Leave-Multiple-Out Informal Benchmarking

Understanding the Behavior of Informal Benchmarking for Multivariate Confounding

Nayden Borodjiev[1]

Supervisor(s): Jesse Krijthe[1], **Matej Havelka**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Nayden Borodjiev
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Matej Havelka, Avishek Anand

Abstract

Informal benchmarking is a popular approach for calibrating sensitivity bounds for hidden confounding by treating observed covariates as if they were unobserved. While leave-one-out (LOO) benchmarking removes a single covariate, leave-multiple-out (LMO) benchmarking removes sets of covariates to approximate multidimensional confounding. In this study, we examine whether LMO benchmarking recovers the confounding strength as the number of features dropped increases. Using a synthetic dataset with bounded covariates and known confounding structure, we compare empirical bounds with an Oracle-like benchmark and the true theoretical value. The theoretical bound increases monotonically as more covariates are omitted, but the empirical LMO bound does not follow this pattern - it plateaus and then declines. The experiments show that this behavior is not explained by estimation error alone. Rather, it is a consequence of informal benchmarking being restricted by the given sample: large bounds are obtained from individuals with certain covariate values. This issue becomes more important as larger subsets are omitted, because the strongest theoretical benchmarks depend on increasingly specific patterns in the omitted covariates. As a result, LMO benchmarking may be more reliable for small omitted subsets, but should be interpreted with increasing caution for larger ones. We conclude that LMO informal benchmarking results should be read as sample-realized benchmarks rather than as the maximum confounding strength possible over the full covariate space.

1 Introduction

Causal machine learning has emerged as an important topic for researchers who seek to go beyond simple statistical correlations and determine cause-and-effect relationships. The standard method to assess such causal effects has been Randomised Controlled Trials (RCTs), which ensure that the allocation of treatment is independent of patient characteristics (Braga et al., 2025). However, RCTs are often too expensive, time-consuming, or ethically problematic to carry out (Zabor et al., 2020). This leads to a reliance on observational data in many domains, where researchers use many diverse sources such as electronic health records or economic indicators to estimate causal effects.

The validity of causal conclusions made from such observational data rests on the ignorability (or unconfoundedness) assumption (Baitairian et al., 2026). It demands that all variables that affect both treatment and outcome are observed and included in the model. This is rarely realistic in practice and we expect the existence of hidden confounders. Factors such as socioeconomic status, dietary habits, or environmental exposure are rarely measured. Their absence from the model can severely bias estimated treatment effects, potentially leading researchers to identify false causal relationships.

To address the inevitable presence of unobserved variables, researchers employ sensitivity analysis (Rosenbaum, 2005). It relaxes the unconfoundedness assumption to evaluate how sensitive a causal conclusion is to the existence of an unobserved confounder. One of the more popular techniques is the Marginal Sensitivity Model (MSM), proposed by Tan (2006), that fundamentally relies on the sensitivity parameter Γ. This parameter is the maximum factor by which units similar on observed covariates would differ in their odds of treatment if an unobserved confounder was measured. Researchers can move from a point estimate to partially identified sets by varying Γ, which correspond to the range of treatment effects compatible with a given level of hidden bias.

Although Γ is useful in principle, it is difficult to interpret in practice. Often, it is not obvious how to decide if a particular value is a "large" or "small" degree of hidden bias. In addition, Γ is an unobservable metric, its value is tied to the covariates that we have not measured.

To fill this interpretative gap, the field has turned to informal benchmarking (IB), a calibration strategy, as described by Baitairian et al. (2026), that uses observed data to ground the choice of Γ. The logic of IB is to treat known covariates as if they were unobserved, to see how much they would change the estimated treatment effect. Within this framework, two primary procedures exist: the leave-one-out approach that investigates the impact of removing one feature at a time, and the leave-multiple-out approach, the multi-covariate form of informal benchmarking that removes sets of features. Both are conceptually similar but aim to answer different questions and possibly behave differently under certain conditions.

