + \sqrt{\frac{2}{\pi\mathcal{I}}},$$

which proves the upper bound.

For the support lower bound, take two functions with common anchor value zero and target values $+Ld$ and $-Ld$. Both satisfy the Lipschitz restriction, and the archive data law is identical under them. The same triangle-inequality argument as in Theorem 5.1 yields minimax risk at least Ld.

For the sampling lower bound, restrict further to constant functions $\mu_-(m) = -a$ and $\mu_+(m) = a$, where $a = (2\sqrt{\mathcal{I}})^{-1}$. Let P_- and P_+ be the resulting product Gaussian laws. Their Kullback–Leibler divergence is

$$D_{\mathrm{KL}}(P_-\|P_+) = \frac{1}{2}\sum_i \frac{(2a)^2}{s_i^2} = 2a^2\mathcal{I} = \frac{1}{2}.$$

Pinsker's inequality gives $\|P_- - P_+\|_{\mathrm{TV}} \le 1/2$. Any estimator T induces the nearest-parameter test. When that test errs, the absolute estimation error is at least a. The average of the two risks is therefore at least a times the minimum average testing error, namely

$$a\frac{1 - \|P_- - P_+\|_{\mathrm{TV}}}{2} \ge \frac{a}{4} = \frac{1}{8\sqrt{\mathcal{I}}}.$$

The maximum risk is at least the average risk. Combining the support and sampling constructions proves the lower bound in (5.5). □

F.7 Bridge complementarity and log-determinant information

Proof of Theorem 5.5. Without bridges, $\beta \in [-1,1]^2$ permits $\theta = \beta_1 \in [-1,1]$, so $w(\varnothing) = 2$. Bridge A imposes $\beta_1 + \beta_2 = 0$; each $\beta_1 \in [-1,1]$ remains feasible by taking $\beta_2 = -\beta_1$, hence $w(\{A\}) = 2$. Bridge B imposes $\beta_2 = 0$ and leaves β_1 unrestricted in $[-1,1]$, hence $w(\{B\}) = 2$. Together the equations imply $\beta_1 = \beta_2 = 0$, so $w(\{A,B\}) = 0$. The marginal gains are therefore exactly those in (C.8). Since a marginal gain increases when the conditioning set grows, F_{PI} is not submodular. □

Lemma F.2 (Posterior covariance ordering)**.** *If $S \subseteq T$, then $V_T \preceq V_S$ in the Loewner order.*

Proof. Equation (C.10) gives $V_T^{-1} = V_S^{-1} + A$ for a positive semidefinite matrix A. Congruence by $V_S^{1/2}$ yields

$$V_T = V_S^{1/2}(I + V_S^{1/2} A V_S^{1/2})^{-1} V_S^{1/2} \preceq V_S,$$

because $(I+B)^{-1} \preceq I$ for $B \succeq 0$. □

Proof of Theorem 5.6. The matrix determinant lemma gives

$$\begin{aligned}\Delta(b \mid S) &= \frac{1}{2}\log\frac{\det(V_S^{-1} + \sigma_b^{-2} x_b x_b^\top)}{\det(V_S^{-1})} \\ &= \frac{1}{2}\log\{1 + \sigma_b^{-2} x_b^\top V_S x_b\},\end{aligned}$$

which proves (C.12). This quantity is nonnegative, so $F_{\log\det}$ is monotone, and $F_{\log\det}(\varnothing) = 0$. If $S \subseteq T$, Lemma F.2 implies $x_b^\top V_T x_b \le x_b^\top V_S x_b$. Therefore $\Delta(b \mid T) \le \Delta(b \mid S)$, proving submodularity.

For completeness, let S_t be the greedy set after t selections and let $O = S^*$ with $|O| \le B$. Monotonicity and submodularity imply

$$\begin{aligned}F(O) - F(S_t) &\le F(O \cup S_t) - F(S_t) \\ &\le \sum_{b \in O \setminus S_t} \Delta(b \mid S_t) \\ &\le B\{F(S_{t+1}) - F(S_t)\},\end{aligned}$$

because greedy chooses a largest remaining marginal. Thus the gap $g_t = F(O) - F(S_t)$ satisfies $g_{t+1} \le (1 - 1/B) g_t$. Iteration gives

$$F(O) - F(S_B) \le (1 - 1/B)^B F(O) \le e^{-1} F(O),$$

which is (C.13). □

Proof of Corollary C.4. Let $0 < \lambda_1 \le \cdots \le \lambda_p$ be the eigenvalues of V_S. The condition-number bound gives $\lambda_p \le \varkappa \lambda_1$. Since $\lambda_1 \le (\prod_j \lambda_j)^{1/p}$,

$$x_\star^\top V_S x_\star \le \lambda_p \|x_\star\|^2 \le \varkappa \|x_\star\|^2 \{\det(V_S)\}^{1/p}.$$

Taking square roots proves (C.14).

Let $q_0 = 1 - 1/e$. The greedy guarantee is equivalent to

$$\log \frac{\det V_0}{\det V_{S_g}} \geq q_0 \log \frac{\det V_0}{\det V_{S^*}}.$$

Exponentiating and rearranging yields

$$\det V_{S_g} \leq (\det V_0)^{1-q_0} (\det V_{S^*})^{q_0} = (\det V_0)^{1/e} (\det V_{S^*})^{1-1/e}.$$

Substitution into (C.14) proves (C.15). □

G Experiment-Level Estimation and Bias Accounting

This appendix records the study-level estimation machinery that produces the summary decomposition $\widehat{\tau}_i = \tau_i + \xi_i + b_i$ from unit-level randomized data.

For one study, let $e_0(x) = \mathbb{P}(A = 1 \mid X = x)$ and $g_{0,a}(x) = \mathbb{E}(Y \mid A = a, X = x)$. Given nuisance functions $\widehat{e}, \widehat{g}_0, \widehat{g}_1$ trained on an independent fold, define

$$\begin{aligned} \phi(W; \widehat{\eta}) = \widehat{g}_1(X) - \widehat{g}_0(X) + \frac{A\{Y - \widehat{g}_1(X)\}}{\widehat{e}(X)} \\ - \frac{(1-A)\{Y - \widehat{g}_0(X)\}}{1 - \widehat{e}(X)}. \end{aligned} \tag{G.1}$$

The fold estimate averages (G.1) on its evaluation observations, and cross-fitting averages the fold estimates.

Lemma G.1 (Randomized-study unbiasedness)**.** *Under Assumption 3.1, if the randomized propensity e_0 is known and $\widehat{e} = e_0$, then, conditional on the nuisance-training data,*

$$\mathbb{E}\{\phi(W; \widehat{\eta}) \mid \widehat{\eta}\} = \mathbb{E}\{g_{0,1}(X) - g_{0,0}(X)\} = \tau. \tag{G.2}$$

This equality holds for arbitrary square-integrable outcome regressions $\widehat{g}_a$.

