Proceedings of Machine Learning Research 323:1572–1603, 2026 5th Conference on Causal Learning and Reasoning

Sharp Bounds for Treatment Effect Generalization under Outcome Distribution Shift

Amir Asiaee AMIR.ASIAEETAHERI@VUMC.ORG
Samhita Pal SAMHITA.PAL@VUMC.ORG
Cole Beck COLE.BECK@VUMC.ORG
Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

Jared D. Huling HULING@UMN.EDU
Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, MN, USA

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Generalizing treatment effects from a randomized trial to a target population requires the assumption that potential outcome distributions are invariant across populations after conditioning on observed covariates. This assumption fails when unmeasured effect modifiers are distributed differently between trial participants and the target population. We develop a sensitivity analysis framework that bounds how much conclusions can change when this transportability assumption is violated. Our approach constrains the likelihood ratio between target and trial outcome densities by a scalar parameter $\Lambda \geq 1$, with $\Lambda = 1$ recovering standard transportability. For each Λ, we derive sharp bounds on the target average treatment effect—the tightest interval guaranteed to contain the true effect under all data-generating processes compatible with the observed data and the sensitivity model. We show that the optimal likelihood ratios have a simple threshold structure, leading to a closed-form greedy algorithm that requires only sorting trial outcomes and redistributing probability mass. The resulting estimator runs in $O(n \log n)$ time and is consistent under standard regularity conditions. Simulations demonstrate that our bounds achieve nominal coverage when the true outcome shift falls within the specified Λ, provide substantially tighter intervals than worst-case bounds, and remain informative across a range of realistic violations of transportability.

Keywords: Generalizability; Transportability; Sensitivity analysis; Partial identification; Randomized trials; Treatment effect heterogeneity

1. Introduction

Randomized controlled trials (RCTs) are widely regarded as the gold standard for estimating causal effects because randomization yields strong internal validity. Yet trial participants often differ systematically from the real-world patients or communities to which results are ultimately applied, raising concerns about *external validity*. A large methodological literature studies how to *generalize* or *transport* treatment effects from a trial to a target population by combining trial data with covariate information from the target population, typically via outcome-model-based standardization (g-formula), inverse-odds-of-participation weighting, or augmented/doubly robust estimators that combine both (Cole and Stuart, 2010; Stuart et al., 2011; Buchanan et al., 2018; Dahabreh et al., 2019). Closely related identification questions appear in the causal graphical literature under *transportability* and *data fusion* (Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2016). Recent reviews summarize the resulting toolkit and its applied use across disciplines (Degtiar and Rose, 2023; Ling et al., 2023; Colnet et al., 2024).

We study a canonical trial generalization design with two data sources: a randomized trial and a sample intended to represent the target population. For individuals enrolled in the trial, we observe baseline covariates, randomized treatment assignment, and the outcome. For individuals in the target population sample, we observe the same baseline covariates but not treatment assignments or outcomes. Our inferential goal is the *target average treatment effect* (TATE), i.e., the average causal effect of the trial intervention in the target population. We defer formal notation and the precise definition of the estimand to Section 3.

Standard generalization estimators rely on assumptions that link the trial to the target population. Informally, a common sufficient condition is that, after adjusting for the observed baseline covariates, trial participation is independent of the potential outcomes, so that remaining differences between the trial and the target population are fully explained by their covariate distributions. In addition, identification requires adequate overlap in covariate support between the two populations and the usual causal assumptions for a randomized trial (e.g., consistency and randomization). Under these conditions, the TATE is identified either by reweighting trial participants to represent the target population or by outcome-model-based standardization to the target covariate distribution (Cole and Stuart, 2010; Stuart et al., 2011; Buchanan et al., 2018; Dahabreh et al., 2021). Terminology and precise assumptions vary across communities; see Dahabreh and Hernán (2019) and the synthesis in Degtiar and Rose (2023).

In practice, however, conditional exchangeability across participation is not testable from the observed data and can fail even when rich covariates are available. Multiple mechanisms can cause the conditional outcome distribution to differ between trial and target populations. The most commonly discussed is *unmeasured effect modification*: trial participation may depend on variables that also modify treatment effects, and whose distributions differ across populations. But outcome shift can also arise from *treatment version variability* (the intervention implemented in routine care may differ from the trial protocol (VanderWeele, 2009)) or *network interference* (outcomes in the target may depend on community-level treatment patterns absent in the trial (Hudgens and Halloran, 2008)). All three mechanisms produce the same observable consequence—a shift in conditional outcome distributions between populations—and motivate *sensitivity analyses* that quantify how conclusions change under controlled departures from outcome exchangeability. Several recent proposals develop sensitivity analyses for omitted moderators—including settings where a moderator is observed in the trial but unavailable in the target dataset (Nguyen et al., 2018; Nie et al., 2021; Dahabreh et al., 2023; Huang, 2024). These approaches work by positing a sensitivity model bounding how unobserved moderators relate to selection and/or treatment effect heterogeneity.

We develop a complementary *distributional* sensitivity analysis for trial generalization that directly constrains how much the *conditional outcome distribution* may differ between the trial and target populations after adjusting for observed covariates and treatment arm. Specifically, we introduce an *outcome-shift* model indexed by a single sensitivity parameter $\Lambda \geq 1$. The model restricts the target conditional outcome distribution to lie within a pointwise likelihood-ratio envelope around the trial distribution. This allows flexible, outcome-dependent changes (including shifts concentrated in the tails), while ruling out arbitrarily sharp local spikes that would make sensitivity intervals uninformative. In particular, this is not "uniform density inflation/deflation": the distortion can vary with the outcome value within the envelope. The restriction also implies a corresponding bound on how much conditional means can shift, but not vice versa; we formalize these relationships after introducing notation. At its baseline value $\Lambda = 1$, the model reduces to the

usual outcome-invariance condition underpinning standard generalization methods; as Λ increases, it allows progressively larger departures in a controlled way.

Compared to sensitivity models that bound the density ratio of an unmeasured effect modifier across populations (Nie et al., 2021), our restriction is placed directly on the induced change in the observed outcome distribution within treatment arms. The model is natural in multi-site clinical trials (where site-level differences in care patterns may alter outcome distributions), in policy generalization (where local context may moderate treatment effects), and in health equity applications (where underrepresented populations may have different outcome distributions due to unmeasured social determinants). This perspective is closely related in spirit to likelihood-ratio/odds-ratio sensitivity models for unmeasured confounding in observational studies (Rosenbaum, 2002; Tan, 2006; Zhao et al., 2019; Dorn and Guo, 2023) and to recent work on sharp bounds under generalized causal sensitivity models (Franks et al., 2020; Frauen et al., 2023), but our target is *external validity*: uncertainty enters through possible differences in target-versus-trial outcome distributions rather than by unmeasured confounding of treatment assignment.

Our main contributions are as follows. We derive sharp identification bounds for the target average treatment effect under distributional (pointwise density-ratio) outcome shift between trial and target populations, and we show these bounds can be computed by a closed-form $O(n \log n)$ greedy algorithm. While the mathematical techniques we employ—threshold arguments, linear programming, greedy algorithms—are related to tools used in the marginal sensitivity literature for unmeasured confounding (Tan, 2006; Dorn and Guo, 2023), the problem formulation is different: combining RCT randomization, covariate transport weighting, and distributional outcome shift into a single framework targets external validity rather than internal validity. Concretely:

- **Sharp population-level bounds.** We give a convex characterization of the identified set for target mean potential outcomes and the target ATE under the outcome-shift sensitivity model. We show that the extremal outcome distributions have a simple threshold structure: the optimal likelihood ratio takes only the boundary values Λ^{-1} and Λ, switching at a single outcome threshold. This yields closed-form expressions for the sharp bounds and clarifies how interval width depends on Λ and the variability of trial outcomes.

- **A simple estimation algorithm.** At the sample level, we derive a discrete optimization formulation for the worst-case target means and show that the corresponding linear programs admit a greedy closed-form solution. The algorithm requires only sorting trial outcomes, computing generalization weights, and redistributing probability mass toward extreme outcomes subject to the likelihood ratio constraint. This yields $O(n \log n)$ estimators for the bounds, with consistency following from standard empirical process arguments.

- **Interpretable robustness summary.** We introduce a *tipping-point* sensitivity level $\Lambda^\star$: the smallest outcome-shift magnitude at which the identified set first includes the null. This provides a robustness metric analogous to Rosenbaum's Γ for observational studies (Rosenbaum, 2005): reporting $\Lambda^\star$ answers "how large would the outcome distribution shift have to be to overturn the treatment effect conclusion?" without requiring the analyst to commit to a specific Λ. We distinguish $\Lambda^\star$ (a population identification quantity) from the empirical "breakeven" Λ_{out} reported in our simulations (the smallest Λ achieving $\geq 95\%$ coverage), which additionally incorporates finite-sample estimation error.

2. Related work

Generalizability and transportability of treatment effects. Methods for extending causal conclusions beyond the experimental sample are often framed as generalizability or transportability problems. In the potential outcomes literature, identification and estimation typically proceed by modeling or reweighting the trial sample to match the target covariate distribution using a sampling score (trial participation propensity) (Cole and Stuart, 2010; Stuart et al., 2011; Buchanan et al., 2018), and by clarifying how available data sources (nested vs. non-nested designs) affect identification and estimation (Dahabreh et al., 2019, 2021). In the structural causal model literature, transportability is studied via selection diagrams and data fusion, which characterize when and how causal relations can be moved across environments (Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2016). Comprehensive overviews and applied perspectives appear in Degtiar and Rose (2023), Ling et al. (2023), and Colnet et al. (2024). A related line of work uses observational data to improve trial-based treatment effect estimation under distribution shift between trial and external populations (Asiaee et al., 2023; Asiaee and Pal, 2025; Pal et al., 2025; Karlsson et al., 2025).

Sensitivity analysis for external validity violations. All generalization/transport estimators rely on assumptions that connect trial and target populations, most notably conditions ensuring that the set of observed covariates suffices to adjust for selection into the trial. Because this is inherently uncertain in applications, sensitivity analyses have been proposed to assess robustness to omitted moderators or departures from conditional exchangeability across participation. Nguyen et al. (2018) develop sensitivity analyses when an effect modifier is observed in the trial but not in the target population. Nie et al. (2021) develop a covariate-balancing sensitivity analysis for extrapolating randomized trials across locations, using a bounded conditional density-ratio model for the unmeasured effect modifier across populations to obtain optimization-based identification intervals tightened via covariate moment balancing. Dahabreh et al. (2023) propose bias-function-based sensitivity analyses that directly parameterize violations of generalizability/transportability assumptions in terms of deviations in target counterfactual means. Huang (2024) proposes a sensitivity analysis for weighted generalization estimators in which the key sensitivity parameters can be expressed as a correlation and an explained-variation measure; these are standardized and scale-invariant, and the paper develops benchmarking tools for plausible parameter values. In a complementary direction, Asiaee et al. (2026) develop an omitted-variable-bias decomposition for external validity that yields closed-form sensitivity bounds parameterized by partial R^2 values. Our approach instead bounds the discrepancy between trial and target *conditional outcome distributions* within treatment arms, yielding worst-case identification intervals indexed by a sensitivity parameter Λ.

Partial identification and sharp bounds. Our framework is an instance of partial identification (Manski, 1990, 2003): the sensitivity parameter Λ traces an identification frontier from point identification ($\Lambda = 1$) to worst-case bounds ($\Lambda \to \infty$). We adapt sharp-bound techniques from the marginal sensitivity literature for unmeasured confounding (Rosenbaum, 2002; Tan, 2006; Zhao et al., 2019; Dorn and Guo, 2023; Franks et al., 2020; Frauen et al., 2023), which bounds treatment-assignment odds via pointwise likelihood-ratio constraints, to the external validity setting. The problem formulation is new: combining RCT randomization, covariate transport weighting, and distributional outcome shift into a single framework for the target ATE requires identification arguments specific to generalizability.

3. Problem Formulation and Notation

We adopt the potential outcomes framework for multi-population causal inference, following Dahabreh et al. (2019) with notation adapted for our sensitivity analysis. Let $S \in \{r, o\}$ index the *source* randomized trial ($S = r$) and the *target* population ($S = o$). For unit i in population s, we observe a covariate vector $X_i^s \in \mathcal{X} \subset \mathbb{R}^p$, a binary treatment $A_i^s \in \{-1, +1\}$, and an outcome $Y_i^s \in \mathcal{Y} \subset \mathbb{R}$. Potential outcomes are denoted $(Y_i(+1), Y_i(-1))$, with the usual consistency assumption $Y_i = Y_i(A_i)$.

Throughout, we use the study indicator $s \in \{r, o\}$ as a superscript to denote quantities specific to each population. When we write P^s, $\mathbb{E}^s[\cdot]$, or μ_a^s, this should be understood as the distribution, expectation, or parameter *conditional on membership in study* s—that is, we are characterizing the data-generating process within each population separately.

For treatment level $a \in \{-1, +1\}$, we define the conditional and marginal mean potential outcomes in population s as $\mu_a^s(x) := \mathbb{E}^s[Y(a) \mid X = x]$ and $\mu_a^s := \mathbb{E}^s[Y(a)]$. The average treatment effect (ATE) in population s is $\tau^s := \mu_{+1}^s - \mu_{-1}^s = \mathbb{E}^s[Y(+1) - Y(-1)]$. Our goal is to *generalize* the trial-based treatment effect to the target population—specifically, to learn about $\tau^o = \mathbb{E}^o[Y(+1) - Y(-1)]$—using data from the trial together with covariate information from the target population.

We observe an i.i.d. sample $\mathcal{D}^r := \{(X_i^r, A_i^r, Y_i^r)\}_{i=1}^{n^r}$ from the trial, where treatment is randomized, and a separate i.i.d. sample of covariates $\mathcal{D}^o := \{X_j^o\}_{j=1}^{n^o}$ from the target population. Crucially, we do not observe outcomes or treatment assignments in the target sample, so the conditional outcome distribution in the target, $P^o(Y \mid A, X)$, is not directly identified from the data.

3.1. Identification under standard transportability.

To isolate the role of outcome shift, we assume the usual conditions for internal validity of the trial.

Assumption 1 (Trial internal validity) *For the trial population $S = r$: (i) Consistency: $Y = Y(A)$ a.s.; (ii) Randomization: $(Y(+1), Y(-1)) \perp A \mid X, S = r$; (iii) Positivity: there exists $\eta > 0$ such that $\mathbb{P}^r(A = a \mid X = x) \geq \eta$ for all $a \in \{-1, +1\}$ and almost all x.*

Under Assumption 1, the conditional mean potential outcome is identified from trial data as $\mu_a^r(x) = \mathbb{E}^r[Y \mid A = a, X = x]$.