To our knowledge, there is a gap in the literature regarding the reliability of this multi-covariate benchmarking procedure. While LOO has been widely used and studied, LMO is less understood. In this paper we explore the performance and the limits of the LMO procedure in order to assess its reliability as a validation technique for causal machine learning. The research questions addressed are:

- How does the empirical LMO informal benchmark behave as the number of features dropped increases?
- How does the structure of the data affect the reliability of LMO informal benchmarking as the number of features dropped increases?

By investigating these questions through experimental simulations, we aim to provide practitioners with a clearer understanding of when LMO benchmarking provides a reliable estimate of confounding strength and when it may be prone to statistical artifacts.

The rest of this paper is organised as follows. Section 2 introduces the causal-inference background, including propensity scores, unmeasured confounding and the MSM. Section 3 describes the LMO informal benchmarking framework, simulation design, and evaluation metrics used in the experiments. Section 4 presents the experimental results, Section 5 discusses the main findings and their implications, Section 6 reflects on responsible research considerations, and Section 7 concludes the research with a summary and key takeaways.

2 Background

This section establishes the theoretical foundations for addressing unobserved confounding. We first formalize causal effects and standard assumptions. Then we review propensity score estimation and finally we introduce the Marginal Sensitivity Model to quantify the robustness of causal estimates against hidden bias.

2.1 The Neyman-Rubin Potential Outcome Framework

We use the Neyman-Rubin potential outcome framework (Splawa-Neyman et al., 1990; Rubin, 1974) to formally describe causal effects in observational situations. For every unit/patient i, we have a binary treatment $T_i \in \{0, 1\}$ and two possible outcomes: $Y_i(1)$, which represents the result if the unit was treated, and $Y_i(0)$, which represents the result if the unit was not treated. This is formalised as:

$$Y = TY(1) + (1 - T)Y(0)$$

where Y is the observed outcome. Because the individual treatment effect $(Y_i(1) - Y_i(0))$ is unobservable, researchers focus on population-level estimands, such as the Average Treatment Effect (ATE):

$$\text{ATE} = E[Y(1) - Y(0)]$$

which represents the expected difference in outcomes if the entire population was treated versus if none were treated.

2.2 Assumptions under Unconfoundedness

Because the ATE takes into account unobserved counterfactuals, the estimated effects based on observational data require specific assumptions (Baitairian et al., 2026). The X-ignorability assumption asserts that treatment assignment is independent of possible outcomes after X is taken into account:

$$(Y(0), Y(1)) \perp T \mid X$$

This isolates the causal relationship between treatment and potential outcomes. Additionally, we assume positivity (or overlap), requiring that every unit has a non-zero probability of receiving either treatment state ($0 < P(T = 1 \mid X) < 1$). Under these conditions, we can calculate the propensity score, $e(X) = P(T = 1 \mid X)$, which represents the conditional probability of treatment given the observed vector (Rosenbaum and Rubin, 1983). One possible choice for estimating the mean potential outcomes that form the ATE is Inverse Probability Weighting (IPW) (Hirano et al., 2003). For example, the mean outcome under treatment can be estimated as:

$$\hat{\theta}(1) = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)}$$

with an analogous estimate for $\hat{\theta}(0)$ and the IPW estimate of the ATE is then $\hat{\theta}(1) - \hat{\theta}(0)$.

2.3 Estimating Propensity Scores

In practical applications of causal inference, the true propensity score $e(X)$ is rarely known and must be inferred from the observed data to construct the IPW estimator. This typically involves training a predictive model to estimate the probability of treatment assignment based on the observed covariates X.

One standard and widely used approach is logistic regression (Austin, 2011). Under this model, the probability of treatment is expressed as a linear function of the observed covariates mapped through the logistic function:

$$\hat{e}(X) = \frac{1}{1 + \exp(-X^\top \hat{\beta})}$$

Logistic regression is favored for its low computational complexity, interpretability, and well-understood properties. While highly flexible machine learning algorithms are increasingly popular for capturing complex relationships (Yao et al., 2021), logistic regression serves as the foundational baseline and, as shown by Baitairian et al. (2026), it outperforms other methods in simpler scenarios.