Proof. Condition on X and the training data. Randomization and consistency give

$$\begin{aligned} \mathbb{E}\left[\frac{A\{Y - \widehat{g}_1(X)\}}{e_0(X)} \middle| X, \widehat{\eta}\right] &= g_{0,1}(X) - \widehat{g}_1(X), \\ \mathbb{E}\left[\frac{(1-A)\{Y - \widehat{g}_0(X)\}}{1 - e_0(X)} \middle| X, \widehat{\eta}\right] &= g_{0,0}(X) - \widehat{g}_0(X). \end{aligned}$$

Substitution into (G.1) leaves $g_{0,1}(X) - g_{0,0}(X)$. Averaging over X proves the result. □

When the propensity is estimated, direct conditioning gives the exact population remainder

$$\begin{aligned} &\mathbb{E}\{\phi(W; \widehat{\eta}) - g_{0,1}(X) + g_{0,0}(X) \mid \widehat{\eta}\} \\ &\quad = \mathbb{E}\left[(\widehat{e} - e_0)\left\{\frac{\widehat{g}_1 - g_{0,1}}{\widehat{e}} + \frac{\widehat{g}_0 - g_{0,0}}{1 - \widehat{e}}\right\} \middle| \widehat{\eta}\right]. \end{aligned} \tag{G.3}$$

If $\widehat{e} \in [\pi_0, 1 - \pi_0]$, Cauchy–Schwarz yields

$$|\text{remainder}| \leq \pi_0^{-1} \left\|\widehat{e} - e_0\right\|_{L_2} \left(\left\|\widehat{g}_1 - g_{0,1}\right\|_{L_2} + \left\|\widehat{g}_0 - g_{0,0}\right\|_{L_2}\right). \tag{G.4}$$

Equation (G.4) becomes an observable finite-sample certificate only when the norms used as β_i have external or high-probability upper bounds. Our exact finite-sample simulations use known Gaussian scales and $\beta_i = 0$; NSW standard errors are descriptive/asymptotic inputs and are not used in a finite-sample theorem claim.

Table 4: Reproducibility manifest for the theorem-linked synthetic experiments. All entries are Monte Carlo results from prespecified synthetic data-generating processes.

Panel	Seed	Monte Carlo units	Diagnostic value
A	2026091601	20,000 per shift	min. certified coverage 0.974
A	2026091601	20,000 per shift	semantic coverage at boundary 0.151
B	2026091602	50,000 per K	pointwise/simultaneous at $K = 128$: 0.002/0.953
C	2026091603	40,000 on 27×27 grid	risk/rate range [0.796, 1.171]
D	2026091604	200 instances	min. greedy ratio 0.947
D	2026091604	250 random/instance	mean random ratio 0.704

H Synthetic Protocols and Reproducibility

H.1 Panel A: hidden-mechanism coverage

The source anchor effect is zero and the target effect is $\tau_\star = h$ for hidden shift $h \in \{0, 0.02, \ldots, 0.50\}$. Recorded source and target descriptors are identical. Four source experiments of size 25 have an aggregate Gaussian standard error 0.1. With $L = 1$ and declared set $U_\star = [-0.3, 0.3]$, the operational interval is

$$[\widehat{\theta} - \{0.3 + z_{0.975}(0.1)\}, \widehat{\theta} + \{0.3 + z_{0.975}(0.1)\}], \tag{H.1}$$

whereas the semantic ablation removes the 0.3 term. Common Gaussian draws are used across adjacent h values. Wilson intervals quantify Monte Carlo uncertainty in coverage.

H.2 Panel B: post-selection coverage

For $K \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, draw $Z_1, \ldots, Z_K$ independently from $\mathsf{N}(0, 1)$ and select $\widehat{j} = \arg\max_j |Z_j|$. The pointwise interval covers zero when $|Z_{\widehat{j}}| \le z_{0.975}$, whose exact probability is

$$\{2\Phi(z_{0.975}) - 1\}^K = 0.95^K. \tag{H.2}$$

The Gaussian Bonferroni interval uses $z_{1-0.05/(2K)}$, giving exact coverage

$$\{2\Phi(z_{1-0.05/(2K)}) - 1\}^K \ge 0.95. \tag{H.3}$$

The simulation independently verifies both expressions.

H.3 Panel C: canonical minimax phase diagram

For each pair (a, σ) on a 27×27 grid from 10^{-2} to 1, draw $Z \sim \mathsf{N}(0, \sigma^2)$ and evaluate the endpoint risk

$$R(a, \sigma) = \mathbb{E}|Z - a|, \tag{H.4}$$

where $a = Ld$ and $\sigma = \mathcal{I}^{-1/2}$. Symmetry makes the other endpoint identical. Panel C plots the Monte Carlo estimate of (H.4); it is a risk surface for the upper-bound estimator, not a numerical computation of the exact minimax constant.

H.4 Panel D: log-determinant bridge design

Each instance draws a six-dimensional positive-definite prior covariance with log-uniform eigenvalues in $[0.35, 3]$ and a random orthogonal eigenbasis. Twelve normalized Gaussian candidate vectors receive independent lognormal scales, and each observation has noise standard deviation 0.5. For budget four, exhaustive search evaluates 495 subsets. Greedy and random methods use the identical candidate pool and prior. All formula and invariant checks in the script must pass before artifacts are written.

H.5 COMMITTED ARTIFACT CHECKS

The script asserts that empirical probabilities lie in $[0, 1]$, the exact pointwise coverage equals 0.95^K to numerical tolerance, the simultaneous exact coverage is at least 0.95, every greedy ratio is at least $1 - 1/e$ up to floating-point tolerance, and the PI counterexample has widths $(2, 2, 2, 0)$. It exports a vector PDF with embedded fonts, a high-resolution PNG, panel-level CSV files, and the manifest table. Running the script overwrites only its declared generated artifacts.

I LEGACY SYNTHETIC AND NSW AUDITS

These tables report the earlier point-representation prototype and its weaker results. The mechanism-set calibration and finite-sample claims belong to the set-valued method.

Table 5: Stronger summary-data baselines on 300 common targets per synthetic condition. Cells are all-target MAE (Monte Carlo SE).

Method	Nominal	Moderate	Severe	High noise
ATLAS, no rejection	0.139 (0.006)	0.231 (0.009)	0.707 (0.025)	0.162 (0.007)
Semantic forced	0.249 (0.011)	0.491 (0.015)	1.133 (0.022)	0.267 (0.011)
Inverse-variance pooling	0.285 (0.013)	0.573 (0.015)	1.260 (0.017)	0.297 (0.014)
Random-effects pooling	0.282 (0.013)	0.569 (0.015)	1.256 (0.017)	0.294 (0.013)
Full-representation nearest	0.456 (0.020)	0.541 (0.022)	0.793 (0.033)	0.487 (0.022)
Ridge meta-regression	0.141 (0.006)	0.199 (0.008)	0.328 (0.016)	0.172 (0.007)
RBF kernel ridge	0.120 (0.005)	0.181 (0.008)	0.396 (0.019)	0.177 (0.008)

All-target ATLAS uses the no-rejection point estimate. Every method uses the same draws; ridge and kernel tuning uses source outcomes only. Point-only baselines have no reported confidence interval.

The stronger-baseline table shows that point accuracy is not uniformly dominated. In the nominal synthetic condition, RBF kernel ridge has MAE 0.120 versus 0.139 for the no-refusal prototype; under severe mismatch, ridge meta-regression has MAE 0.328 versus 0.707. These predictors do not supply the certificate studied here, so we report their point performance separately.

The semi-synthetic NSW-covariate study contains several comparators with smaller point error. Its targets overlap and its response surfaces are imposed by simulation, so it serves as a summary-archive debugging audit. The independent-study protocol below provides the separate source-holdout evaluation.

New NSW real-covariate and causal-truth audit. We add a participant-disjoint audit that reads only the eight pretreatment NSW covariates for its real-data calibration component. A fixed 40/30/30 development–calibration–validation split yields 14 calibration and 13 validation entities; the outcome-blind sup-norm proxy radius is 1.066, with validation profile coverage 0.846. This is a within-trial observed-covariate proxy result, not latent mechanism-set coverage or an independent study-level calibration archive. A separate 100-replicate experiment uses the original NSW assignment and covariates but frozen response surfaces with shared unit noise, giving an exact finite-neighborhood causal truth. Across 600 target evaluations per surface, ATLAS intervals cover the known truth in all evaluated cases while refusing all targets; the forced semantic intervals cover only 0.015, 0.0417, and 0.0933 on the constant, smooth, and interaction surfaces. Target outcomes and target truth are joined only for scoring, and the target-effect flip audit leaves every prediction and interval unchanged. The audit evaluates known causal effects on real covariates under synthetic outcomes. Real cross-intervention mechanism calibration requires an independent archive.