No outcome shift. The standard approach to generalization further assumes *outcome transportability*: the conditional distribution of potential outcomes is identical across populations,

$$P^r(Y(a) \mid X = x) = P^o(Y(a) \mid X = x) \quad \text{for all } x \text{ and } a, \tag{1}$$

sometimes written compactly as $Y(a) \perp S \mid X$ (Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2016). Under this assumption, combining trial outcomes with the target covariate distribution identifies the target effect. Let $w(x) := dP^o/dP^r(x)$ denote the density ratio between the target and trial covariate distributions. Then

$$\tau^o = \mathbb{E}^o\big[\mu_{+1}^r(X) - \mu_{-1}^r(X)\big] = \mathbb{E}^r\big[w(X) \cdot (\mu_{+1}^r(X) - \mu_{-1}^r(X))\big]. \tag{2}$$

This identity suggests two estimation strategies (Stuart et al., 2011; Dahabreh et al., 2019). The *outcome modeling* approach fits regression models $\hat{\mu}_a^r(x)$ on trial data and averages over the target

covariate sample: $\hat{\tau}^o_{\text{om}} = (n^o)^{-1} \sum_{j=1}^{n^o} [\hat{\mu}^r_{+1}(X^o_j) - \hat{\mu}^r_{-1}(X^o_j)]$. The *inverse probability weighting* approach estimates the density ratio $\hat{w}(x)$ typically via logistic regression of the study indicator S on X using pooled data and computes arm-specific weighted means, $\hat{\mu}^o_{a,\text{ipw}} := \frac{\sum_{i=1}^{n^r} \hat{w}(X^r_i)\, \mathbf{1}\{A^r_i = a\}\, Y^r_i}{\sum_{i=1}^{n^r} \hat{w}(X^r_i)\, \mathbf{1}\{A^r_i = a\}}$, then forms $\hat{\tau}^o_{\text{ipw}} := \hat{\mu}^o_{+1,\text{ipw}} - \hat{\mu}^o_{-1,\text{ipw}}$, with denominators computed separately within each arm.

In practice, outcome transportability (1) may fail even when selection into the trial and treatment assignment are well controlled. This occurs whenever unmeasured effect modifiers have different distributions across populations. Our goal is to relax (1) and develop a sensitivity model that quantifies how violations of outcome transportability affect conclusions about τ^o.

3.2. Unmeasured effect modification

To understand when and why outcome transportability (1) fails, it helps to consider a structural model for potential outcomes. Suppose $Y(a) = m_a(X, U, \varepsilon), a \in \{-1, +1\}$, where U is a (possibly multidimensional) unobserved effect moderator and ε is idiosyncratic noise. We impose no restriction on the joint distribution of (X, U, ε) beyond measurability and integrability, and we keep the structural functions m_a fixed across populations. Critically, we allow the conditional distribution of U given X to differ between trial and target: $P^r(U \mid X) \neq P^o(U \mid X)$. This is the key violation: even after conditioning on observed covariates X, the distribution of the unobserved moderator U may differ across populations.

Under Assumption 1, the conditional mean potential outcome in population s is $\mu^s_a(x) = \int m_a(x, u, \varepsilon)\, dP^s(u, \varepsilon \mid X = x)$. Integrating out ε and defining $\bar{m}_a(x, u) := \mathbb{E}[m_a(x, u, \varepsilon) \mid X = x, U = u]$, we obtain $\mu^s_a(x) = \int \bar{m}_a(x, u)\, dP^s(u \mid X = x)$. The *conditional outcome shift* at covariate value x and treatment a is

$$\Delta_a(x) := \mu^o_a(x) - \mu^r_a(x) = \int \bar{m}_a(x, u)\, d(P^o - P^r)(u \mid X = x), \tag{3}$$

which is nonzero whenever the modifier distribution differs across populations in directions correlated with $\bar{m}_a(x, \cdot)$.

Because the trial is randomized, unmeasured *confounding* of treatment assignment is irrelevant; our concern is unmeasured *effect modification*, which shifts outcome distributions across populations even after adjusting for X.[1] When the conditional mean is linear in a scalar moderator U—i.e., $\mathbb{E}[Y(a) \mid X = x, U = u] = \mu^r_a(x) + \beta_a(x) \cdot u$—equation (3) yields $\Delta_a(x) = \beta_a(x)(\mathbb{E}^o[U \mid X = x] - \mathbb{E}^r[U \mid X = x])$, an omitted-variable-bias decomposition (Nguyen et al., 2018; Huang, 2024). In general, neither $\beta_a(x)$ nor the shift in $P^s(U \mid X)$ is identified. To connect Λ to effect-modification magnitude: for a Gaussian location shift of size δ with variance σ^2 on a bounded support $[L, U]$, the required sensitivity parameter grows as $\Lambda = \exp\big(\delta \max(|L - \mu|, |U - \mu|)/\sigma^2 - \delta^2/(2\sigma^2)\big)$.

4. Outcome-Shift Sensitivity Model

Rather than using a sensitivity model that aims to characterize the behavior of the unobserved moderator U, we bound the discrepancy between outcome distributions in the trial and target populations using a likelihood ratio constraint. This parallels marginal sensitivity models for unmeasured confounding (Tan, 2006; Zhao et al., 2019), but applies the bound to the outcome rather than treatment assignment.

1. For example, if APOE-ε4 carrier frequency differs between trial and target, it can modify drug response without confounding treatment assignment.

4.1. The sensitivity model

Let $f^s(y \mid a, x)$ denote the conditional density of Y given $(A = a, X = x)$ in population $s \in \{r, o\}$. Under randomization in the trial, $f^r(y \mid a, x)$ is identified from the observed data $(Y, A, X, S = r)$.

Definition 2 (Outcome-shift sensitivity model) *Fix a sensitivity parameter $\Lambda \geq 1$. We say that the pair (P^r, P^o) satisfies the* outcome-shift sensitivity model *with parameter Λ if for each $a \in \{-1, +1\}$ and almost all x, the likelihood ratio*

$$r_a(x, y) := \frac{f^o(y \mid a, x)}{f^r(y \mid a, x)}$$

satisfies: (i) $\Lambda^{-1} \leq r_a(x, y) \leq \Lambda$ for all $y \in \mathcal{Y}$ (bounded likelihood ratio); (ii) $\int_{\mathcal{Y}} r_a(x, y)\, f^r(y \mid a, x)\, dy = 1$ for P^r_X-almost every x (normalization).

The normalization condition holds separately for each covariate value x, ensuring that $f^o(\cdot \mid a, x) = r_a(x, \cdot)\, f^r(\cdot \mid a, x)$ is a valid density for each (a, x). Together, the two conditions imply $f^o(y \mid a, x) = r_a(x, y)\, f^r(y \mid a, x)$ with $\Lambda^{-1} \leq r_a(x, y) \leq \Lambda$. The sensitivity parameter Λ controls the degree of departure from transportability. When $\Lambda = 1$, the model enforces $f^o(\cdot \mid a, x) = f^r(\cdot \mid a, x)$, recovering the standard transportability assumption (1). As Λ increases, progressively larger outcome shifts are permitted. For interpretation, $\Lambda = 2$ means the target population can have at most twice (or at least half) the density of any outcome value compared to the trial, conditional on (A, X).

Mean-shift implication. The LR restriction implies $|\mu^o_a(x) - \mu^r_a(x)| \leq (\Lambda - 1)\, \mathbb{E}^r[|Y| \mid A = a, X = x]$, since $|r_a(x, y) - 1| \leq \Lambda - 1$. The converse is false: a mean-shift bound alone permits arbitrary distributional distortions.

Relationship to f-divergence constraints. Our pointwise LR constraint is an L^∞ bound on the density ratio, equivalent to bounding the Rényi ∞-divergence: $D_\infty(P^o \| P^r) \leq \log \Lambda$. This is strictly stronger than any f-divergence constraint, since $D_f(P^o \| P^r) \leq \max_{t \in [1/\Lambda, \Lambda]} f(t)$ for any convex f with $f(1) = 0$, while the converse fails. We note the distinction from Frauen et al. (2023), whose GMSM uses pointwise LR bounds for unmeasured confounding (internal validity); our constraint addresses outcome distributions across populations (external validity).

Given $r_a(x, y)$, the conditional mean potential outcome in the target population is

$$\mu^o_a(x) = \int y\, f^o(y \mid a, x)\, dy = \int y\, r_a(x, y)\, f^r(y \mid a, x)\, dy = \mathbb{E}^r\big[Y \cdot r_a(X, Y) \mid A = a, X = x\big], \tag{4}$$

and the marginal mean is $\mu^o_a = \mathbb{E}^o[\mu^o_a(X)]$. Since $r_a(x, y)$ is not identified from data, neither is μ^o_a. Our goal is to characterize the set of values μ^o_a can take as r_a ranges over all functions satisfying Definition 2.

To connect this to estimation from trial data, let $w(x) := dP^o/dP^r(x)$ denote the density ratio between target and trial covariate distributions, and let $\pi^r(a \mid x)$ denote the (known) trial randomization probability. The target mean can be written as an expectation over the trial distribution:

$$\mu^o_a = \mathbb{E}^o[\mu^o_a(X)] = \mathbb{E}^r\big[w(X) \cdot \mu^o_a(X)\big] = \mathbb{E}^r\left[w(X) \cdot r_a(X, Y) \cdot Y \cdot \frac{\mathbf{1}\{A = a\}}{\pi^r(a \mid X)}\right]. \tag{5}$$

This identity is key for estimation: it expresses the target quantity as a weighted average of trial outcomes, where the weights combine the covariate shift correction $w(X)$, the outcome shift $r_a(X, Y)$, and the inverse propensity weight $\mathbf{1}\{A = a\}/\pi^r(a \mid X)$. Since r_a is unknown, we optimize over admissible likelihood ratios to obtain bounds.

4.2. Sharp identification bounds

For a fixed $\Lambda \geq 1$, define the class of admissible likelihood ratio functions

$$\mathcal{R}_a(\Lambda) := \Big\{ r_a : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \;\Big|\; \Lambda^{-1} \leq r_a(x, y) \leq \Lambda,\ \mathbb{E}^r[r_a(X, Y) \mid X, A = a] = 1 \text{ a.s.} \Big\}.$$

The identified set for μ_a^o is $\mathcal{M}_a(\Lambda) := \{\mu_a^o(r_a) : r_a \in \mathcal{R}_a(\Lambda)\}$, and the identified set for the target ATE is $\mathcal{T}(\Lambda) := \{\mu_{+1}^o(r_{+1}) - \mu_{-1}^o(r_{-1}) : r_a \in \mathcal{R}_a(\Lambda)\}$.

To characterize these sets, we first derive bounds on the conditional mean $\mu_a^o(x)$ for fixed (x, a), then aggregate over the covariate distribution. Define the optimization problems

$$\mu_a^{o,+}(x; \Lambda) := \sup_{r_a \in \mathcal{R}_a(\Lambda)} \int y\, r_a(x, y)\, f^r(y \mid a, x)\, dy \quad \text{and} \tag{6}$$

$$\mu_a^{o,-}(x; \Lambda) := \inf_{r_a \in \mathcal{R}_a(\Lambda)} \int y\, r_a(x, y)\, f^r(y \mid a, x)\, dy. \tag{7}$$

Structure of optimal solutions. The optimization problems (6)–(7) have a special structure: the objective is linear in the likelihood ratio $r_a(x, y)$, and the constraint set is defined by box constraints plus a single linear equality (normalization). A standard result in linear programming implies that optimal solutions to such problems take extreme values—the likelihood ratio equals either Λ^{-1} or Λ almost everywhere, rather than intermediate values.

To build intuition, consider the maximization problem for $\mu_a^o(x)$. We seek a reweighting of the trial outcome distribution that places as much probability mass as possible on large outcome values. The likelihood ratio $r_a(x, y)$ acts as a redistributor of probability: setting $r_a(x, y) = \Lambda$ inflates the probability of outcome y by a factor of Λ, while $r_a(x, y) = \Lambda^{-1}$ deflates it. The normalization constraint enforces that total probability remains one, so inflation at some values must be offset by deflation elsewhere.

Given this trade-off, the optimal strategy is clear: inflate the largest outcomes as much as possible ($r_a = \Lambda$) and deflate the smallest outcomes as much as possible ($r_a = \Lambda^{-1}$). Any intermediate value of r_a wastes capacity that could be better allocated to the extremes. The following lemma formalizes this.

Lemma 3 (Threshold structure) *Fix x and a, and suppose $Y \mid (A = a, X = x, S = r)$ has support contained in a bounded interval $[L, U]$. Any maximizer of (6) takes the threshold form $r_a^\star(x, y) = \Lambda$ for $y > t$, $r_a^\star(x, y) = \Lambda^{-1}$ for $y < t$, and $r_a^\star(x, t) = r_0 \in [\Lambda^{-1}, \Lambda]$, for some threshold $t \in [L, U]$ determined by the normalization constraint. For the minimization problem (7), the optimal solution has the reversed form. The proof is given in Appendix A.1.*

Lemma 3 implies that the conditional identified set for $\mu_a^o(x)$ is the interval $[\mu_a^{o,-}(x; \Lambda), \mu_a^{o,+}(x; \Lambda)]$, with endpoints attained by threshold likelihood ratios. Aggregating over the covariate distribution yields sharp bounds on the marginal quantities.

Theorem 4 (Sharp bounds for the target ATE) *Suppose Assumption 1 holds, $|Y| \leq C$ almost surely in both populations, and (P^r, P^o) satisfies the outcome-shift sensitivity model with parameter $\Lambda \geq 1$. Then the identified set for the target mean potential outcome is*

$$\mathcal{M}_a(\Lambda) = \big[\mu_a^{o,-}(\Lambda),\, \mu_a^{o,+}(\Lambda)\big], \quad \textit{where} \quad \mu_a^{o,\pm}(\Lambda) := \mathbb{E}^o\big[\mu_a^{o,\pm}(X;\Lambda)\big].$$

The identified set for the target ATE is $\mathcal{T}(\Lambda) = \big[\tau^{o,-}(\Lambda),\, \tau^{o,+}(\Lambda)\big]$, with $\tau^{o,-}(\Lambda) := \mu_{+1}^{o,-}(\Lambda) - \mu_{-1}^{o,+}(\Lambda)$, $\tau^{o,+}(\Lambda) := \mu_{+1}^{o,+}(\Lambda) - \mu_{-1}^{o,-}(\Lambda)$. These bounds are sharp*: for every value in the interval, there exists a pair (P^r, P^o) satisfying the sensitivity model and consistent with the observed trial data for which the target ATE equals that value. The proof is provided in Appendix A.2.*

Theorem 4 establishes that the outcome-shift sensitivity model yields a one-dimensional sensitivity analysis indexed by Λ. When $\Lambda = 1$, we recover point identification under standard transportability: $\tau^{o,-}(1) = \tau^{o,+}(1) = \tau^o$. As $\Lambda \to \infty$, the interval $\mathcal{T}(\Lambda)$ expands toward the range of values consistent with the trial data and covariate distribution alone, with no restriction on outcome transportability.