2.4 The Threat of Unmeasured Confounding

In practice, X-ignorability is often unrealistic. If there are unmeasured confounders U that influence both T and Y, the ignorability assumption fails, and standard IPW estimates will be biased. To account for this, we relax the assumption to (X, U)-ignorability, acknowledging that treatment is only independent of potential outcomes when both observed and unobserved factors are considered:

$$(Y(0), Y(1)) \perp T \mid (X, U)$$

Because U is by definition unobserved, it is impossible to obtain a single point estimate for the treatment effect. Instead, researchers use sensitivity analysis to evaluate how robust their conclusions are to potential hidden bias.

2.5 The Marginal Sensitivity Model

The MSM (Tan, 2006) provides a flexible framework for sensitivity analysis by bounding the influence of U on the treatment assignment mechanism. It assumes that two individuals who share identical observed covariates X may still differ in their odds of receiving treatment by at most a factor of Γ due to the unmeasured confounder U. Formally, the MSM is defined by the set of possible propensity scores $e(x, u)$ such that:

$$\Gamma^{-1} \leq \text{OR}(e(x, u), e(x)) \leq \Gamma$$

where the odds ratio function is defined as:

$$\text{OR}(a, b) = \frac{a/(1 - a)}{b/(1 - b)}$$

When $\Gamma = 1$, the model reduces to standard unconfoundedness; as Γ increases, the model allows for progressively stronger hidden bias.

One of the most important metrics is the critical value Γ_c, defined as the minimum confounding strength required for the sensitivity bounds (or their confidence intervals) to include the null effect. If Γ_c is high, the causal findings are

considered robust because it would take a very strong unobserved confounder to nullify the effect. If Γ_c is low, the study is sensitive to hidden bias, and the causal conclusions are less reliable.

3 Methodology

This section outlines the experimental methodology used to assess the limits of informal benchmarking. We begin by defining the extended Leave-Multiple-Out Informal Benchmarking Framework. Next, we introduce a synthetic Data Generating Process utilizing three distinct covariate distributions to isolate the impact of data geometry. Finally, we establish the theoretical ground-truth benchmark and the metrics used to validate our findings.

3.1 The Leave-Multiple-Out Informal Benchmarking Framework

Informal benchmarking addresses the problematic interpretability of unobserved confounding Γ by calculating an empirical bound $\hat{\Gamma}$ based on the observed covariates (Baitairian et al., 2026). The core idea is to treat known covariates as if they were unobserved, to see how much they would move the estimated treatment effect.

Traditionally, the sensitivity of a treatment effect is calculated via a leave-one-out procedure, where a single covariate is dropped to measure the resulting shift in the odds of treatment. However, this limits the benchmarking to a single dimension of confounding, which may not capture the true multidimensional nature of hidden bias. To address this, we use a generalized LMO framework that systematically evaluates the confounding strength of all possible subsets of observed covariates, providing a more comprehensive sensitivity analysis.

Let $X \in \mathbb{R}^{N \times p}$ represent the fully observed covariate matrix. We define $S \subset \{1, 2, \ldots, p\}$ as a specific subset of covariates chosen to be omitted, where $|S| = m$ is the size of the dropped subset. This in turn makes LOO the base case where $m = 1$.