J INDEPENDENT-SOURCE BENCHMARK AND EXTENSION PROTOCOL

This appendix distinguishes the executed Many Labs 2 source-holdout benchmark from the broader validations that remain. Within-trial subgroups do not provide independent cross-intervention validation, and one standardized multi-source experiment cannot calibrate the mechanism sets needed for heterogeneous interventions.

Table 6: NSW held-out local-contrast reconstruction. All-target MAE uses all 1,680 raw reconstructions; Released MAE uses released targets only. Inclusion refers to a noisy held-out local contrast, not a latent subgroup effect; monetary quantities are thousands of U.S. dollars.

Method	All-target MAE	Released MAE	Median AE	Sign	Inclusion	Width	Rejection
Causal ATLAS	0.862	0.864	0.699	0.854	0.974	5.630	0.232
ATLAS, no rejection	0.862	0.862	0.699	0.854	0.970	5.275	0.000
Semantic forced	1.169	1.169	0.937	0.754	0.605	2.854	0.000
Nearest semantic	1.242	1.242	1.031	0.780	0.992	8.311	0.000
Global mean	1.679	1.679	1.349	0.784	0.446	2.475	0.000

Table 7: NSW-covariate semi-synthetic validation against the exact target average effect. Cells are all-target MAE (Monte Carlo SE), in thousands of dollars.

Method	Constant	Smooth	Interaction
ATLAS, no rejection	0.504 (0.024)	0.488 (0.023)	0.493 (0.022)
Semantic forced	0.393 (0.027)	0.457 (0.029)	0.458 (0.024)
Inverse-variance pooling	0.393 (0.029)	0.445 (0.029)	0.500 (0.024)
Random-effects pooling	0.393 (0.029)	0.445 (0.029)	0.499 (0.024)
Full-representation nearest	0.752 (0.030)	0.779 (0.028)	0.751 (0.028)
Ridge meta-regression	0.520 (0.028)	0.532 (0.029)	0.519 (0.031)
RBF kernel ridge	0.542 (0.026)	0.539 (0.027)	0.556 (0.024)
Unit-data ridge T-learner	0.422 (0.022)	0.426 (0.021)	0.417 (0.021)

Each surface has 100 independent assignment/outcome replicates and six overlapping targets. MC SE uses the 100 replicate means. The T-learner has source individual data; the other methods have archive summaries. ATLAS raw point estimates equal its no-rejection counterpart.

J.1 Seven-site microcredit benchmark

1. **Unit of independence.** Each of the seven independently conducted randomized trials in Meager (2019) is one archive object. Repeated outcomes within a site are not treated as new studies; the executed, versioned feasibility analyses below remain distinct from the planned seven-site benchmark.
2. **Estimand freeze.** Select a treatment-contrast definition, follow-up window, currency/price normalization, and outcome family common enough to be scientifically interpretable. Publish the harmonization map before leave-one-site-out evaluation.
3. **Split.** For fold j, seal all outcome summaries from site j. Fit or calibrate every representation, mechanism set, smoothness sensitivity path, and baseline using the other six sites plus non-outcome target metadata.
4. **Prediction and refusal.** Produce a point/interval only if the release radius is below a predeclared outcome-scale loss tolerance. Otherwise produce the identified envelope and bridge recommendation.
5. **Evaluation.** Unseal the randomized target estimate and its sampling uncertainty. Treat the estimate as a noisy reference rather than causal truth; report width, release, sign error, and risk–coverage only where the corresponding quantities are defined and the analysis is qualified.

J.2 Published-summary microcredit feasibility pilot

The source-corrected primary uses published community-access intention-to-treat profit summaries for Mexico (Table 3: 0 with assignment-clustered SE 39 in two-week pesos) and India endline 1 (Table 3A: 354 with weighted area-clustered SE 314 in 30-day rupees). Conversion gives 0 (4.003323) and 15.341040 (13.607589), respectively, in 2009 USD PPP per fortnight. Each holdout has one training study. The fixed-effect sampling-only baseline computes two targets and releases neither at half-width tolerance 0.20; one prediction interval contains the full noisy-reference sampling interval and one partially overlaps it. Random effects and the training range are `not_identified`; design meta-regression and descriptive ATLAS are `not_qualified` because design-coordinate provenance remains unresolved.

The separate post-hoc population sensitivity adds the all-sample weighted Morocco Table 8B estimate, -476 with village-clustered SE 1252 in annual MAD, converted to -4.247724 (11.172586) on the same scale. The old high-propensity-to-borrow stratum cell is excluded from the primary. With two training studies per target, fixed effects and the training range each compute three targets and release none at tolerance 0.20. Fixed-effect intervals contain one full noisy-reference interval and partially overlap two; training-range intervals partially overlap all three. Random effects remain `not_identified`; the design methods remain `not_qualified`. Every computed rule refuses at all four frozen tolerances 0.10, 0.20, 0.40, and 0.80. Released-target risk is undefined with no releases; uncomputed methods are not counted as successful refusals.

Evaluation uses noisy randomized references with unobserved causal truth. PPP and period conversions standardize units; equivalence of profit constructs, recall periods, populations, and control access requires separate justification. Overlapping holdouts are not independent repetitions. Target-effect and SE perturbations leave saved predictions and releases unchanged under fixed inputs and qualification gates; coordinate provenance requires a separate outcome-blind audit. An independent real joint mechanism-set calibration archive and completed human double coding remain unavailable, limiting interpretation to descriptive feasibility and refusal rather than real joint 95% coverage or calibrated ATLAS. The supplement records source cells, qualification decisions, protocols, input/code hashes, and run manifests. The seven-to-eighth-market example motivates future use beyond this pilot.

J.3 Executed Many Labs 2 source-holdout benchmark

The Many Labs 2 paper deposits data and materials at OSF project `8cd4r`; the public `ManyLabsOpenScience/ManyLabs2` repository links that project. We pin commit `acef63fc397b8dce7f0b00f863bcea78d324bea8` and verify SHA-256 `15898b5c241696adc7fd91a638839ba49a9f9be64c78da9151b39018d66bfbd3` for the cleaned Tversky.1 global file. The analysis retains rows marked for inclusion with factor in {Cheap, Expensive}, response in {Yes, No}, a nonmissing source, and both randomized arms within that source. No row is duplicated across source holdouts.

For source s and arm a, let $\widehat{p}_{sa}$ be the sample response rate and n_{sa} its arm size. The effect and unbiased plug-in sampling variance used by the meta-analytic baseline are

$$\widehat{\tau}_s = \widehat{p}_{s,\text{Cheap}} - \widehat{p}_{s,\text{Expensive}}, \tag{J.1}$$

$$\widehat{v}_s = \frac{\widehat{p}_{s,\text{Cheap}}(1-\widehat{p}_{s,\text{Cheap}})}{n_{s,\text{Cheap}}-1} + \frac{\widehat{p}_{s,\text{Expensive}}(1-\widehat{p}_{s,\text{Expensive}})}{n_{s,\text{Expensive}}-1}. \tag{J.2}$$

In fold s, the fixed-effect baseline uses weights $\widehat{v}_j^{-1}$ on $j \neq s$. The random-effects baseline estimates between-source variance by the nonnegative DerSimonian–Laird moment estimator using only $j \neq s$, then uses weights $(\widehat{v}_j + \widehat{\tau}^2_{\text{DL}})^{-1}$. To score prediction of the noisy held-out source estimate, both intervals add the outcome-blind worst-case target variance

$$v_{s,\max} = \frac{1}{4n_{s,\text{Cheap}}} + \frac{1}{4n_{s,\text{Expensive}}}. \tag{J.3}$$

The random-effects interval uses a small-sample t critical value. These are working-model predictive intervals, not finite-sample certificates.