5. Estimation

We now derive sample analogues of the sharp bounds in Theorem 4 and provide an efficient algorithm for computing them.

5.1. Discrete optimization formulation

Fix a treatment level $a \in \{-1, +1\}$ and let $\mathcal{I}_a := \{i \in \{1, \ldots, n^r\} : A_i^r = a\}$ denote the indices of units in the trial receiving treatment a, with $n_a := |\mathcal{I}_a|$. Let $\hat{w}_i := \hat{w}(X_i^r)$ be estimated generalization weights (e.g., from logistic regression of the study indicator S on covariates X using pooled data). Define normalized baseline weights

$$p_i^{(a)} := \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}_a} \hat{w}_j}, \qquad i \in \mathcal{I}_a. \tag{8}$$

These weights sum to one and represent the trial outcome distribution reweighted to match the target covariate distribution.

In the finite-sample setting, the outcome-shift sensitivity model reduces to finding multipliers $\lambda_i^{(a)} \in [\Lambda^{-1}, \Lambda]$ such that the reweighted probabilities $q_i^{(a)} := p_i^{(a)} \lambda_i^{(a)}$ sum to one. The sample bounds on the target mean solve the linear programs

$$\hat{\mu}_a^{o,+}(\Lambda) := \max_{\lambda} \sum_{i \in \mathcal{I}_a} p_i^{(a)} \lambda_i^{(a)} Y_i^r, \qquad \hat{\mu}_a^{o,-}(\Lambda) := \min_{\lambda} \sum_{i \in \mathcal{I}_a} p_i^{(a)} \lambda_i^{(a)} Y_i^r, \tag{9}$$

subject to $\Lambda^{-1} \leq \lambda_i^{(a)} \leq \Lambda$ for all $i \in \mathcal{I}_a$ and $\sum_{i \in \mathcal{I}_a} p_i^{(a)} \lambda_i^{(a)} = 1$. The sample ATE bounds are

$$\hat{\tau}^{o,-}(\Lambda) := \hat{\mu}_{+1}^{o,-}(\Lambda) - \hat{\mu}_{-1}^{o,+}(\Lambda), \qquad \hat{\tau}^{o,+}(\Lambda) := \hat{\mu}_{+1}^{o,+}(\Lambda) - \hat{\mu}_{-1}^{o,-}(\Lambda).$$

5.2. A greedy algorithm

Although (9) is a linear program, its special structure admits a closed-form solution. The feasible set is a polytope defined by box constraints plus a single equality constraint (normalization). A standard result in linear programming implies that extreme points of this polytope have at most one coordinate strictly between its bounds—all other coordinates are either Λ^{-1} or Λ. Combined with the threshold structure from Section 4.2, this means the optimal solution assigns maximum weight Λ to the largest outcomes and minimum weight Λ^{-1} to the smallest, with at most one outcome receiving an intermediate weight to satisfy normalization.

This leads to a simple greedy algorithm. Sort the outcomes in treatment arm a in ascending order: $Y^r_{(1)} \leq Y^r_{(2)} \leq \cdots \leq Y^r_{(n_a)}$, with corresponding baseline weights $p^{(a)}_{(1)}, \ldots, p^{(a)}_{(n_a)}$. Initialize all reweighted probabilities at their lower bound: $q^{\text{low}}_{(j)} := p^{(a)}_{(j)}\Lambda^{-1}$. The total mass at the lower bound is Λ^{-1}, leaving a "budget" of $B := 1 - \Lambda^{-1}$ to distribute. Each outcome j can absorb additional mass up to its capacity $C_{(j)} := p^{(a)}_{(j)}(\Lambda - \Lambda^{-1})$.

To maximize the mean, allocate the budget starting from the largest outcome: fill $q_{(n_a)}$ to capacity, then $q_{(n_a-1)}$, and so on until the budget is exhausted. To minimize the mean, allocate starting from the smallest outcome. The following theorem confirms this procedure is optimal.

Theorem 5 (Greedy algorithm) *Let $Y^r_{(1)} \leq \cdots \leq Y^r_{(n_a)}$ be the sorted outcomes in arm a with baseline weights $p^{(a)}_{(j)}$. Define $q^{\text{low}}_{(j)} := p^{(a)}_{(j)}\Lambda^{-1}$, $C_{(j)} := p^{(a)}_{(j)}(\Lambda - \Lambda^{-1})$, and $B := 1 - \Lambda^{-1}$.*

(Upper bound) *Set $q^{\max}_{(j)} := q^{\text{low}}_{(j)}$ for all j. For $j = n_a, n_a - 1, \ldots, 1$: set $\delta_{(j)} := \min\{C_{(j)}, B\}$, update $q^{\max}_{(j)} \leftarrow q^{\max}_{(j)} + \delta_{(j)}$ and $B \leftarrow B - \delta_{(j)}$; stop when $B = 0$. Then $\hat{\mu}^{o,+}_a(\Lambda) = \sum_{j=1}^{n_a} q^{\max}_{(j)} Y^r_{(j)}$.*

(Lower bound) *Reset $B := 1 - \Lambda^{-1}$ and $q^{\min}_{(j)} := q^{\text{low}}_{(j)}$. For $j = 1, 2, \ldots, n_a$: set $\delta_{(j)} := \min\{C_{(j)}, B\}$, update $q^{\min}_{(j)} \leftarrow q^{\min}_{(j)} + \delta_{(j)}$ and $B \leftarrow B - \delta_{(j)}$; stop when $B = 0$. Then $\hat{\mu}^{o,-}_a(\Lambda) = \sum_{j=1}^{n_a} q^{\min}_{(j)} Y^r_{(j)}$. The proof is presented in Appendix A.3.*

Algorithm 1: Outcome-shift sensitivity bounds

Input: Trial data $\{(X^r_i, A^r_i, Y^r_i)\}_{i=1}^{n^r}$, weights $\hat{w}_i$, $\Lambda \geq 1$
Output: Bounds $[\hat{\tau}^{o,-}, \hat{\tau}^{o,+}]$ on target ATE
for $a \in \{-1, +1\}$ **do**
  $\mathcal{I}_a \leftarrow \{i : A^r_i = a\}$, $n_a \leftarrow |\mathcal{I}_a|$;
  $p^{(a)}_i \leftarrow \hat{w}_i / \sum_{j \in \mathcal{I}_a} \hat{w}_j$ for $i \in \mathcal{I}_a$;
  Sort: $Y^r_{(1)} \leq \cdots \leq Y^r_{(n_a)}$ with weights $p^{(a)}_{(j)}$;
  for $j = 1, \ldots, n_a$ **do**
    $q^{\text{low}}_{(j)} \leftarrow p^{(a)}_{(j)}\Lambda^{-1}$, $C_{(j)} \leftarrow p^{(a)}_{(j)}(\Lambda - \Lambda^{-1})$;
  end
  $B \leftarrow 1 - \Lambda^{-1}$, $q^{\max} \leftarrow q^{\text{low}}$;
  for $j = n_a$ **to** 1 **do**
    $\delta \leftarrow \min\{C_{(j)}, B\}$;
    $q^{\max}_{(j)} \mathrel{+}= \delta$, $B \mathrel{-}= \delta$;
    if $B = 0$ **then break**;
  end
  $\hat{\mu}^{o,+}_a \leftarrow \sum_{j=1}^{n_a} q^{\max}_{(j)} Y^r_{(j)}$;
  $B \leftarrow 1 - \Lambda^{-1}$, $q^{\min} \leftarrow q^{\text{low}}$;
  for $j = 1$ **to** n_a **do**
    $\delta \leftarrow \min\{C_{(j)}, B\}$;
    $q^{\min}_{(j)} \mathrel{+}= \delta$, $B \mathrel{-}= \delta$;
    if $B = 0$ **then break**;
  end
  $\hat{\mu}^{o,-}_a \leftarrow \sum_{j=1}^{n_a} q^{\min}_{(j)} Y^r_{(j)}$;
end
return $\hat{\tau}^{o,-} = \hat{\mu}^{o,-}_{+1} - \hat{\mu}^{o,+}_{-1}$,
  $\hat{\tau}^{o,+} = \hat{\mu}^{o,+}_{+1} - \hat{\mu}^{o,-}_{-1}$

The algorithm requires $O(n_a \log n_a)$ time for sorting plus $O(n_a)$ for the greedy allocation, giving overall complexity $O(n^r \log n^r)$. Algorithm 1 summarizes the complete procedure. We note that ties in outcomes do not affect correctness: when $Y_{(j)} = Y_{(j+1)} = \cdots = Y_{(j+k)}$, any allocation of budget among tied outcomes yields the same objective value, since all share the same outcome value.

5.3. Consistency

The sample bounds are consistent for their population counterparts under standard conditions.

Theorem 6 (Consistency) *Suppose Assumption 1 holds, $|Y| \leq C$ almost surely, and the generalization weights satisfy $\hat{w}_i = w(X_i^r) + o_p(1)$ uniformly, where $w(x) = dP^o/dP^r(x)$ is bounded and strictly positive on the support of $P^{r,X}$. Then as $n^r, n^o \to \infty$, $\hat{\mu}_a^{o,\pm}(\Lambda) \xrightarrow{p} \mu_a^{o,\pm}(\Lambda)$, $\hat{\tau}^{o,\pm}(\Lambda) \xrightarrow{p} \tau^{o,\pm}(\Lambda)$, for each $a \in \{-1, +1\}$. The sample identified set $[\hat{\tau}^{o,-}(\Lambda), \hat{\tau}^{o,+}(\Lambda)]$ converges in Hausdorff distance to the population identified set $[\tau^{o,-}(\Lambda), \tau^{o,+}(\Lambda)]$. The proof is in Appendix A.4.*

6. Experiments

We evaluate the proposed bounds via simulation, examining coverage, sharpness, and comparison to alternatives. We use four data-generating processes (DGPs) with unmeasured effect modification: DGP 1 (linear Gaussian), DGP 2 (nonlinear), DGP 3 (binary outcomes), and DGP 4 (heavy-tailed errors). Full DGP specifications, additional experiments, and extended results appear in Appendix B.

6.1. Sensitivity Envelopes

Figure 1 displays sensitivity envelopes—the sharp bounds as a function of Λ—for a single large dataset ($n^r = 2000$, $n^o = 5000$) from each DGP. As Λ increases from 1 (standard transportability), the identified set widens monotonically. The true target ATE falls within the bounds once Λ is sufficiently large. Binary outcomes (DGP 3) yield notably tighter intervals due to bounded support, while heavy-tailed errors (DGP 4) produce wider intervals for the same Λ.

Practical interpretation. The sensitivity envelopes can be summarized via (i) the *sensitivity curve* showing how bounds widen with Λ; (ii) the *tipping-point* $\Lambda^\star := \inf\{\Lambda \geq 1 : 0 \in [\tau^{o,-}(\Lambda), \tau^{o,+}(\Lambda)]\}$, the smallest outcome-shift magnitude at which the identified set includes the null; and (iii) *benchmarking* against observed modifiers by comparing outcome densities across modifier strata within treatment arms. The interval may fail to cover the true ATE for two distinct reasons: (a) Λ is set below the true density-ratio supremum $\Lambda_{\min}$, so the true distribution lies outside the model (an identification limitation); or (b) finite-sample estimation error, which vanishes by Theorem 6. The parameter Λ bounds the aggregate shift from all unmeasured sources combined.

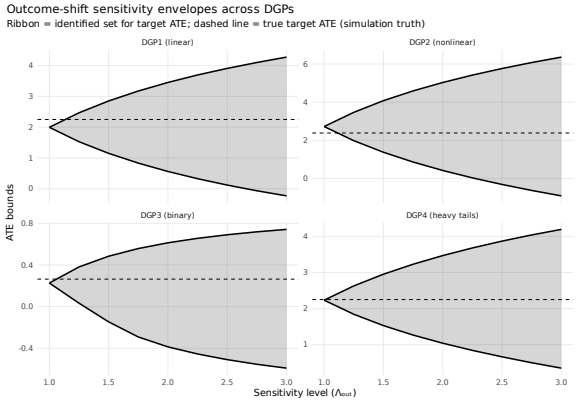

Figure 1: Sensitivity envelopes for DGPs 1–4. Each panel shows the sharp bound interval $[\hat{\tau}^{o,-}(\Lambda), \hat{\tau}^{o,+}(\Lambda)]$ as Λ varies, with the true target ATE marked. Binary outcomes (DGP 3) yield tighter bounds.

6.2. Coverage and Sharpness

We assess repeated-sampling properties using DGP 1 with $(n^r, n^o) = (500, 1000)$ and $R = 1000$ replications. Figure 2 shows coverage and mean interval width as functions of Λ. At $\Lambda = 1$, coverage is near zero because transportability is violated. Coverage increases smoothly, reaching 0.98 at $\Lambda = 1.4$ and 1.00 by $\Lambda = 1.6$. The bounds are sharp: the sharpness ratio (finite-sample width divided by large-n oracle width) exceeds 0.98 across the Λ grid (Table 1).

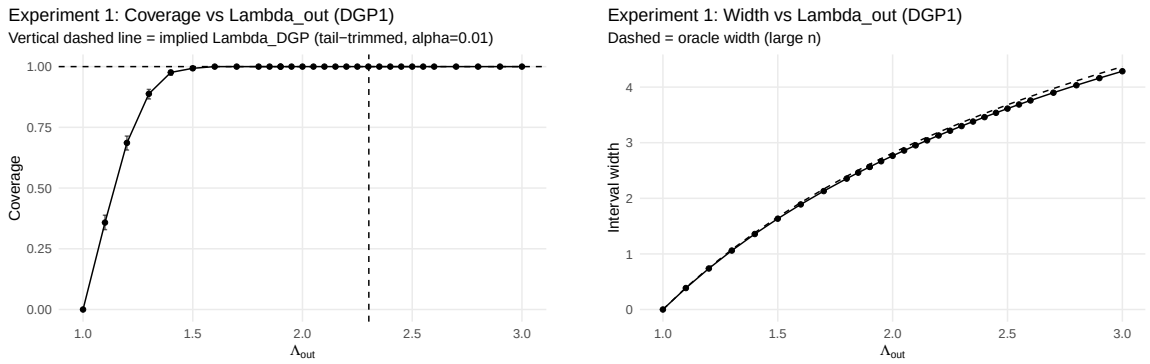

Figure 2: Coverage (left) and mean width (right) vs. Λ for DGP 1 ($R = 1000$). Coverage transitions from undercoverage at $\Lambda = 1$ to nominal levels by $\Lambda \approx 1.5$. Oracle width (dashed) confirms sharpness.

6.3. Comparison to Alternatives

We compare our sharp bounds to three alternatives: (i) the naive transported point estimate ($\Lambda = 1$), (ii) a nonparametric bootstrap CI for the transported estimator, and (iii) worst-case bounds assuming only bounded outcomes. Figure 3 presents results for DGP 1 at $\Lambda = 2.0$ with $R = 500$ replications. The naive point estimate achieves 0% coverage. The bootstrap CI achieves approximately 70% coverage with width 0.80, correctly quantifying sampling but not identification uncertainty. Our sharp bounds achieve 100% coverage with width 2.76. Worst-case bounds also achieve 100% coverage but with width 14.6—over five times wider—demonstrating the value of the structured sensitivity model.

Table 1: Selected operating points for DGP 1 ($R = 1000$). Sharpness ratio confirms finite-sample bounds approximate population bounds.

Λ	Coverage	Mean width	Oracle width	Sharpness
1.4	0.976	1.358	1.382	0.983
1.5	0.993	1.633	1.662	0.983
2.0	1.000	2.764	2.815	0.982

6.4. Identification vs. Estimation

A key conceptual point is that the limitation under violated transportability is *identification*, not estimation. We fix DGP 1 with $n^o = 5000$ and vary $n^r \in \{200, 500, 1000, 2000, 5000\}$ over $R = 200$ replications. Figure 4 shows a striking pattern: as n^r grows, the naive bootstrap CI shrinks but coverage *worsens*, dropping from approximately 50% at $n^r = 200$ to under 10% at $n^r = 5000$. The estimator becomes precisely wrong. Our sharp bounds (at $\Lambda = 2.0$) maintain near-perfect coverage with widths that stabilize rather than collapse—exactly the behavior expected under partial identification.

Table 2: Coverage and width at $\Lambda = 2.0$ across DGPs ($R = 1000$). Complete results are provided in Appendix Table 7.

DGP	Cov.	Width
1 (Linear)	1.000	2.769
2 (Nonlinear)	1.000	4.260
3 (Binary)	1.000	0.977
4 (Heavy-tailed)	1.000	2.545

6.5. Robustness Across Outcome Types

Table 2 summarizes coverage and width across all four DGPs at $\Lambda = 2.0$. The method achieves near-nominal coverage in each case. Binary outcomes (DGP 3) yield the tightest intervals due to bounded support. Nonlinear effect modification (DGP 2) and heavy tails (DGP 4) require somewhat wider intervals, but coverage remains high, suggesting robustness to diverse outcome distributions.

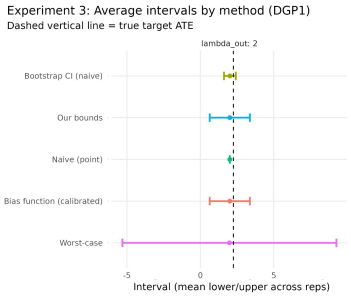

Figure 3: Comparison at $\Lambda = 2.0$ for DGP 1. Naive estimator and bootstrap CI fail to cover; worst-case bounds are uninformatively wide.

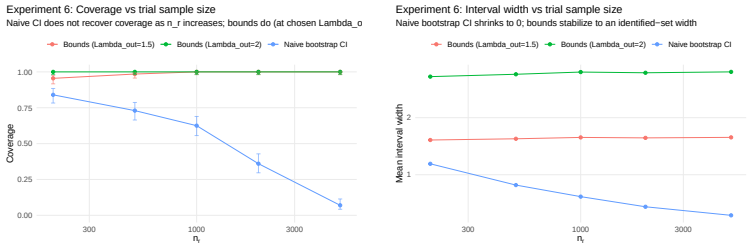

Figure 4: Identification vs. estimation (DGP 1). As n^r grows, naive bootstrap CI shrinks but coverage deteriorates (left). Sharp bounds maintain coverage with stable widths (right).

7. Conclusion and Future Work

We developed a sensitivity analysis framework for generalizing treatment effects when transportability may fail due to unmeasured effect modifiers, bounding the likelihood ratio between target and trial outcome densities by a parameter Λ to obtain sharp identified sets. The key insight is that extremal distributions have a threshold structure—optimal likelihood ratios take only boundary values Λ^{-1} and Λ—enabling a closed-form $O(n \log n)$ greedy algorithm. Simulations confirm nominal coverage when Λ encompasses the true shift, sharpness in finite samples, substantially tighter intervals than worst-case bounds, and correct diagnosis of violated transportability as an identification (not estimation) problem.

Discussion and future work. The present paper establishes identification and consistent estimation of sharp bounds, but does not provide confidence intervals that simultaneously account for sampling variability and partial identification. Because the bound endpoints are functionals of the weighted outcome distribution (resembling weighted tail expectations), they are natural candidates for functional delta-method and bootstrap arguments based on Hadamard differentiability under suitable regularity conditions (Shapiro et al., 2009); developing a complete inferential theory that accounts for estimated generalization weights is valuable future work. Several extensions are also natural. One can allow a *heterogeneous* sensitivity level $\Lambda(a, x)$ to vary by subgroup, retaining convexity because the identification problems are conditional in (a, x); or one can replace the pointwise LR envelope with an *f-divergence* constraint (e.g., a KL ball), which yields a larger ambiguity set (and typically wider bounds) but requires different computation than the greedy algorithm. Finally, our sharp bounds rely on bounded outcome support; with unbounded outcomes, pointwise LR restrictions can lead to vacuous worst-case means, motivating either tail trimming (as we do in simulations) or additional tail/moment restrictions. Practical performance also depends on weight estimation and overlap: uniform weight convergence is strong under poor overlap, so trimming or normalization is often helpful. A real-data case study applying the sensitivity bounds to a concrete trial generalization problem is an important next step.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which substantially improved the paper. This work was supported in part by the Patient-Centered Outcomes Research Institute (PCORI) award ME-2023C1-32148. Code is available at github.com/AsiaeeLab/marginal-sensitivity-for-treatment-effect-generalization.

References

Amir Asiaee and Samhita Pal. Improving RCT-based CATE estimation under covariate mismatch via calibrated alignment. *arXiv preprint arXiv:2603.19186*, 2025.

Amir Asiaee, Chiara Di Gravio, Cole Beck, Yuting Mei, Samhita Pal, and Jared D. Huling. Improving precision of RCT-based CATE estimation using data borrowing with double calibration. *arXiv preprint arXiv:2306.17478*, 2023.

Amir Asiaee, Samhita Pal, and Jared D. Huling. Omitted-variable sensitivity analysis for generalizing randomized trials. *arXiv preprint arXiv:2603.27788*, 2026.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113.

Ashley L. Buchanan, Michael G. Hudgens, Stephen R. Cole, Katie R. Mollan, Paul E. Sax, Eric S. Daar, Adaora A. Adimora, Joseph J. Eron, and Michael J. Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1193–1209, 2018. doi: 10.1111/rssa.12357.

Stephen R. Cole and Elizabeth A. Stuart. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115, 2010. doi: 10.1093/aje/kwq084.

Bénédicte Colnet, Ilyes Mayer, Gaël Varoquaux, Erwan Scornet, and Julie Josse. Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191, 2024. doi: 10.1214/23-STS889.

Issa J. Dahabreh and Miguel A. Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, 2019. doi: 10.1007/s10654-019-00533-2.

Issa J. Dahabreh, Sarah E. Robertson, Eric J. Tchetgen Tchetgen, Elizabeth A. Stuart, and Miguel A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019. doi: 10.1111/biom.13009.

Issa J. Dahabreh, Sebastien J.-P. A. Haneuse, James M. Robins, Sarah E. Robertson, Ashley L. Buchanan, Elizabeth A. Stuart, and Miguel A. Hernán. Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 190(8):1632–1642, 2021. doi: 10.1093/aje/kwaa270.

Issa J. Dahabreh, James M. Robins, Sebastien J.-P. A. Haneuse, Iman Saeed, Sarah E. Robertson, Elizabeth A. Stuart, and Miguel A. Hernán. Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *Statistics in Medicine*, 42(13):2029–2043, 2023. doi: 10.1002/sim.9550.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023. doi: 10.1146/annurev-statistics-042522-103837.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023.

Alexander M. Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115(532):1730–1746, 2020. doi: 10.1080/01621459.2019.1604369.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. *Advances in Neural Information Processing Systems*, 36:40556–40586, 2023.

Melody Y. Huang. Sensitivity analysis for the generalization of experimental results. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 187(4):900–918, 2024. doi: 10.1093/jrsssa/qnae012.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Rickard Karlsson, Piersilvio De Bartolomeis, Issa J. Dahabreh, and Jesse H. Krijthe. Robust estimation of heterogeneous treatment effects in randomized trials leveraging external data. *arXiv preprint arXiv:2507.03681*, 2025.

Albee Y. Ling, Maria E. Montez-Rath, Manisha Desai, Bernard Sebastien, Iman Saeed, and Elizabeth A. Stuart. An overview of current methods for real-world applications to generalize or transport clinical trial findings to target populations of interest. *Epidemiology*, 34(5):627–636, 2023. doi: 10.1097/EDE.0000000000001633.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Charles F Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Trang Quynh Nguyen, Benjamin Ackerman, Ian Schmid, Stephen R. Cole, and Elizabeth A. Stuart. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLOS ONE*, 13(12):e0208795, 2018. doi: 10.1371/journal.pone.0208795.

Xinkun Nie, Guido Imbens, and Stefan Wager. Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*, 2021.

Samhita Pal, Jared D. Huling, and Amir Asiaee. Improving RCT-based CATE estimation under covariate mismatch via double calibration. *arXiv preprint arXiv:2603.17066*, 2025.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 247–254. AAAI Press, 2011.

Paul R. Rosenbaum. *Observational Studies*. Springer, New York, 2 edition, 2002.

Paul R Rosenbaum. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*, 4:1809–1814, 2005.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

Elizabeth A. Stuart, Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011. doi: 10.1111/j.1467-985X.2010.00673.x.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Tyler J VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

Qingyuan Zhao, Dylan S. Small, and Bhaswar B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

Appendix A. Proofs

A.1. Proof of Lemma 3

Write P for the conditional distributions of $Y \mid (A = a, X = x, S = r)$ on $[L, U]$, so $dP(y) = f^r(y \mid a, x)\, dy$. Consider the maximization problem; the minimization case is analogous with the ordering reversed.

Define the feasible set

$$\mathcal{C} := \Big\{ r : [L, U] \to [\Lambda^{-1}, \Lambda] \;\Big|\; \int r(y)\, dP(y) = 1 \Big\}.$$

This set is convex and weak-$\star$ compact in $L^\infty(P)$, and the objective $r \mapsto \int y\, r(y)\, dP(y)$ is linear and continuous, so a maximizer exists.

Step 1 (bang–bang/extreme-point structure). We claim that any extreme point of $\mathcal{C}$ takes values in $\{\Lambda^{-1}, \Lambda\}$ P-a.s., possibly except on a P-null set. Indeed, if $r \in \mathcal{C}$ satisfies $\Lambda^{-1} < r(y) < \Lambda$ on a set E with $P(E) > 0$, we can choose measurable $E_1, E_2 \subseteq E$ with $P(E_1) = P(E_2) > 0$ and define for small $\epsilon > 0$,

$$r_\pm(y) := r(y) \pm \epsilon(\mathbf{1}_{E_1}(y) - \mathbf{1}_{E_2}(y)).$$

For ϵ small enough, $r_\pm \in \mathcal{C}$ and $r = (r_+ + r_-)/2$, so r is not extreme.

Step 2 (no inversions $\Rightarrow$ threshold form). Let r be a maximizer that is an extreme point, hence $r(y) \in \{\Lambda^{-1}, \Lambda\}$ P-a.s. Define $H := \{y : r(y) = \Lambda\}$ and $Lw := \{y : r(y) = \Lambda^{-1}\}$. Suppose for contradiction that there exist $y_1 < y_2$ with $y_1 \in H$ and $y_2 \in Lw$, with both points belonging to sets of positive P-mass. Then there exist measurable sets $E_1 \subseteq H$ and $E_2 \subseteq Lw$ with $P(E_1) = P(E_2) > 0$ and such that $\sup E_1 < \inf E_2$ (take small neighborhoods and trim). Define $\tilde{r}$ by swapping the values on E_1 and E_2:

$$\tilde{r}(y) := \begin{cases} \Lambda^{-1}, & y \in E_1, \\ \Lambda, & y \in E_2, \\ r(y), & \text{otherwise.} \end{cases}$$

Then $\tilde{r} \in \mathcal{C}$ because $\int \tilde{r}\, dP = \int r\, dP$ (the swap changes the integral by $(\Lambda^{-1} - \Lambda)P(E_1) + (\Lambda - \Lambda^{-1})P(E_2) = 0$). Moreover, the objective strictly increases:

$$\int y\, \tilde{r}(y)\, dP(y) - \int y\, r(y)\, dP(y) = (\Lambda - \Lambda^{-1})\Big(\int_{E_2} y\, dP(y) - \int_{E_1} y\, dP(y) \Big) \;>\; 0,$$

since all values in E_2 exceed all values in E_1. This contradicts optimality.

Therefore, up to P-null sets, the set $H = \{r = \Lambda\}$ must be an upper tail: if $y \in H$ and $y' > y$, then $y' \in H$ P-a.s. Hence there exists a threshold $t \in [L, U]$ such that $r(y) = \Lambda$ for $y > t$ and $r(y) = \Lambda^{-1}$ for $y < t$, P-a.s.

Step 3 (atoms/ties at the threshold). If $P(\{t\}) = 0$, the normalization constraint determines t uniquely (up to null sets). If $P(\{t\}) > 0$, the normalization constraint may require assigning an intermediate value on the atom at t: set $r(t) = r_0 \in [\Lambda^{-1}, \Lambda]$ so that $\int r\, dP = 1$. This yields exactly the stated threshold form. ■

A.2. Proof of Theorem 4

Interval structure. Fix $a \in \{-1, +1\}$ and $x \in \mathcal{X}$. By Lemma 3, the conditional mean $\mu_a^o(x) = \int y\, r_a(x, y)\, f^r(y \mid a, x)\, dy$ is maximized and minimized over $\mathcal{R}_a(\Lambda)$ by threshold likelihood ratios, yielding the interval $[\mu_a^{o,-}(x; \Lambda), \mu_a^{o,+}(x; \Lambda)]$. Since the map $r_a \mapsto \mu_a^o(x)$ is linear and continuous, and the constraint set $\mathcal{R}_a(\Lambda)$ is convex and compact (in the weak-$\star$ topology on L^∞), every value in this interval is attained by some admissible r_a.