The script evaluates every half-width threshold on a fixed grid and highlights $\delta = 0.15$ only as an illustrative operating point. It writes source effects, all 57 holdout predictions, the complete selective frontier, global meta-analysis diagnostics, figure files, and a machine-generated report. A deterministic checksum check runs before analysis, and rerunning the script reproduces every reported number.

J.4 Executed Many Labs 2 effect-family extension

Many Labs 2 contains multiple repeated effect protocols across geographically distinct samples (Klein et al., 2018). We use the author-level original-effects summary pinned at commit `acef63fc397b8dce7f0b00f863bcea78d324bea8` (SHA-256 `78b3432bebb798595ebb08ee10fb68d3f6f59cd5e0ab49606cfb6835eb226406`).

The fixed eligibility rule retains finite Cohen's d, positive reported variance, and $N > 1$; it yields 23 eligible families and 28 valid study-level effects. The confirmatory panel holds out all rows from the five complete families Hauser, Huang, Miyamoto, Ross, and Savani, giving ten targets. The sensitivity panel holds out every eligible family in turn, so all 28 effects are scored once. A laboratory-by-effect cell is not counted as an independent study for uncertainty.

For each held-out family, the training pool excludes the entire family before any fit or radius calculation. The transport baseline is a weighted context meta-regression using the pre-existing WEIRD/NONWEIRD design metadata. The hierarchical baseline is a Normal–Normal random-effects predictive interval with DerSimonian–Laird heterogeneity. Robust partial identification uses the training effect envelope plus an N-only sampling allowance. Study-level conformal calibration leaves out the unique `study.analysis` id; family-level conformal calibration leaves out the complete family. The first historical run mistakenly used `study.name`, which is a family label in this summary; that output remains archived for audit, while the corrected panel uses `study.analysis`.

Table 8 reports the complete threshold frontier. The corrected five-family panel has noisy-reference inclusion 0.800, 0.900, 1.000, 1.000, and 0.800 for transport, hierarchical, robust, study-conformal, and family-conformal methods, respectively; at thresholds 0.20–1.00, the first two release at most $2/10$ and the other three release none. The all-family sensitivity panel gives inclusion fractions 0.929, 0.964, 1.000, 1.000, and 0.929 over 28 targets; transport and hierarchical methods release at most $2/28$, while the other methods release $0/28$. Target-effect flip audits have maximum prediction and radius changes of zero. These inclusion rates do not establish causal coverage: the public summary lacks real mechanism labels, an independent real calibration archive, and noise-free target truth.

Table 8: Cross-effect-family baseline summary. Inclusion is scored against the noisy held-out standardized-effect reference; release is the maximum number released over the prespecified half-width thresholds 0.20–1.00.

Panel	Transport	Hierarchical	Robust PI	Study conformal	Family conformal
Five-family ($n = 10$)	.800	.900	1.000	1.000	.800
All-family ($n = 28$)	.929	.964	1.000	1.000	.929
Max releases, five-family	2	2	0	0	0
Max releases, all-family	2	2	0	0	0

J.5 Baselines and evaluation criteria

The baseline suite includes fixed-effect and DerSimonian–Laird pooling, hierarchical Bayes when a faithful implementation is available, ridge and kernel meta-regression, semantic nearest neighbor, graphical/target-population transport estimators when their required covariates and assumptions are present, and target-only experimentation. We evaluate the framework by joint coverage, release across calibrated tolerances, interval width, target-outcome isolation, and common-estimand harmonization.

J.6 Composition-only ExAtlas-style stress audit

We implemented a composition-only ExAtlas-style baseline (Zhang et al., 2026) using the same data-generating processes, target shifts, and fixed random seeds as the Causal ATLAS evaluation. The audit covers an in-support condition, three observable target shifts, and three hidden moderator shifts, with 200 replications per condition. For each replication, we record whether the archive is judged composable by the baseline and whether Causal ATLAS issues a release certificate.

This controlled audit isolates ExAtlas-style composition from its enrichment, reconciliation, and bridge-generation modules. The complete protocol, per-replication records, summary statistics, and scripts are provided in the accompanying repository.

The composition baseline is more permissive in-support, but its composability declines sharply under observable and hidden shifts. The results distinguish semantic composition from causal transportability; comparisons with the complete ExAtlas system remain outside this audit.

Table 9: Composition-only ExAtlas-style stress audit. Entries are release or composability rates over 200 fixed-seed replications.

Scenario	ExAtlas-style	Causal ATLAS
In-support	82.5%	41.0%
Observable shift 0.25	39.0%	29.5%
Observable shift 0.50	4.5%	7.0%
Observable shift 0.75	0.5%	2.0%
Hidden shift 0.25	77.5%	39.5%
Hidden shift 0.50	25.5%	27.0%
Hidden shift 0.75	10.0%	11.5%

K Machine-Checked Deterministic Core

The module `LeanProofs/CausalAtlasBridge.lean` was added to the existing LeanProofs project and imported by its root module. Lean 4.34.0-rc1 and Lake 5.0.0 verify the following deterministic kernels:

- `simplex_weighted_noise_bound` and `simplex_weighted_bias_bound`, corresponding to the coordinate and nuisance aggregation steps in Theorem 5.2;
- `selection_substitution`, which checks that an eventwise bound quantified over all admissible weights can be evaluated at a data-dependent admissible choice;
- `two_world_absolute_loss`, the deterministic loss inequality used by Theorems 5.1 and 5.4;
- `barycentric_cancellation`, the finite-dimensional identity that removes the first-order term in Lemma C.1;
- `interval_intersection_endpoint_characterization` and four interval-projection lemmas; and
- `complementarity_pi_width_table` plus `complementarity_pi_width_violates_diminishing_returns`, which check widths $(2, 2, 2, 0)$ and the strict failure of diminishing returns.

Three independent checks passed: the lean-prover audit script with warnings treated as errors, `lake env lean -error=warning` on the module, and a complete `lake build` of 8,743 jobs. The source contains no `sorry`, `admit`, user-added `axiom`, or `unsafe` declaration. The `#print axioms` audit for all 12 outputs reports only Mathlib's standard `propext`, `Classical.choice`, and `Quot.sound` dependencies.

Mechanization covers the deterministic kernels listed above. Probability spaces, conditional sub-Gaussian concentration, AIPW expectations, Lindeberg–Feller convergence, Gaussian KL divergence, and the McShane extension theorem are treated by the conventional proofs above.

L Complete Calibration–Transport–Bridge Workflow

L.1 Question, units, and information available to the policy

The preceding synthetic panels test individual claims. This supplement evaluates the complete finite-budget sequence: jointly calibrate mechanism sets, fit an archive composition, release or refuse, acquire a bridge trial, refit, and stop at the first admissible release. The question is whether targeted acquisition improves release probability or acquisition cost relative to random bridges, nearest-mechanism bridges, no acquisition, and a new trial at the target. The comparison includes direct experimentation as a competing policy.

The experimental unit is an independently generated world. Three effect surfaces crossed with three support geometries, each with 200 worlds, give 1,800 independent worlds. Six policies and two frozen tolerances, $\delta \in \{0.2, 0.3\}$, are paired within each world. Four reporting budgets, $0, 96, 192, 288$ new participants, produce 86,400 policy–tolerance–budget records; those records are not independent replicates. The seed is 2026091801. Historical trials contain 800 participants each; each new bridge

or target batch contains 96 participants with equal allocation to the two arms. The four historical trials are treated as sunk costs. Mechanism-audit and intervention-deployment costs are not included in the participant count.

The planning interface receives proxy centers, calibrated radii, noise scales, (L, H), and candidate sample sizes. It does not receive true mechanisms, true effects, or outcomes from any trial. Candidate selection and weights are planned from this information. Observations are generated and read only for trials actually acquired. The target's true effect is used by the scorer, not by the planner. Potential outcomes are shared across methods for paired comparisons without revealing unacquired observations to a policy.

L.2 Data-generating processes and joint calibration

Let $\mathcal{M} = [-1, 1]^2$. The surfaces and analytically valid smoothness bounds are

$$\begin{aligned}
\mu_1(x, y) &= .55x + .30y, & (L_1, H_1) &= (\sqrt{.55^2 + .30^2}, 0),\\
\mu_2(x, y) &= .50x + .20y + .06(x^2 + y^2), & (L_2, H_2) &= (\sqrt{.62^2 + .32^2}, .12),\\
\mu_3(x, y) &= .45x + .20y + .08\sin(2x), & (L_3, H_3) &= (\sqrt{.61^2 + .20^2}, .32).
\end{aligned}$$

Each balanced trial observes independent Gaussian responses with variance $.5^2$ in each arm, control mean zero, and treatment mean $\mu(m)$. Its difference in arm means has exact standard deviation $1/\sqrt{n}$ and zero nuisance bias. The initial archive has four mechanisms and the candidate menu has six. Coordinates in Table 10 receive independent $N(0, .025^2)$ jitter and projection onto $\mathcal{M}$; the target coordinate stays fixed.

Table 10: Support geometries before source jitter. The labels describe the candidate menu, not a guaranteed release decision.

Geometry	Target	Historical archive
Supported	$(.20, .12)$	$(.35, .10), (.15, .30), (.25, -.05), (.05, .15)$
Bridgeable	$(0, -.05)$	$(-.75, -.55), (-.65, -.35), (-.55, -.70), (-.40, -.50)$
Unreachable	$(.88, .82)$	$(-.85, -.65), (-.72, -.55), (-.62, -.75), (-.48, -.58)$
Candidate coordinates, in the same scenario order:		
$(.2, .1), (-.1, .25), (.4, -.2), (-.35, .15), (.05, -.35), (-.2, -.15)$		
$(-.25, -.20), (0, -.05), (.25, .10), (.45, .30), (.65, .45), (.75, .60)$		
$(-.80, -.40), (-.65, -.30), (-.50, -.22), (-.42, -.12), (-.30, -.18), (-.20, -.05)$		

Each world has 199 independent labelled audit archives. Each audit score is the maximum Euclidean norm of 11 independent standard bivariate Gaussian proxy errors, covering all ten possible sources and the target jointly. With $\eta = .05$, the conformal order statistic is $\lceil 200(1 - .05) \rceil = 190$. If this value is q, all sets have the form $U_i = B(c_i, .012q) \cap \mathcal{M}$, where c_i is the projected true mechanism plus independent Gaussian proxy noise of scale .012. Projection is nonexpansive. This synthetic audit assumes access to mechanism labels; it is not an empirical calibration procedure already demonstrated on real interventions.

L.3 Certified bounds, optimization, and finite-round error budget

For simplex weights a, write $\bar{c}_a = \sum_i a_i c_i$ and $D_{ij} = \|c_i - c_j\| + r_i + r_j$ for $i \neq j$, with $D_{ii} = 0$. The implementation uses the smaller of two valid causal-error bounds,

$$B_{\mathrm{bar}}(a) = L\big(\|c_\star - \bar{c}_a\| + r_\star + \sum_i a_i r_i\big) + \frac{H}{4}\sum_{i,j} a_i a_j D_{ij}^2, \tag{L.1}$$

$$B_{\mathrm{Lip}}(a) = L\sum_i a_i(\|c_\star - c_i\| + r_\star + r_i). \tag{L.2}$$

The first bounds the paper's $\sup G$ by the pairwise weighted-variance identity and triangle inequalities. The second bounds the causal approximation error directly, and need not bound G itself. Taking

their minimum is an additional valid tightening on the same mechanism-coverage event, not an exact computation of the paper's original $B_{\mathcal{U}}$. Labels R_2 and R_∞ below identify the stochastic penalty appended to this verified geometric bound.

SLSQP uses analytic gradients, uniform and best-single-source starts, 70 iterations and tolerance 10^{-8}. The best feasible vertex, uniform vector, and previous weight vector extended by zero are retained as fallbacks. The evaluated radius is valid at the saved feasible weights; global optimality is not certified. Targeted selection adds the candidate yielding the smallest prospective feasible radius, then collects its trial. This is a one-step heuristic and is not covered by the log-determinant approximation guarantee. Random and nearest-center orderings use the same R_2 refit; static R_2 acquires nothing.

All policies have a common lower bound of .90 on coverage of their entire output path under this DGP. For R_∞, $\eta = .05$ is paired with a .05 simultaneous coordinate-noise budget over all ten sources. For R_2, .025 covers coordinate noise for refusal intervals and .025 is split over the four design-measurable weighted estimates. The weighted multiplier is $\sqrt{2\log(2 \cdot 4/.025)}$; the refusal multiplier is $\sqrt{2\log(2 \cdot 10/.025)}$. The direct target trial requires no mechanism calibration and splits its .10 budget over three cumulative looks, using $\sqrt{2\log(2 \cdot 3/.10)}$. Correlation between cumulative looks is allowed by the union bound. These are finite-round, per-policy guarantees, not arbitrary-horizon or simultaneous-across-method guarantees.

After a refused release, the reported interval is the intersection over observed sources of $[\widehat{\tau}_i \pm \{q_{\mathrm{coord}} s_i + L(\|c_i - c_\star\| + r_i + r_\star)\}]$. The reported interval is a Lipschitz outer region and may be wider than the sharp smooth-class identified set. An empty intersection is labelled inconsistent. A policy stops when its radius is at most δ; later reporting budgets carry forward that estimate and its actual cost without further acquisition. The target-only policy starts with the whole real line and pools up to three new target batches. Reported path coverage includes both released intervals and refused outer regions.

L.4 Results, adverse cases, and uncertainty

Table 11: Bridgeable menu: final release probability and actual mean new participants at a maximum budget of 288. Each row pools 600 paired worlds, equally across three surfaces.

δ	Policy	Release (%)	Participants
0.2	Targeted R_2	66.50	227.04
0.2	Targeted R_∞	0.00	288.00
0.2	Random R_2	38.50	261.44
0.2	Nearest R_2	0.00	288.00
0.2	Static R_2	0.00	0.00
0.2	Target trial	100.00	288.00
0.3	Targeted R_2	100.00	117.92
0.3	Targeted R_∞	34.67	221.60
0.3	Random R_2	100.00	159.68
0.3	Nearest R_2	100.00	241.60
0.3	Static R_2	0.00	0.00
0.3	Target trial	100.00	96.00

At $\delta = .3$, targeted R_2 saves 41.76 participants (26.15%) relative to random bridges at the same 100% release rate; the paired 95% bootstrap interval for savings is $[37.44, 46.24]$. At $\delta = .2$, its release advantage is 28.00 percentage points, interval $[25.50, 30.50]$, and it saves 34.40 participants, interval $[30.72, 37.92]$. Bootstrap resampling uses 2,000 draws paired by world and stratified by surface. These are exploratory, individual intervals, not familywise claims of superiority.

Direct experimentation is an important competing policy. At $\delta = .3$ it releases in every world using 96 participants, saving 21.92 relative to targeted R_2, interval $[19.84, 24.00]$. At $\delta = .2$ it achieves full release with 288 participants, whereas targeted R_2 refuses 33.50% of worlds. The comparison therefore favors direct experimentation at the stated equal per-participant cost in these settings.