For the marginal mean, we have

$$\mu_a^{o,+}(\Lambda) = \mathbb{E}^o[\mu_a^{o,+}(X; \Lambda)] = \int \mu_a^{o,+}(x; \Lambda)\, dP^{o,X}(x),$$

and similarly for $\mu_a^{o,-}(\Lambda)$. Since $|Y| \leq C$ almost surely, the conditional bounds satisfy $|\mu_a^{o,\pm}(x; \Lambda)| \leq C$ for all x, so the integrals are well-defined. The identified set for μ_a^o is thus $\mathcal{M}_a(\Lambda) = [\mu_a^{o,-}(\Lambda), \mu_a^{o,+}(\Lambda)]$.

For the target ATE, note that $\tau^o = \mu_{+1}^o - \mu_{-1}^o$ where μ_{+1}^o and μ_{-1}^o can be chosen independently (the likelihood ratios r_{+1} and r_{-1} are separate). Hence the identified set is

$$\mathcal{T}(\Lambda) = \{\mu_{+1}^o - \mu_{-1}^o : \mu_a^o \in \mathcal{M}_a(\Lambda),\, a \in \{-1, +1\}\} = \left[\mu_{+1}^{o,-} - \mu_{-1}^{o,+},\, \mu_{+1}^{o,+} - \mu_{-1}^{o,-}\right].$$

Sharpness. Fix $a \in \{-1, +1\}$. Let r_a^+ and r_a^- be measurable maximizers/minimizers of the conditional problems (they exist by compactness/continuity as above), so that $\mathbb{E}^o[\mu_a^o(X; r_a^+)] = \mu_a^{o,+}(\Lambda)$ and $\mathbb{E}^o[\mu_a^o(X; r_a^-)] = \mu_a^{o,-}(\Lambda)$.

For any $\alpha \in [0, 1]$, define the convex combination

$$r_{a,\alpha}(x, y) := \alpha r_a^+(x, y) + (1 - \alpha) r_a^-(x, y).$$

Since the constraints defining $\mathcal{R}_a(\Lambda)$ are pointwise bounds and a linear normalization condition, $r_{a,\alpha} \in \mathcal{R}_a(\Lambda)$ for all $\alpha \in [0, 1]$. Moreover, the induced marginal mean is the corresponding convex combination:

$$\overset{o}{\mathbb{E}}\big[\mu_a^o\big(X; r_{a,\alpha}\big)\big] = \alpha\, \mu_a^{o,+}(\Lambda) + (1 - \alpha)\, \mu_a^{o,-}(\Lambda).$$

Hence every μ_a^o in the interval $[\mu_a^{o,-}(\Lambda), \mu_a^{o,+}(\Lambda)]$ is attainable by choosing $\alpha = (\mu_a^o - \mu_a^{o,-})/(\mu_a^{o,+} - \mu_a^{o,-})$ (and any α if the endpoints coincide).

For the ATE, note r_{+1} and r_{-1} are chosen independently, so any pair (μ_{+1}^o, μ_{-1}^o) in the product of intervals is attainable, and therefore any difference $\tau = \mu_{+1}^o - \mu_{-1}^o$ in $[\tau^{o,-}(\Lambda), \tau^{o,+}(\Lambda)]$ is attainable. ■

A.3. Proof of Theorem 5

Work in q-variables: $q_{(j)} := p_{(j)}^{(a)} \lambda_{(j)}$. The constraints $\Lambda^{-1} \leq \lambda_{(j)} \leq \Lambda$ and $\sum_j p_{(j)}^{(a)} \lambda_{(j)} = 1$ are equivalent to

$$q_{(j)} \in [q_{(j)}^{\text{low}}, q_{(j)}^{\text{high}}], \qquad \sum_{j=1}^{n_a} q_{(j)} = 1,$$

where $q_{(j)}^{\text{low}} := p_{(j)}^{(a)} \Lambda^{-1}$ and $q_{(j)}^{\text{high}} := p_{(j)}^{(a)} \Lambda$. The objective is $\sum_j q_{(j)} Y_{(j)}^r$.

Feasibility of the greedy construction. The algorithm starts at $q_{(j)}^{\text{low}}$ (which sums to Λ^{-1}) and distributes the remaining mass $\Delta_\Lambda = 1 - \Lambda^{-1}$ by adding at most $C_{(j)} = q_{(j)}^{\text{high}} - q_{(j)}^{\text{low}}$ to each coordinate, so the resulting q stays within the box constraints and sums to one.

Optimality (upper bound). Let q be any feasible point. Suppose there exist indices $i < j$ such that $Y^r_{(i)} < Y^r_{(j)}$, $q_{(i)} > q^{\text{low}}_{(i)}$, and $q_{(j)} < q^{\text{high}}_{(j)}$. Define

$$\varepsilon := \min\{q_{(i)} - q^{\text{low}}_{(i)},\ q^{\text{high}}_{(j)} - q_{(j)}\} > 0,$$

and set $q'_{(i)} := q_{(i)} - \varepsilon$, $q'_{(j)} := q_{(j)} + \varepsilon$, leaving all other coordinates unchanged. Then q' remains feasible (sum preserved; still within bounds) and strictly improves the objective:

$$\sum_k q'_{(k)} Y^r_{(k)} - \sum_k q_{(k)} Y^r_{(k)} = \varepsilon\,(Y^r_{(j)} - Y^r_{(i)}) \ > \ 0.$$

Therefore, at any maximizer there cannot exist such a pair (i, j): whenever a larger outcome j has not been filled to its upper bound, all smaller outcomes must sit at their lower bounds. This forces the "fill from the top" structure implemented by the greedy algorithm, with at most one partially filled coordinate (where the remaining mass runs out).

Optimality (lower bound). The minimization case is identical after reversing the ordering: if a smaller outcome has not been filled while a larger one has received extra mass above its lower bound, shifting mass downward decreases the objective. Hence the greedy "fill from the bottom" procedure is optimal. ■

A.4. Proof of Theorem 6

Fix $a \in \{-1, +1\}$ and $\Lambda \geq 1$. Let $\mathcal{I}_a = \{i : A^r_i = a\}$. Define the (arm-specific) weighted empirical measure of outcomes

$$\hat{P}_{n,a} := \sum_{i \in \mathcal{I}_a} p^{(a)}_i\, \delta_{Y^r_i}, \qquad p^{(a)}_i := \frac{\hat{w}_i}{\sum_{j \in \mathcal{I}_a} \hat{w}_j}.$$

Let P^w_a denote the population analogue obtained by reweighting the trial arm-a distribution by the true density ratio $w(X) = dP^{o,X}/dP^{r,X}(X)$: for any Borel set $B \subseteq [-B, B]$,

$$P^w_a(B) := \frac{\mathbb{E}^r\big[w(X)\,\mathbf{1}\{A = a\}\,\mathbf{1}\{Y \in B\}\big]}{\mathbb{E}^r\big[w(X)\,\mathbf{1}\{A = a\}\big]}.$$

The denominator is strictly positive by positivity of $\mathbb{P}^r(A = a)$ and bounded, strictly positive w on the support of $P^{r,X}$.

Step 1: $\hat{P}_{n,a} \Rightarrow P^w_a$ (in a strong one-dimensional sense). Let $\tilde{p}^{(a)}_i := w(X^r_i)/\sum_{j \in \mathcal{I}_a} w(X^r_j)$ and $\tilde{P}_{n,a} := \sum_{i \in \mathcal{I}_a} \tilde{p}^{(a)}_i \delta_{Y^r_i}$. Uniform consistency $\sup_i |\hat{w}_i - w(X^r_i)| = o_p(1)$ and boundedness of w imply

$$\sum_{i \in \mathcal{I}_a} \big|p^{(a)}_i - \tilde{p}^{(a)}_i\big| = o_p(1), \qquad \text{hence} \qquad d_{\text{TV}}(\hat{P}_{n,a}, \tilde{P}_{n,a}) = o_p(1).$$

Moreover, since the weights $w(X)$ are bounded and the class $\{\mathbf{1}\{y \leq t\} : t \in \mathbb{R}\}$ is VC, a weighted Glivenko–Cantelli theorem yields

$$\sup_{t \in \mathbb{R}} \Big|\tilde{P}_{n,a}((-\infty, t]) - P^w_a((-\infty, t])\Big| \xrightarrow{p} 0.$$

Because Y is supported on $[-B, B]$, uniform convergence of CDFs implies convergence in W_1 (Wasserstein-1) distance, and therefore $W_1(\hat{P}_{n,a}, P^w_a) \xrightarrow{p} 0$.

Step 2: continuity of the sharp-bound functionals. For any probability measure P supported on $[-B, B]$, define

$$\phi_\Lambda^+(P) := \sup_{\lambda:\ \Lambda^{-1}\le\lambda\le\Lambda,\ \int\lambda\,dP=1} \int y\,\lambda(y)\,dP(y), \qquad \phi_\Lambda^-(P) := \inf_{\lambda:\ \Lambda^{-1}\le\lambda\le\Lambda,\ \int\lambda\,dP=1} \int y\,\lambda(y)\,dP(y).$$

The sample LP in (9) is exactly $\hat{\mu}_a^{o,\pm}(\Lambda) = \phi_\Lambda^\pm(\hat{P}_{n,a})$, and the population sharp bounds satisfy $\mu_a^{o,\pm}(\Lambda) = \phi_\Lambda^\pm(P_a^w)$.

Let $Q_P(u) := \inf\{y : P((-\infty, y]) \ge u\}$ be the generalized quantile function and set $\rho := 1/(\Lambda+1)$. A standard rearrangement/knapsack argument (the continuous analogue of Theorem 5) gives the representation

$$\phi_\Lambda^+(P) = \Lambda^{-1}\int y\,dP(y) + (\Lambda - \Lambda^{-1})\int_{1-\rho}^1 Q_P(u)\,du, \quad \phi_\Lambda^-(P) = \Lambda^{-1}\int y\,dP(y) + (\Lambda - \Lambda^{-1})\int_0^\rho Q_P(u)\,du,$$

where the generalized quantile handles atoms and corresponds to allowing one partial weight at the threshold.

In one dimension, $W_1(P,Q) = \int_0^1 |Q_P(u) - Q_Q(u)|\,du$, and also $|\int y\,dP(y) - \int y\,dQ(y)| \le W_1(P,Q)$ for measures on a bounded interval. Therefore,

$$|\phi_\Lambda^\pm(P) - \phi_\Lambda^\pm(Q)| \le \Lambda\, W_1(P,Q),$$

so $\phi_\Lambda^\pm$ is continuous (indeed Lipschitz) in W_1.

Combining with Step 1 yields $\hat{\mu}_a^{o,\pm}(\Lambda) = \phi_\Lambda^\pm(\hat{P}_{n,a}) \xrightarrow{p} \phi_\Lambda^\pm(P_a^w) = \mu_a^{o,\pm}(\Lambda)$.

Step 3: ATE bounds and Hausdorff convergence. By definition, $\hat{\tau}^{o,-}(\Lambda) = \hat{\mu}_{+1}^{o,-}(\Lambda) - \hat{\mu}_{-1}^{o,+}(\Lambda)$ and $\hat{\tau}^{o,+}(\Lambda) = \hat{\mu}_{+1}^{o,+}(\Lambda) - \hat{\mu}_{-1}^{o,-}(\Lambda)$, so convergence of the means implies $\hat{\tau}^{o,\pm}(\Lambda) \xrightarrow{p} \tau^{o,\pm}(\Lambda)$. Finally, for intervals on $\mathbb{R}$ the Hausdorff distance equals the maximum endpoint error, hence

$$d_H([\hat{\tau}^{o,-}(\Lambda), \hat{\tau}^{o,+}(\Lambda)], [\tau^{o,-}(\Lambda), \tau^{o,+}(\Lambda)]) \xrightarrow{p} 0$$

. ■

Appendix B. Extended Simulation Study

This section empirically validates the outcome-shift sensitivity bounds, illustrates the coverage–informativeness tradeoff as the sensitivity parameter varies, and documents robustness to nonlinearity, discrete outcomes, heavy tails, and weight estimation.

B.1. Estimand, data structure, and sensitivity parameter

We consider a randomized trial conducted in a *trial population* (study indicator $S = r$) and a *target population* ($S = o$). In the trial we observe i.i.d. draws $\{(X_i, A_i, Y_i)\}_{i=1}^{n^r}$, where $A \in \{-1, +1\}$ is randomized with $\mathbb{P}(A = +1) = 1/2$, and in the target we observe baseline covariates $\{X_j\}_{j=1}^{n^o}$ (no outcomes).

The target estimand is the target average treatment effect (ATE)

$$\tau^o \;=\; \mathbb{E}[\,Y^o(+1) - Y^o(-1)\,],$$

where $Y^o(a)$ denotes the potential outcome under treatment level a in the target population.

Table 3: Simulation design summary (master table; can be trimmed).

Exp.	Goal	DGP(s)	n^r	n^o	Reps	Key outputs
Main	Envelopes vs. Λ_{out}	1–4	2000	5000	1 dataset	Fig. 5
1	Validation + sharpness	1	500	1000	1000	Fig. 6, Tab. 4
2	Tradeoff vs. moderator shift	1	500	1000	1000	Fig. 7, Tab. 5
3	Baseline comparison	1	500	1000	50	Fig. 8, Tab. 6
4	Scaling in n^r (fixed Λ_{out})	1	100–5000	1000	300	Fig. 9
5	Robustness across outcome types	1–4	500	1000	1000	Fig. 10, Tab. 7
6	Weight estimation sensitivity	1	500	1000	150	Fig. 11
6b	Identification vs estimation	1	200–5000	5000	200	Fig. 12
7	Bounded-support DGP (LR holds literally)	7	500	1000	200	Fig. 13
8	"Bang–bang" optimizer structure	1	1 dataset	–	–	Fig. 14

Outcome-shift sensitivity model and Λ_{out}. The outcome-shift model in section 4 and theorem 2 constrains how much the *conditional outcome distribution* in the target can differ from the trial. In our simulation code and figures we denote the sensitivity parameter by $\Lambda_{\text{out}} \geq 1$ (this is the same Λ as in the main text, with an "out" subscript only to emphasize outcome shift). Concretely, for each treatment arm $a \in \{-1, +1\}$ and covariate profile x, we bound the conditional likelihood ratio by

$$\frac{1}{\Lambda_{\text{out}}} \leq \frac{f^o_{Y|A,X}(y \mid a, x)}{f^r_{Y|A,X}(y \mid a, x)} \leq \Lambda_{\text{out}} \qquad \text{for all } y,$$

where $f^s_{Y|A,X}$ is the conditional outcome density/mass in population $s \in \{r, o\}$. When $\Lambda_{\text{out}} = 1$, this reduces to standard outcome transportability and the bounds collapse to the usual transported point estimate.

Generalization weights. To transport trial outcomes to the target covariate distribution we use inverse odds weights (see eq. (8)). In most experiments we use *oracle* weights $w(x) = f^o_X(x)/f^r_X(x)$ (available in simulation) to isolate outcome-shift effects from weight-estimation error. We separately study estimated weights in appendix B.9.

Metrics. Across Monte Carlo replications we report: (i) *coverage* of the true τ^o by the bound interval; (ii) *mean interval width*; (iii) a *sharpness ratio* defined as (sample mean width)/(oracle width), where "oracle width" is the width obtained by running the same procedure on an extremely large simulated trial to approximate the population sharp interval; and (iv) runtime where relevant. For coverage curves we include Wilson binomial intervals (where plotted; see tables for corresponding [lo, hi]).

B.2. Data-generating processes (DGPs 1–4)

All DGPs share a covariate shift between trial and target and an unmeasured moderator whose distribution differs across populations. Let $p = 5$ and write $X = (X_1, \ldots, X_p)^\top$. We generate

$$X^r \sim \mathcal{N}(0_p, I_p), \qquad X^o \sim \mathcal{N}(\mu_{\text{shift}} \cdot \mathbf{1}_p, I_p),$$

with $\mu_{\text{shift}} = 0.5$ in the main experiments. The unmeasured moderator satisfies

$$U^s \mid X \sim \mathcal{N}(\gamma_s X_1, 1), \qquad s \in \{r, o\},$$

and we set $\gamma_r = 0$ and (unless otherwise stated) $\gamma_o = 0.5$, so that the target has a different distribution of the effect modifier even after conditioning on X.

DGP 1 (linear outcome with Gaussian moderator). Potential outcomes are

$$Y(a) \;=\; \beta_0 + \beta_x^\top X + a \cdot \big(\tau_0 + \beta_u U\big) + \varepsilon_a, \qquad \varepsilon_a \sim \mathcal{N}(0, \sigma^2),$$

with $\beta_0 = 1$, $\beta_x = (0.3, 0.2, 0.1, 0.1, 0.1)^\top$, $\tau_0 = 1$, $\beta_u = 0.5$, and $\sigma = 1$. Because $A \in \{-1, +1\}$, the individual treatment effect is $Y(+1) - Y(-1) = 2(\tau_0 + \beta_u U)$, hence the target ATE has a closed form:

$$\tau^o \;=\; 2\tau_0 + 2\beta_u\, \mathbb{E}^o[U] \;=\; 2\tau_0 + 2\beta_u\, \gamma_o\, \mathbb{E}^o[X_1] \;=\; 2\tau_0 + 2\beta_u\, \gamma_o\, \mu_{\text{shift}} \;=\; 2 + \gamma_o \mu_{\text{shift}} \;=\; 2.25.$$

DGP 2 (nonlinear effect modification). Covariates and U are generated as in DGP 1, but the outcome model is nonlinear:

$$Y(a) \;=\; \beta_0 + \sin(\pi X_1/2) + X_2^2 \;+\; a \cdot \Big(\tau_0 + \beta_u\, |U| \cdot \operatorname{sign}(X_1)\Big) + \varepsilon_a, \qquad \varepsilon_a \sim \mathcal{N}(0, \sigma^2).$$

This preserves the same covariate shift and moderator shift but induces nonlinear treatment effect heterogeneity. The true τ^o is approximated via Monte Carlo with $n_{\text{truth}} = 10^6$ draws from the target joint distribution of (X, U).

DGP 3 (binary outcomes). We generate binary outcomes via

$$Y(a) \sim \text{Bernoulli}(p_a(X, U)), \qquad \operatorname{logit}\big(p_a(X, U)\big) = \beta_0 + \beta_x^\top X + a \cdot (\tau_0 + \beta_u U),$$

with $\beta_0 = -0.5$, $\beta_x = (0.3, 0.2, 0.1, 0.1, 0.1)^\top$, $\tau_0 = 0.5$, and $\beta_u = 0.3$. The true τ^o is approximated via Monte Carlo with $n_{\text{truth}} = 10^6$ draws.

DGP 4 (heavy-tailed outcomes). This matches DGP 1 except the error is heavy-tailed:

$$Y(a) = \beta_0 + \beta_x^\top X + a \cdot (\tau_0 + \beta_u U) + \varepsilon_a, \qquad \varepsilon_a \sim t_3 \text{ scaled to variance } \sigma^2,$$

so that rare large residuals occur. The closed-form τ^o remains the same as DGP 1.

B.3. Main envelope figure across DGPs

Before turning to repeated-sampling experiments, we plot a single-dataset "sensitivity envelope" for each DGP: the sharp lower and upper bounds as Λ_{out} varies, using a large trial ($n^r = 2000$) and target covariate sample ($n^o = 5000$). Figure 5 shows the expected monotone widening of the identified set as the analyst allows larger outcome shift. For bounded outcomes (DGP 3), widths are substantially smaller; for heavy tails (DGP 4), widths grow faster with Λ_{out}.

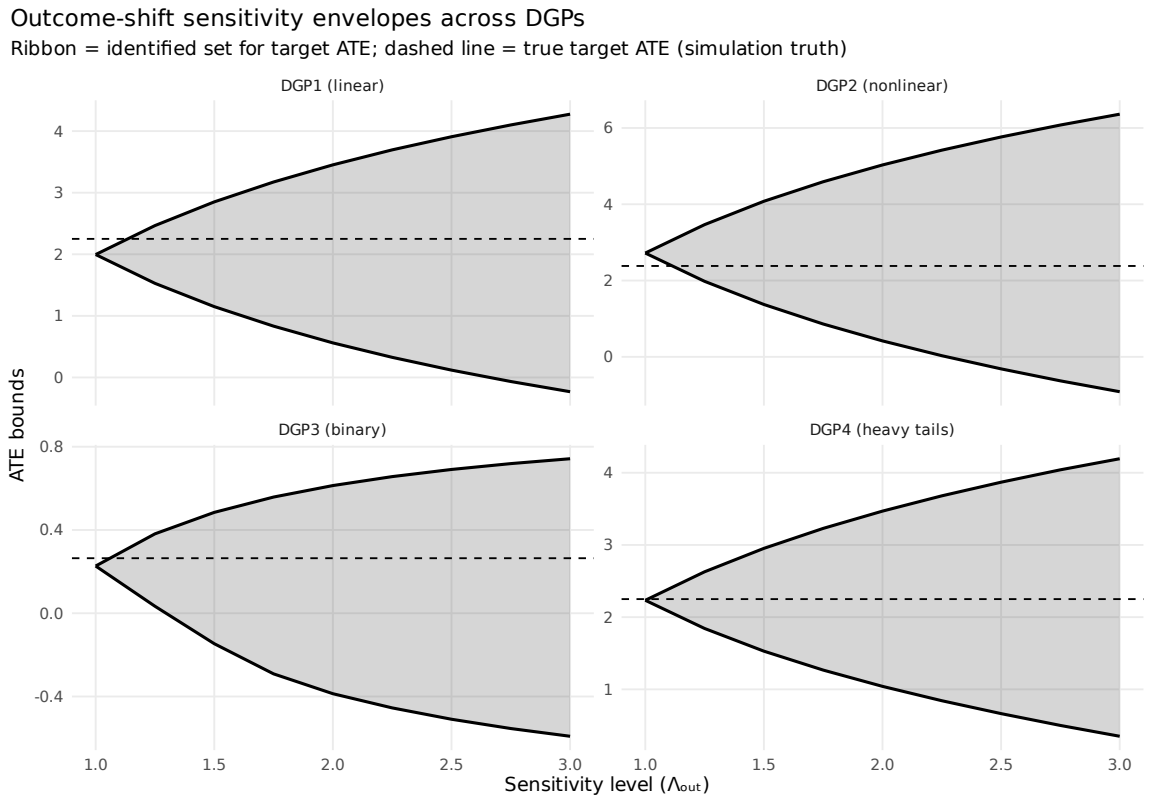

Figure 5: Sensitivity envelopes on a single large simulated dataset for DGPs 1–4. Each panel plots the sharp bound interval $[\hat{\tau}^-(\Lambda_{\text{out}}), \hat{\tau}^+(\Lambda_{\text{out}})]$ as a function of Λ_{out}, with the true target ATE indicated for reference. DGP 3 (binary outcomes) yields notably tighter envelopes due to bounded outcomes.

B.4. Experiment 1: validation and sharpness (DGP 1)

Design. We simulate DGP 1 with $(n^r, n^o) = (500, 1000)$ for $R = 1000$ replications. For each replication and each value on a grid $\Lambda_{\text{out}} \in [1, 3]$, we compute sharp bounds and record: coverage of $\tau^o = 2.25$, mean width, and a sharpness ratio relative to an "oracle width" computed on an extremely large simulated dataset ($n^r = 100{,}000$) using the same procedure.

Because Gaussian likelihood ratios are technically unbounded, we also compute a *tail-trimmed* implied $\Lambda_{\text{DGP}}(\alpha)$ by restricting to central $(1 - \alpha)$ outcome mass (here $\alpha = 0.01$) and taking the smallest Λ that upper-bounds the conditional likelihood ratio on that truncated region. This provides a calibration reference (conservative for functionals like the ATE).

Results. Figure 6 shows a clean transition from severe undercoverage near $\Lambda_{\text{out}} = 1$ (transportability assumed) to near-nominal coverage once Λ_{out} is moderately above 1. Coverage is already ≈ 0.98 at $\Lambda_{\text{out}} = 1.4$ and ≈ 0.99 at $\Lambda_{\text{out}} = 1.5$, reaching 1.00 by $\Lambda_{\text{out}} = 1.6$. Importantly, widths closely track oracle widths: the sharpness ratio is about 0.98 across the grid (Tab. 4), indicating the finite-sample procedure is nearly as tight as the population sharp interval. The tipping point distribution in Figure 6 (rightmost panel) concentrates around $\Lambda_{\min} \in [1.1, 1.3]$, the smallest sensitivity level needed for a given replicate to cover.

Table 4: Experiment 1 (DGP 1): selected operating points. Coverage is across $R = 1000$ replications; oracle widths use a large-n approximation.

Λ_{out}	Coverage	Mean width	Oracle width	Sharpness ratio
1.2	0.686	0.738	0.750	0.983
1.4	0.976	1.358	1.382	0.983
1.5	0.993	1.633	1.662	0.983
1.6	1.000	1.890	1.924	0.983
2.0	1.000	2.764	2.815	0.982
3.0	1.000	4.285	4.369	0.981

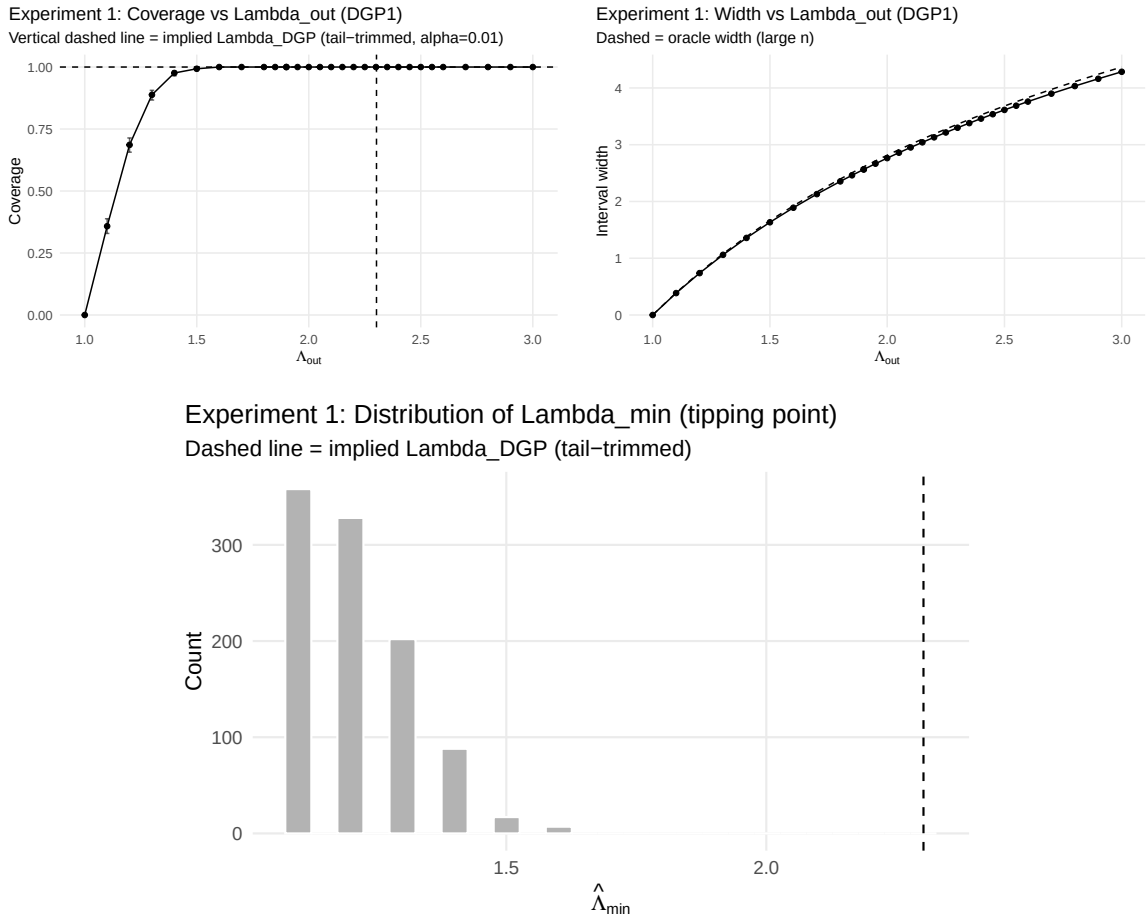

Figure 6: Experiment 1 (DGP 1): validation and sharpness. *Top-left:* coverage of the sharp interval for τ^o vs. Λ_{out} (with binomial uncertainty). *Top-right:* mean interval width vs. Λ_{out}, overlaid with an oracle large-n width. *Bottom:* distribution of $\Lambda_{\min}$, the smallest sensitivity level for which a replicate's interval covers τ^o.

B.5. Experiment 2: coverage–informativeness tradeoff vs. moderator shift (DGP 1)

Design. We vary the magnitude of population shift in the unmeasured moderator by setting $\gamma_o \in \{0.25, 0.5, 0.75, 1.0\}$ (with $\gamma_r = 0$), while keeping $(n^r, n^o) = (500, 1000)$ and $R = 1000$ replica-

Table 5: Experiment 2 (DGP 1): breakeven Λ_{out} (minimum achieving ≥ 0.95 coverage) as a function of moderator-shift strength γ_o. We also report the tail-trimmed $\Lambda_{\text{DGP}}(\alpha{=}0.01)$ computed from the conditional likelihood ratio, which is substantially more conservative for the ATE functional.