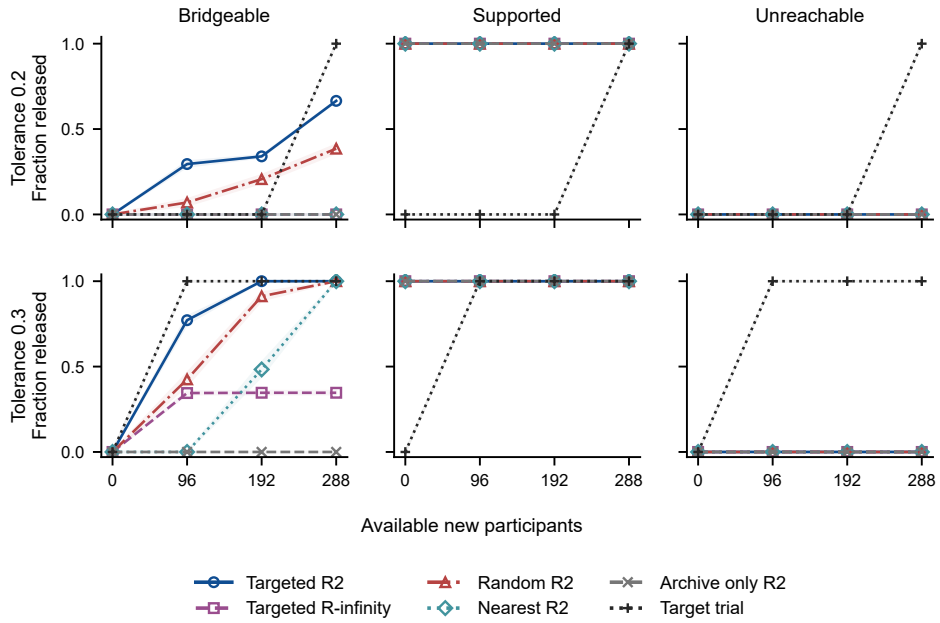

Figure 5: Workflow release curves across all three geometries (columns) and both tolerances (rows), showing all six policies. Stopped policies are carried forward; new-participant budgets and actual expenditure are distinct. Several policies coincide at zero or full release. Table 11 gives the corresponding numerical comparisons.

With supported archives, every archive policy releases at zero new cost for both tolerances. With unreachable menus, every acquiring archive policy spends all 288 participants and still refuses, whereas the static policy refuses without spending. Direct target experimentation succeeds in both geometries. These results motivate a menu-futility rule and a switch to direct experimentation in future implementations.

Table 12: Unpooled bridgeable results, 200 independent worlds per surface. Release columns compare targeted (T) and random (R) bridges. The last column reports target-trial path coverage; archive path coverage is reported in the saved cell summaries.

Surface	δ	T rel. (%)	R rel. (%)	T cost	R cost	Target cov. (%)
Linear	0.2	100.0	89.0	107.04	208.32	100.0
Linear	0.3	100.0	100.0	96.00	132.96	100.0
Quadratic	0.2	99.5	26.5	286.08	288.00	100.0
Quadratic	0.3	100.0	100.0	96.00	153.12	100.0
Sinusoidal	0.2	0.0	0.0	288.00	288.00	99.5
Sinusoidal	0.3	100.0	100.0	161.76	192.96	99.5

Across all 1,800 worlds, the joint mechanism sets cover 94.8889% against a nominal 95%. The minimum observed path coverage among surface–geometry–policy–tolerance cells is 97.5%. Cell-level Wilson intervals, including nonzero upper limits for zero observed failures, are retained in the output. The path statement concerns joint coverage; conditional-on-release coverage is evaluated separately, and the competing methods are not assumed to have identical actual coverage.

For a fixed illustrative linear/bridgeable world (replicate 0, $\delta = .2$), the initial radius is .5831. The first acquired bridge reduces it to .2058, so the method still refuses. A second bridge reduces it to .1687 and triggers release at 192 participants. At the next reporting budget the released estimate and cost remain unchanged. This trace illustrates acquisition and stopping; the population comparison uses every world, not this example.

The rerunnable artifact includes the configuration, 30,600 planning states, 32,883 paired acquisition-log records, world audits, weighted estimates, refusal regions and paired comparisons. Seven workflow tests and 21 output checks cover simplex feasibility, recomputed bounds and estimates, source availability, participant accounting, post-stop immutability and finite-round calibration diagnostics. The results concern dimension two, known Gaussian noise, analytically specified (L, H), measured synthetic audit labels, and a menu of six candidates; broader outcome-adaptive planning and real cross-intervention mechanism calibration require additional studies.

M Real-Data Supplement: Country Holdout and Competing Estimators

M.1 Design and target-outcome isolation

This post-hoc supplement reanalyzes the real Many Labs 2 framing responses from Section 6. The analysis protocol was fixed before this supplemental run, after the source-holdout benchmark had already been examined. The pinned source commit is `acef63fc397b8dce7f0b00f863bcea78d324bea8`; the raw CSV SHA-256 is `15898b5c241696adc7fd91a638839ba49a9f9be64c78da9151b39018d66bfbd3`. Filtering retains included cases, the Cheap/Expensive arms, Yes/No responses, and nonmissing sources. There are 7,228 participants, 57 source samples, and 26 countries. Each source maps to exactly one country in the public metadata.

Leave-one-source-out (LOSO) excludes only the target source. Leave-one-country-out (LOCO) excludes every source in the target's country; each target source is scored separately after fitting to the remaining countries. This tests whether same-country studies materially drive the source-holdout result. The target's arm counts are observed design information. Its responses and effect estimate are hidden until predictions and interval-radius decisions are fixed. A raw-data integration audit flips every held-out Yes/No response, rebuilds source summaries, and refits the complete pipeline. Predictions and radii must remain unchanged for every method under both protocols.

For source s, let $\widehat{p}_{sa}$ be the observed response proportion and n_{sa} the arm count. The common effect and variance inputs are

$$\widehat{\tau}_s = \widehat{p}_{s1} - \widehat{p}_{s0}, \qquad v_s = \frac{\widehat{p}_{s1}(1-\widehat{p}_{s1})}{n_{s1}-1} + \frac{\widehat{p}_{s0}(1-\widehat{p}_{s0})}{n_{s0}-1}.$$

We compare inverse-variance fixed effects, DerSimonian–Laird random effects, the unweighted source mean, the source median, and a trimmed mean deleting $\lfloor .1K \rfloor$ effects from each tail. All five methods receive the identical admissible training sources. The robust estimators are point baselines; no confidence claim is attached to them without an uncertainty model.

For fixed effects, the predictive radius is $z_{.975}\sqrt{(\sum_s v_s^{-1})^{-1} + v_\star^{\max}}$, with $v_\star^{\max} = 1/(4n_{\star 1}) + 1/(4n_{\star 0})$.

The random-effects estimator uses weights $w_s = (v_s + \widehat{\tau}^2)^{-1}$ and radius $t_{.975,K-2}\sqrt{(\sum_s w_s)^{-1} + \widehat{\tau}^2 + v_\star^{\max}}$, where $\widehat{\tau}^2$ is the nonnegative DerSimonian–Laird heterogeneity estimate. These are model-based intervals for the noisy held-out effect estimate. Neither is an application of Theorem 5.2.

M.2 Results and size-matched control

We report source-weighted MAE and country-balanced MAE, which first averages the errors within each country and then averages the 26 country means. Neither treats overlapping training fits as independent experimental replications. Table 13 retains all prespecified point baselines.

The fixed-effect estimator's source MAE changes from .080043 under LOSO to .080428 under LOCO; the random-effects estimator's MAE changes from .080088 to .080397. The median attains the lowest country-balanced MAE in both splits, and the lowest source MAE under LOSO. In this benchmark, the meta-analytic models show no substantial predictive advantage over the simple robust alternatives.

Table 13: Real framing data: absolute error against noisy held-out source estimates. Influence is the largest change in a fitted prediction after removing one training source and refitting, across all target fits. These errors measure predictive fit; calibrated causal transport requires mechanism-set calibration.