γ_o	Breakeven Λ_{out} (95% cov.)	Tail-trimmed $\Lambda_{\text{DGP}}(0.01)$
0.25	1.4	1.508
0.50	1.5	2.302
0.75	1.6	3.609
1.00	1.7	5.944

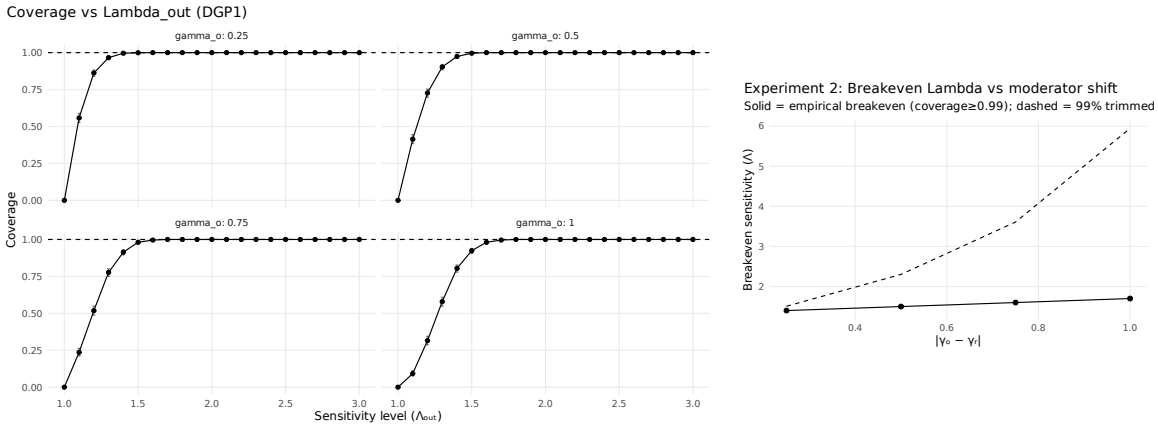

Figure 7: Experiment 2 (DGP 1): increasing moderator shift γ_o pushes the coverage-vs-Λ_{out} curve to the right. *Left:* coverage vs. Λ_{out} for $\gamma_o \in \{0.25, 0.5, 0.75, 1.0\}$. *Right:* breakeven Λ_{out} achieving 95% coverage.

tions. For each γ_o we sweep Λ_{out} and summarize the induced tradeoff between coverage and width. We also compute a *breakeven* sensitivity level, defined as the smallest Λ_{out} achieving at least 95% empirical coverage for that γ_o.

Results. As expected, increasing moderator shift makes transportability more fragile: the coverage curve shifts right, and the breakeven Λ_{out} rises nearly linearly with γ_o (Tab. 5, Figure 7). Concretely, the breakeven level increases from 1.4 at $\gamma_o = 0.25$ to 1.7 at $\gamma_o = 1.0$. This experiment provides an interpretable calibration: analysts can read off the sensitivity level needed to restore nominal coverage under a given degree of moderator shift.

B.6. Experiment 3: comparison to baseline procedures (DGP 1)

Design. We compare our sharp bounds to common alternatives at two sensitivity levels $\Lambda_{\text{out}} \in \{1.5, 2.0\}$ on DGP 1 with $(n^r, n^o) = (500, 1000)$. The baselines are: (i) naive transported point estimate (assumes $\Lambda_{\text{out}} = 1$); (ii) naive nonparametric bootstrap CI for the transported estimator; (iii) a very conservative worst-case bound; and (iv) a calibrated bias-function baseline. (For computational convenience in this experiment, the run uses $R = 50$ replications; the gaps are large enough to be visually stable, and increasing R is straightforward.)

Results. The naive transported estimator (and its bootstrap CI) can be too optimistic: its intervals are narrow but under-cover, while worst-case bounds cover trivially but are far too wide. Our bounds

Table 6: Experiment 3 (DGP 1): baseline comparison at $\Lambda_{\text{out}} \in \{1.5, 2.0\}$. Coverage is across $R = 50$ replications.

Method	Λ_{out}	Coverage	Mean width
Naive (point, Λ_{out}=1)	1.5	0.00	0.00
Bootstrap CI (naive)	1.5	0.70	0.804
Our bounds	1.5	0.98	1.616
Worst-case	1.5	1.00	14.576
Naive (point, Λ_{out}=1)	2.0	0.00	0.00
Bootstrap CI (naive)	2.0	0.70	0.804
Our bounds	2.0	1.00	2.734
Worst-case	2.0	1.00	14.576

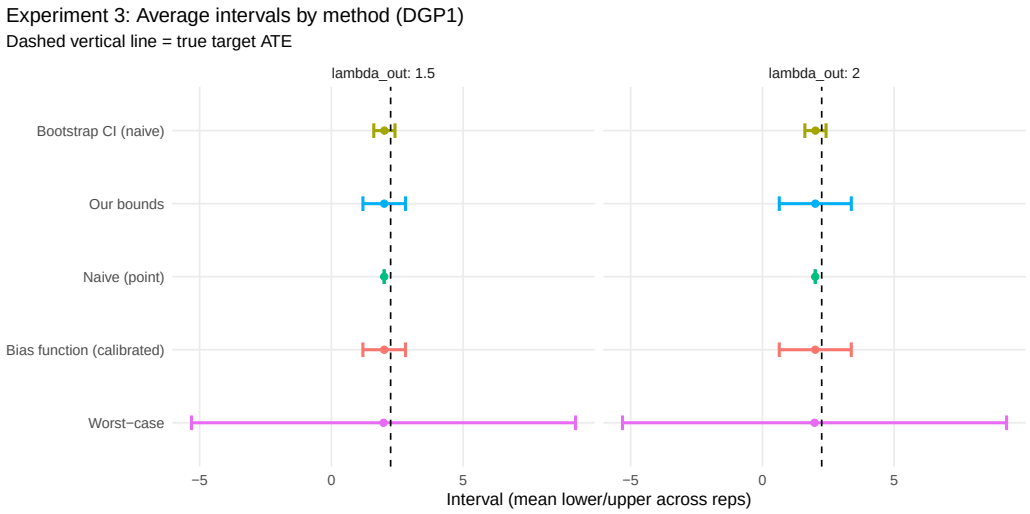

Figure 8: Experiment 3 (DGP 1): forest plot comparing interval procedures at $\Lambda_{\text{out}} \in \{1.5, 2.0\}$. Our bounds remain informative while maintaining high coverage; worst-case bounds are orders of magnitude wider.

achieve near-nominal coverage with substantially smaller width than worst-case, and they coincide with the calibrated bias-function baseline here (Tab. 6).

B.7. Experiment 4: scaling with trial sample size (DGP 1)

Design. Fix $\Lambda_{\text{out}} = 2.0$ and $(n^o = 1000)$, vary the trial size $n^r \in \{100, 200, 500, 1000, 2000, 5000\}$, and run $R = 300$ replications under DGP 1. We record coverage, mean width, sharpness ratio (relative to a large-n oracle width), and runtime.

Results. Figure 9 shows that width and sharpness stabilize quickly as n^r grows, while runtime remains negligible (milliseconds). Even at $n^r = 200$, coverage is already at 1.00 in this setting, and the sharpness ratio approaches 1 as finite-sample noise shrinks. This supports the practical usability of the greedy algorithm in algorithm 1.

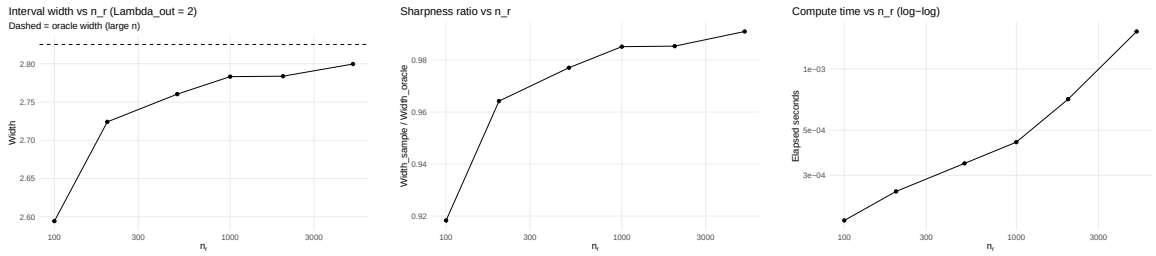

Figure 9: Experiment 4 (DGP 1): scaling with trial size n^r at fixed $\Lambda_{\text{out}} = 2$. *Left:* mean width vs. n^r. *Center:* sharpness ratio vs. n^r (approaches 1). *Right:* runtime vs. n^r (milliseconds).

Table 7: Experiment 5: coverage and width across DGPs.

DGP	Λ_{out}	Coverage	Mean width
DGP1 (linear)	1.5	0.991	1.635
DGP1 (linear)	2.0	1.000	2.769
DGP1 (linear)	3.0	1.000	4.296
DGP2 (nonlinear)	1.5	0.995	2.498
DGP2 (nonlinear)	2.0	1.000	4.260
DGP2 (nonlinear)	3.0	1.000	6.717
DGP3 (binary)	1.5	0.999	0.609
DGP3 (binary)	2.0	1.000	0.977
DGP3 (binary)	3.0	1.000	1.332
DGP4 (heavy tails)	1.5	0.989	1.497
DGP4 (heavy tails)	2.0	1.000	2.545
DGP4 (heavy tails)	3.0	1.000	3.991

B.8. Experiment 5: robustness across outcome types (DGPs 1–4)

Design. We run the same Monte Carlo experiment across DGPs 1–4 with $(n^r, n^o) = (500, 1000)$, $R = 1000$ replications, and $\Lambda_{\text{out}} \in \{1.5, 2.0, 3.0\}$.

Results. Figure 10 and Tab. 7 show that the bounds maintain high coverage across nonlinear, binary, and heavy-tailed settings. Binary outcomes (DGP 3) yield the tightest intervals, while nonlinear and heavy-tailed outcomes require wider intervals for the same Λ_{out}, as expected.

B.9. Experiment 6: sensitivity to weight estimation (DGP 1)

Design. We compare three weighting strategies in DGP 1 with $(n^r, n^o) = (500, 1000)$ and $R = 150$ replications: (i) oracle weights using the known Gaussian density ratio; (ii) estimated logistic membership weights using the correct covariates; and (iii) a misspecified logistic membership model using only X_1. We report coverage and width at $\Lambda_{\text{out}} \in \{1.5, 2.0\}$ and visualize the induced weight distributions.

Results. Figure 11 shows that (in this moderate covariate-shift setting) bounds are stable across weighting strategies: estimated weights closely match oracle performance, and even mild misspec-

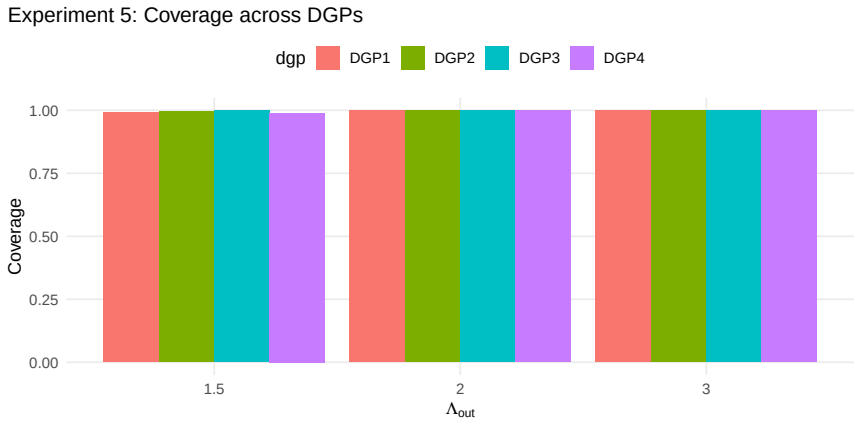

Figure 10: Experiment 5: robustness across DGPs 1–4. Coverage at $\Lambda_{\mathrm{out}} = 1.5$ is already near nominal in all cases and reaches 1.00 by $\Lambda_{\mathrm{out}} = 2.0$ for all DGPs.

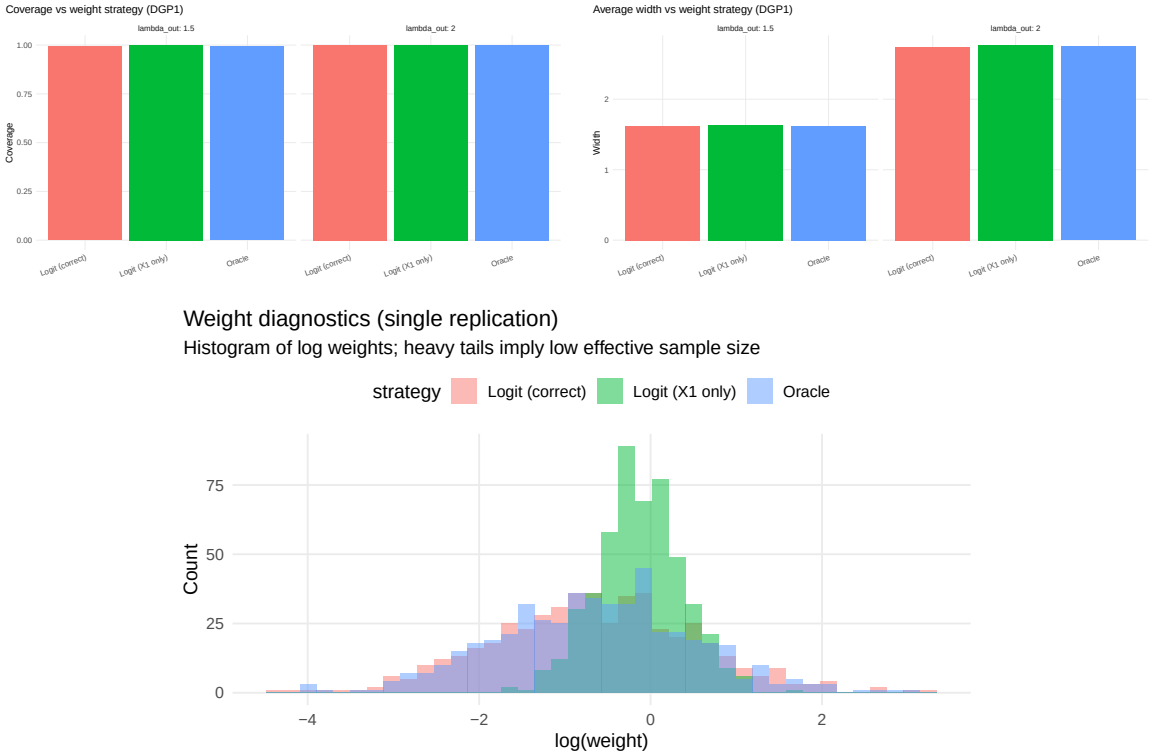

Figure 11: Experiment 6 (DGP 1): sensitivity to weight estimation. *Top-left:* coverage by weighting strategy. *Top-right:* width by weighting strategy. *Bottom:* weight distribution diagnostic (log-scale), illustrating tail behavior across strategies.

ification does not drastically change width or coverage at these Λ_{out} values. The weight histogram highlights how misspecification changes tail behavior (maximum weight and effective sample size), which is useful for diagnostics in applications.