Split	Estimator	Source MAE	Country MAE	Max. error	Influence
LOSO	Fixed	0.0800	0.0872	0.2911	0.0038
LOSO	Random	0.0801	0.0872	0.2911	0.0039
LOSO	Unweighted	0.0807	0.0892	0.2985	0.0054
LOSO	Median	0.0794	0.0832	0.2737	0.0030
LOSO	Trimmed	0.0809	0.0895	0.2960	0.0029
LOCO	Fixed	0.0804	0.0873	0.2911	0.0060
LOCO	Random	0.0804	0.0874	0.2911	0.0057
LOCO	Unweighted	0.0811	0.0895	0.2985	0.0078
LOCO	Median	0.0820	0.0838	0.2737	0.0059
LOCO	Trimmed	0.0812	0.0897	0.2960	0.0046

LOCO also removes more training sources. To separate archive size from country exclusion, for each target we draw 100 outcome-independent subsets from its 56 LOSO training sources, each with exactly the corresponding LOCO training count, using seed 2026091607. The fixed-effect source MAE averaged over those draws is .079987, compared with .080428 for actual country exclusion. The difference is .000442 in risk-difference units. These 5,700 target–subset records measure finite-archive sensitivity across 57 targets and do not constitute independent target replications. Country groups and convenience-sample recruitment also limit any population-generalization interpretation.

M.3 Predictive inclusion and release sensitivity

Under LOSO, both model-based intervals include 56 of 57 noisy reference estimates, with mean widths .4041 (fixed) and .4149 (random). Under LOCO, fixed-effect intervals include 55 of 57 and random-effects intervals include 56 of 57, with widths .4052 and .4220. These counts measure inclusion of noisy references; latent causal-effect coverage requires calibrated mechanism sets and noise-free target truth. Table 14 reports the five fixed radius thresholds; choosing a favorable threshold after observing target errors is not part of the evaluation.

Table 14: Release sensitivity for model-based predictive intervals. MAE is conditional on release and evaluated against noisy held-out estimates. LOCO inclusion is conditional on the displayed released subset; these empirical fractions are not conditional coverage guarantees.

Radius	Model	LOSO rel.	MAE	LOCO rel.	MAE	Inclusion
0.10	Fixed	1/57	0.0210	1/57	0.0004	1.000
0.10	Random	1/57	0.0210	1/57	0.0013	1.000
0.15	Fixed	10/57	0.0593	9/57	0.0575	1.000
0.15	Random	6/57	0.0489	6/57	0.0472	1.000
0.20	Fixed	27/57	0.0630	27/57	0.0630	0.963
0.20	Random	21/57	0.0621	21/57	0.0604	1.000
0.25	Fixed	52/57	0.0788	52/57	0.0792	0.962
0.25	Random	52/57	0.0789	52/57	0.0792	0.981
0.30	Fixed	54/57	0.0783	54/57	0.0787	0.963
0.30	Random	54/57	0.0784	54/57	0.0787	0.981

At radius .15, the random-effects procedure releases only six sources in either protocol. Perfect empirical inclusion in six releases is weak evidence about future selective coverage. At .20, the fixed-effect procedure releases 27 sources in both protocols, but LOCO includes 26 of those 27 noisy estimates, compared with 27 of 27 under LOSO. Reporting the whole threshold grid prevents this change from disappearing behind a single selected operating point.

All 114 raw-response perturbation/refit checks produced zero change in predictions and radii across the five estimators. Source IDs used for every fit, all 570 method–split–target records, influence

diagnostics, the size-matched subsets and file hashes are saved by `run_real_supplement.py`. This analysis checks country dependence, competing estimators, training size, and release sensitivity in the real benchmark. The NSW analysis remains a separate within-trial reconstruction audit; its semi-synthetic component is not counted as an additional real experiment.

N Supplementary Selection and Calibration Audits

These supplementary audits use the completed `full` run of `run_all.py`, dated 16 September 2026. They are an independent implementation of the stated simulation protocols, not a replacement for the numerical results in Figure 2. The source hashes, seeds, configuration, and unrounded outputs are recorded in `run_manifest.json`, `selection.csv`, and `joint_calibration_summary.csv`.

N.1 Simultaneous protection and continuous weight selection

Figure 2B uses a Gaussian Bonferroni interval. Table 15 distinguishes that Gaussian specialization from the theorem's sub-Gaussian certificate. Each of 50,000 repetitions draws 128 independent standard-normal source errors with zero target effect and zero causal discrepancy. Max-selection chooses the source with the largest absolute error. Continuous weights are obtained by applying softmax to the absolute errors divided by the displayed temperature. The mixture combines uniform weights with the max-selected vertex, with mixing weight equal to the logistic transform of the maximum absolute error. Seed 2026091602 is shared across the prespecified grid of archive sizes, from 1 to 128.

Table 15: Selection audit at archive size 128. Coverage refers to the zero target effect; brackets give 95% Wilson Monte Carlo intervals. Width is the mean full interval width. Applying the design-measurable certificate to outcome-selected weights is an invalid diagnostic.

Selection	Interval	Coverage [95% MC interval]	Width
Maximum absolute error	Pointwise Gaussian	.00158 [.00127, .00197]	3.9199
Maximum absolute error	Gaussian Bonferroni	.95260 [.95070, .95443]	7.0927
Maximum absolute error	Theorem simultaneous	.99586 [.99526, .99639]	8.2660
Softmax, temperature .25	Theorem simultaneous	.99652 [.99596, .99700]	8.2660
Softmax, temperature .25	Design-based (invalid)	.49174 [.48736, .49612]	3.0706
Softmax, temperature 1	Theorem simultaneous	1.00000 [.99992, 1.00000]	8.2660
Softmax, temperature 1	Design-based (invalid)	.79786 [.79432, .80136]	.6437
Continuous mixture	Theorem simultaneous	.99692 [.99639, .99737]	8.2660

The simultaneous certificate protects both vertex selection and these continuous outcome-dependent weights, at a substantial width cost in this Gaussian setting. These intervals are evaluated without a release threshold, so this audit reports coverage and width rather than release rate. The narrower design-measurable interval applies only when its weight-selection condition holds.

N.2 Joint calibration and distribution shift

The supplementary calibration audit uses 199 independent labelled calibration archives, 11 entities (four archive sources, six candidate bridges, and one target), two proxy coordinates, and $\eta = .05$. The row-wise joint maximum is calibrated at the 190th order statistic and is then frozen before evaluating 2,000 independent test archives. A marginal negative control calibrates individual entity errors and reuses that radius across the full universe. Nominal tests use the matched proxy-error distribution; the stress test multiplies only test errors by 1.5. The new audit uses seed 2026092401; the earlier source-count grid in Table 16 remains a separate diagnostic. The nominal joint coverage target is 95%.

Table 16: Joint coverage over all source coordinates and the additional target coordinate. The shifted test distribution violates the calibration premise. Parentheses give 95% Wilson intervals for nominal joint-max coverage.

Sources	Joint-max, nominal	Marginal-only, nominal	Gaussian union	Joint-max, shifted
4	.9565 (.9467, .9646)	.7895	.9585	.3425
16	.9525 (.9423, .9610)	.4295	.9530	.0780
64	.9510 (.9406, .9596)	.0585	.9515	.0005

Archive-wide control requires joint calibration. Joint-max calibration approximately attains nominal joint coverage under the matched distribution and loses coverage under the specified scale shift. These synthetic labeled errors characterize the tested distribution; the rank argument still requires exchangeability of calibration and test archives.

N.3 Comparison scope and remaining baselines

The controlled comparisons above isolate simultaneous selection protection and joint calibration. The acquisition comparisons in Appendix L isolate bridge policies under a shared implementation. In the pooled bridgeable comparison, targeted acquisition releases all targets at tolerance .3 with 117.92 new participants on average, compared with 159.68 for random bridges; direct target experimentation uses 96 participants. At tolerance .2, targeted acquisition releases 66.5% with 227.04 participants, while the direct target trial releases all targets with 288 participants. In unreachable menus, archive acquisition can exhaust the budget and still refuse. These results describe the tested acquisition trade-offs.

The earlier semantic and point-representation baselines in Appendix I isolate representation effects rather than the complete workflow. Appendix P adds matched-certificate weight comparisons and computational-envelope ablations. These comparisons isolate weighting and geometric bounds on shared bridge paths.

O Source-Level Effect Estimates

P Matched-Certificate Weight and Geometric-Bound Ablations

We conducted a matched-certificate attribution analysis on the same 1,800 synthetic worlds as the workflow experiment. Each world shares calibration data, proxy noise, and source-specific trial streams across variants. We use the data-generating process, seeds, 199 calibration archives per world, and finite-round error budgets of Appendix L. Each bridge acquires 96 participants, with at most three bridges. We fix either a random or nearest-proxy bridge order, so all variants receive the same source prefix at each budget. Under each of the two certificates separately, we compare optimized, uniform, inverse-variance, and nearest-source weights. All weights use design information only: inverse-variance weights use the declared noise variances, and nearest-source weights select the source whose proxy center is closest to the target. The optimizer uses the same SLSQP settings and feasible fallbacks as Appendix L; no global optimality is claimed. We also optimize each geometric bound separately and evaluate all bounds at the same weights. Fixed-budget scores and first-release stopping costs are reported separately.

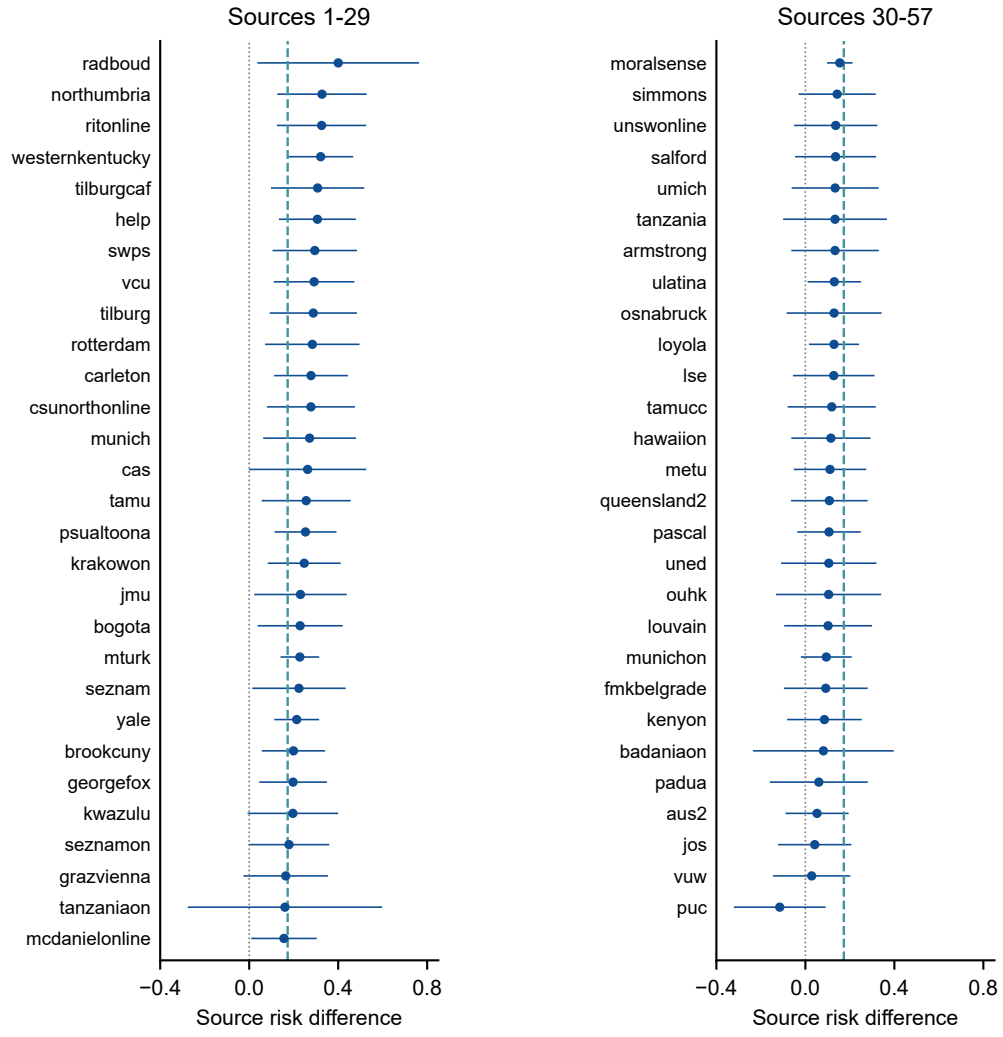

Figure 6: All 57 Many Labs 2 source risk differences, sorted and split across two panels for readability. Bars are source-level 95% normal intervals; the dashed line is the global random-effects mean. These noisy source estimates provide evaluation references, not calibrated mechanism labels.

Table 17: Matched-certificate weighting on the bridgeable menu with a common random bridge order. Each row summarizes 600 paired worlds at a maximum budget of 288 participants. Cost counts actual participants acquired before first release or budget exhaustion.

Certificate	Weights	Tolerance 0.2		Tolerance 0.3	
		Release (%)	Cost	Release (%)	Cost
R_2	Optimized	38.50	261.44	100.00	159.68
R_2	Uniform	1.33	288.00	20.33	283.52
R_2	Inverse variance	0.00	288.00	0.00	288.00
R_2	Nearest source	0.00	288.00	0.00	288.00
R_∞	Optimized	0.00	288.00	31.83	244.96
R_∞	Uniform	0.00	288.00	0.00	288.00
R_∞	Inverse variance	0.00	288.00	0.00	288.00
R_∞	Nearest source	0.00	288.00	0.00	288.00

Table 17 reports the matched-weight comparison. At tolerance 0.3, optimized R_2 weights improve release over uniform weights by 79.67 percentage points, with a paired 95% bootstrap interval of [77.00,82.33], and save 123.84 participants, interval [119.36,128.32]. These are exploratory individual intervals from 2,000 world-paired resamples stratified by response surface. At the common final budget, optimized weights also have lower all-target MAE (0.0334 versus 0.1405) and narrower mean certificates (0.4224 versus 0.7061). Under the nearest-proxy bridge order, optimized R_2 weights release all bridgeable targets at tolerance 0.3 with 241.60 participants on average, while the three simple weight rules release none. At tolerance 0.2, all four rules refuse. The supported menu shows a limit to this benefit: optimized, uniform, and inverse-variance R_2 weights all release without acquisition, and uniform weights have slightly lower MAE. No variant releases on the unreachable menu.

Table 18: Geometric-bound ablation under the same bridgeable random-order protocol with R_2. Each bound is optimized separately using the same stochastic penalty and source prefixes.

Geometry	Tolerance 0.2		Tolerance 0.3	
	Release (%)	Cost	Release (%)	Cost
Barycentric bound	38.50	261.44	100.00	159.68
Lipschitz bound	0.00	288.00	0.00	288.00
Minimum of both	38.50	261.44	100.00	159.68

Table 18 separates the geometric bounds. Across all tested surfaces, support geometries, schedules, certificates, and tolerances, taking the minimum of the two bounds does not improve release probability or stopped cost over the barycentric bound alone. At identical weights the Lipschitz bound sometimes tightens the radius on nonlinear surfaces, mainly in unreachable geometries, without changing release decisions. These results support the role of weighting and barycentric geometry in this DGP; the practical effect of taking the minimum is negligible in these tests. All evaluated policies have zero observed bad releases and full empirical path coverage in this run. Complete cell summaries, fixed-budget scores, saved weights, and paired comparisons are retained with the experiment artifacts.