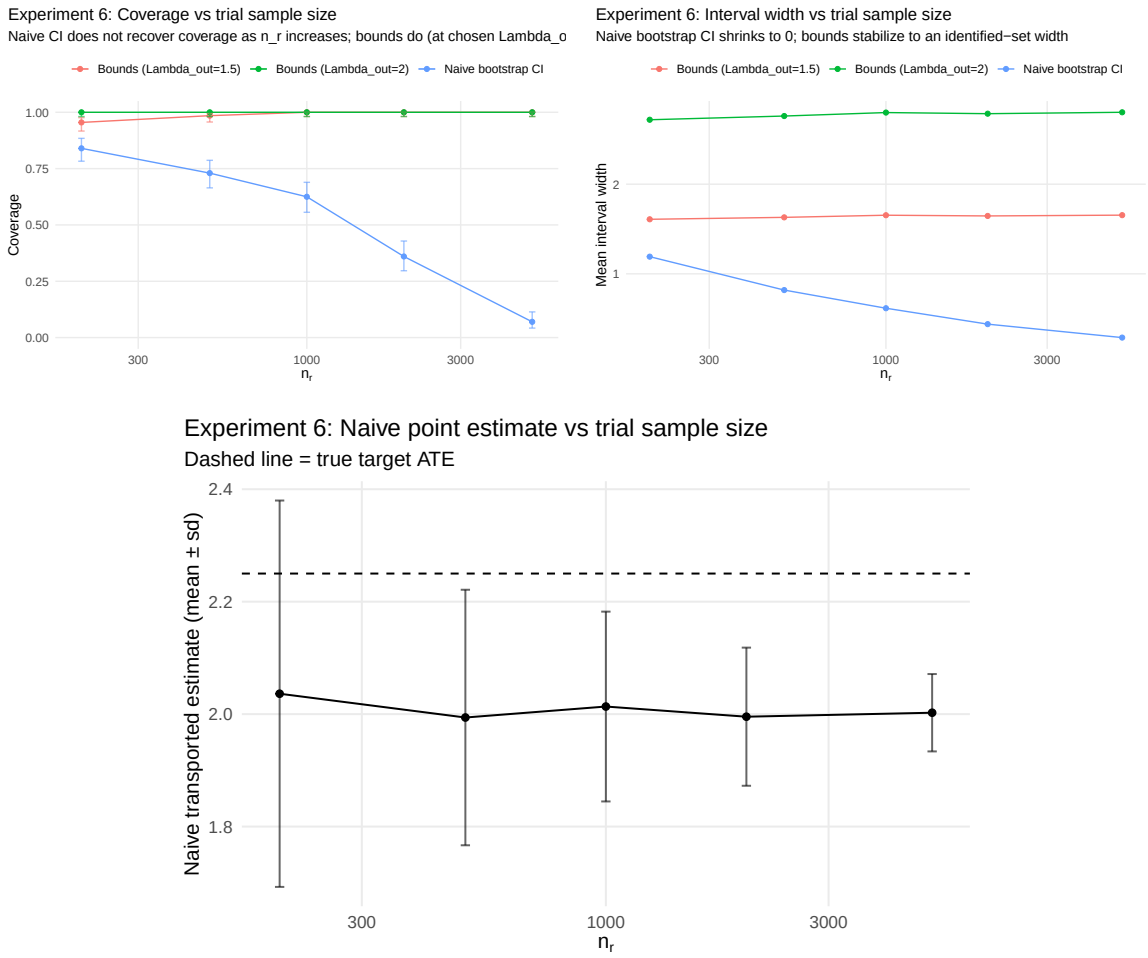

Figure 12: Experiment 6b (DGP 1): identification vs. estimation. As n^r grows, the naive bootstrap CI becomes narrower yet less valid, while sharp bounds retain coverage and stabilize in width. The bottom panel (optional) illustrates convergence of the naive point estimate away from the true τ^o.

B.10. Experiment 6b: identification vs. estimation scaling (DGP 1)

Design. This experiment isolates an important failure mode of naive inference: even with very large n^r, the transported point estimator can be *precisely wrong* when outcome transportability fails. We simulate DGP 1 with a large target covariate sample $n^o = 5000$, vary trial size $n^r \in \{200, 500, 1000, 2000, 5000\}$, and run $R = 200$ replications per n^r. We compare: (i) the naive transported point estimate (equivalently $\Lambda_{\text{out}} = 1$), (ii) a naive nonparametric bootstrap CI for the transported estimator (200 bootstrap resamples), and (iii) our sharp bounds at $\Lambda_{\text{out}} \in \{1.5, 2.0\}$.

Results. Figure 12 shows a stark pattern: as n^r increases, the naive bootstrap CI shrinks rapidly but coverage *worsens* (e.g., dropping to ≈ 0.07 by $n^r = 5000$), demonstrating that the dominant limitation is *identification*, not sample size. In contrast, the sharp bounds maintain near-nominal coverage with widths that do not collapse to zero (as expected under partial identification).

B.11. Experiment 7: bounded-support DGP where the LR model holds literally

A common technical concern is that for Gaussian outcomes the pointwise likelihood ratio can be unbounded. To demonstrate that our method behaves as intended when the model holds literally, we construct a bounded-support DGP (DGP 7) with truncated covariates and truncated outcomes so that the conditional likelihood ratio is finite.

DGP 7 definition (bounded). We draw covariates from truncated normals:

$$X^r \sim \text{TruncNormal}(0, I_p;\ [-B, B]^p), \qquad X^o \sim \text{TruncNormal}(\delta \mathbf{1}_p, I_p;\ [-B, B]^p),$$

with $B = 3$ and $\delta = 0.5$. The moderator is generated as before with $\gamma_r = 0$ and $\gamma_o = 0.2$: $U^s \mid X \sim \mathcal{N}(\gamma_s X_1, 1)$. Outcomes follow a truncated normal:

$$Y(a) \sim \text{TruncNormal}\Big(\beta_0 + \beta_x^\top X + a(\tau_0 + \beta_u U),\ \sigma^2;\ [y_L, y_U]\Big),$$

with $\beta_0 = 1$, $\beta_x = (0.5, 0.3, 0.2, 0.1, 0.1)^\top$, $\tau_0 = 1$, $\beta_u = 0.25$, $\sigma = 1$, and $[y_L, y_U] = [-3, 3]$. For this DGP we also compute a literal global Λ_{DGP} by maximizing the conditional likelihood ratio over (a, x, y) on the bounded support (reported in the corresponding table file).

Design and results. We run an Exp-1-style sweep with $(n^r, n^o) = (500, 1000)$ and $R = 200$ replications over a grid of Λ_{out}. Figure 13 shows the same qualitative pattern as in DGP 1, but here the LR-bounded model is well-defined without trimming: coverage increases monotonically with Λ_{out}, and widths grow smoothly. The tipping points concentrate near $\Lambda_{\min} \approx 1.1$ in this particular setting.

B.12. Experiment 8: "bang–bang" tilting visualization

Finally, we visualize the structure predicted by theorem 3: the extremal solutions that attain the sharp upper/lower bounds correspond to multipliers that largely saturate at the LR constraints, producing a near-threshold ("bang–bang") pattern when ordered by outcome (or the relevant sufficient statistic).

B.13. Summary discussion (what the simulations show)

Across DGPs and experimental designs, the simulations support four main conclusions. First, the proposed procedure achieves empirical coverage once Λ_{out} is large enough to plausibly contain the true outcome shift, and it transitions smoothly from the transported point estimate at $\Lambda_{\text{out}} = 1$ to wider partial-identification intervals as Λ_{out} increases (Exp. 1, Fig. 6). Second, the intervals are *sharp* in the sense that their widths closely match large-n oracle widths, implying little avoidable conservatism in finite samples (Exp. 1, Tab. 4). Third, the method exposes an explicit coverage–informativeness tradeoff and yields interpretable "breakeven" sensitivity levels under varying degrees of moderator shift (Exp. 2, Tab. 5). Fourth, the identification aspect is central: naive transported estimators can become arbitrarily precise yet invalid as n^r grows, whereas sharp bounds maintain coverage and do not collapse (Exp. 6b, Fig. 12). Robustness experiments (Exp. 5) and diagnostics around weight estimation (Exp. 6) further support practical applicability, while Exp. 7 and Exp. 8 address common technical and interpretability questions about LR-bounded models and the structure of the optimizing tilts.

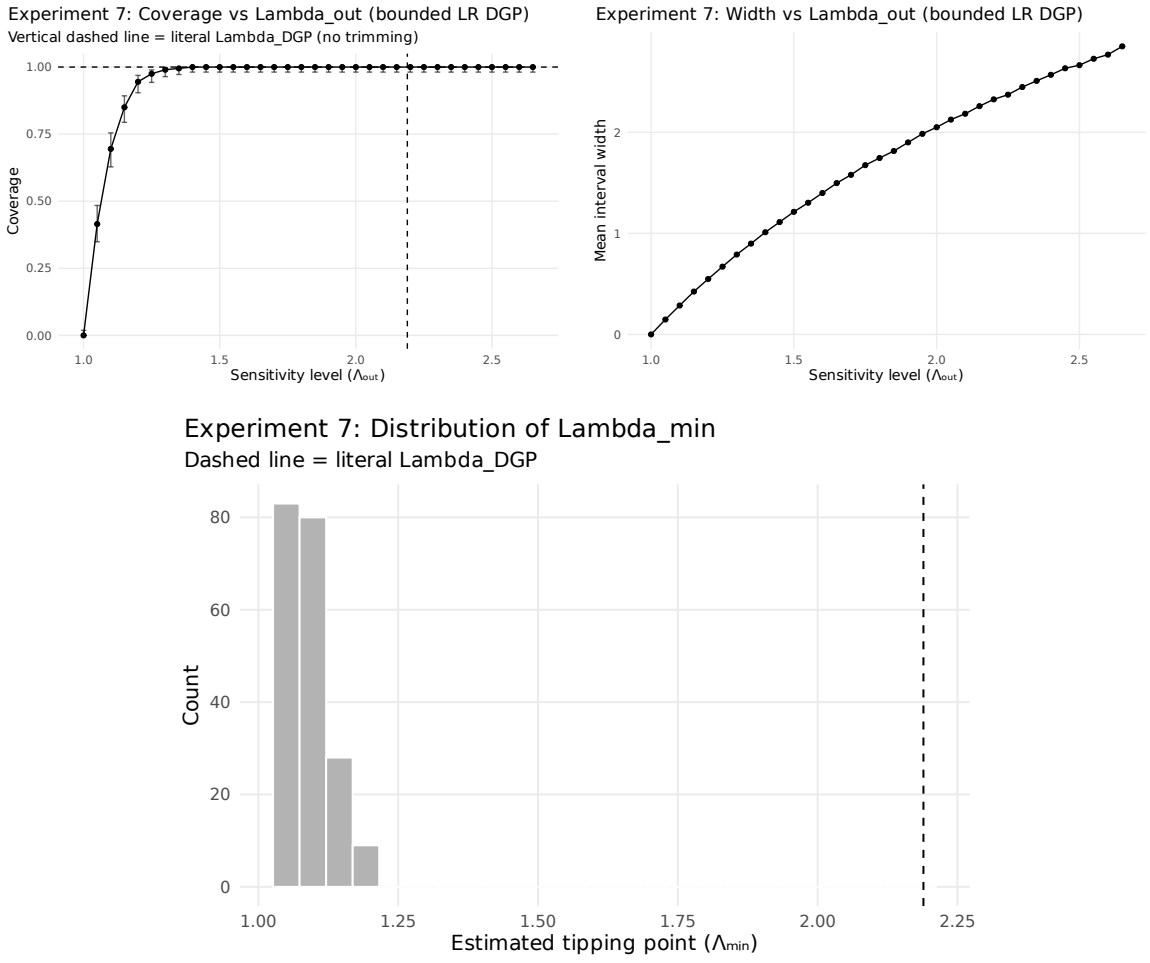

Figure 13: Experiment 7 (bounded-support DGP): when outcomes and covariates have bounded support, the pointwise LR model holds literally. Coverage increases with Λ_{out} and the tipping-point distribution is concentrated.

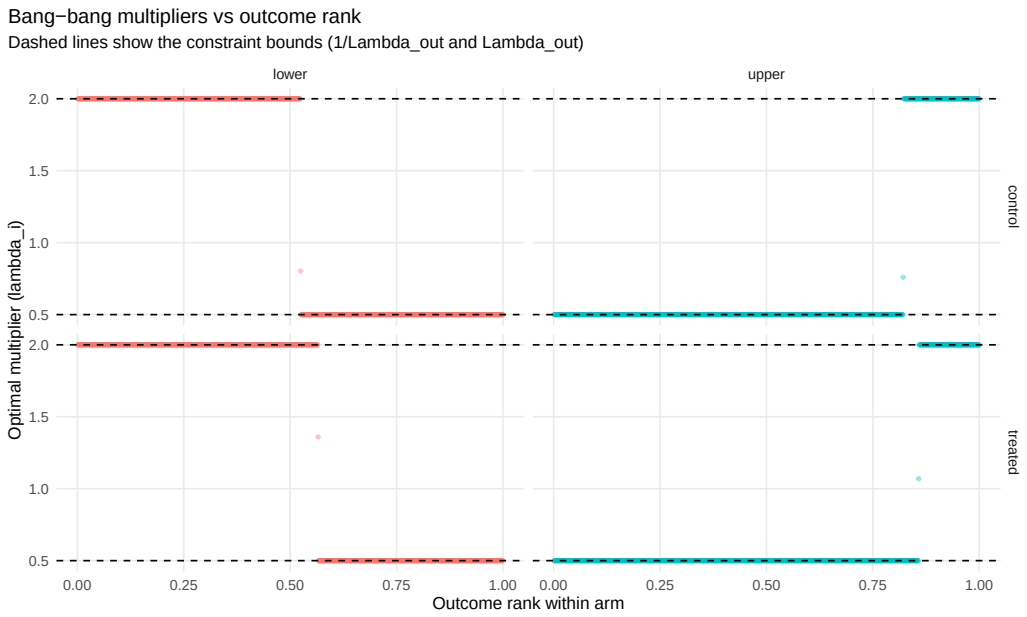

Figure 14: Experiment 8: "bang–bang" structure of the bound-attaining tilts. The estimated multipliers (density ratio adjustments) concentrate near the constraint values and switch sharply around a threshold, consistent with theorem 3.