For any given individual i, the baseline full-information odds of receiving treatment $T_i = 1$ are conditioned on their complete covariate profile X_i:

$$O_{\text{full}}(X_i) = \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} \tag{1}$$

To benchmark against the specific feature subset S, those features are omitted from the dataset, yielding the reduced covariate profile $X_i^{(-S)}$. The reduced-information odds are conditioned strictly on the remaining observed data:

$$O^{(-S)}_{\text{reduced}}(X_i) = \frac{\hat{e}(X_i^{(-S)})}{1 - \hat{e}(X_i^{(-S)})} \tag{2}$$

The implied confounding strength, $\hat{\Gamma}_i^{(S)}$, represents the magnitude of the shift in treatment odds for patient i when the covariates in subset S are treated as unobserved:

$$\hat{\Gamma}_i^{(S)} = \max\left(\frac{O_{\text{full}}(X_i)}{O^{(-S)}_{\text{reduced}}(X_i)}, \frac{O^{(-S)}_{\text{reduced}}(X_i)}{O_{\text{full}}(X_i)}\right) \tag{3}$$

Because the Marginal Sensitivity Model requires bounds to hold for all units simultaneously, the empirical benchmark for the subset S is strictly defined as the maximum of these individual deviations across the entire sample of N individuals:

$$\hat{\Gamma}^{(S)} = \max_{i \in \{1 \ldots N\}} \hat{\Gamma}_i^{(S)} \tag{4}$$

To identify the absolute maximum vulnerability of the study to an unobserved confounder of size m, this maximum must be evaluated across all $\binom{p}{m}$ possible combinations of dropped features.

Algorithm 1 provides a computational implementation of the LMO procedure, which evaluates the confounding strength of all possible subsets of size m and returns the $\hat{\Gamma}^{(S)}$ for each subset S.

3.2 Simulation Design and Data Generating Process

To explore the behavior of the LMO procedure, we designed a simulation with a strictly defined Data Generating Process (DGP). In real observational data, the true confounding mechanism is fundamentally unknowable. However, a synthetic DGP allows us to fix the exact weights $\mathbf{w}$ that influence treatment assignment, providing a theoretical sensitivity limit against which empirical estimates can be benchmarked.

Across the reported experiments, the number of observed covariates is fixed at $p = 10$ with structural treatment weights,

$$\mathbf{w} = (1.0, 0.9, 0.8, \ldots, 0.2, 0.1).$$

Only the absolute values of the weights matter as a negative weight would simply flip the direction of effect on the treatment assignment, without changing the maximum possible magnitude of the log-odds shift.

All covariates are chosen to lie in $[-1, 1]^p$, so the sensitivity bound has a strict analytical ceiling. For an omitted subset S, the dropped structural log-odds contribution is $\sum_{j \in S} X_{ij} w_j$. Because the covariates are bounded, the largest possible shift occurs at the signed corners of the covariate space, where the omitted covariates are close to either 1 or -1 in the relevant coordinates. Therefore, for subset size m, the theoretical maximum is

$$\Gamma^{(m)}_{\text{theory}} = \exp\left(\sum_{k=1}^{m} |w|_{(k)}\right). \tag{5}$$

where $|w|_{(k)}$ denotes the k-th largest absolute structural weight. A more detailed derivation of this ceiling is provided in Appendix A.

The theoretical upper bound is a conservative reference point: it tells us the largest confounding strength possible under the bounded DGP. It is the largest treatment-odds shift allowed anywhere, achievable only for certain extreme configurations of the covariates. This motivates the focus on the geometry of the observed sample: even when the DGP allows such extreme points, the theoretical ceiling is only realized empirically if there are individuals in the sample whose omitted covariates align near these corners.

Algorithm 1 Computational Implementation of LMO Informal Benchmarking

Require: Covariates $X \in \mathbb{R}^{N \times p}$, Treatment $T \in \{0,1\}^N$, Subset size m, Estimator $\hat{E}$

$\hat{e}_{\text{full}} \leftarrow$ Fit and predict propensity scores using $\hat{E}(X, T)$
$\mathbf{O}_{\text{full}} \leftarrow \hat{e}_{\text{full}}/(1 - \hat{e}_{\text{full}})$ ▷ Vector of full-information odds

$\mathcal{C} \leftarrow$ Set of all $\binom{p}{m}$ possible feature combinations
for each subset $S \in \mathcal{C}$ **do**
    $X_{(-S)} \leftarrow X$ with features in S removed
    $\hat{e}_{\text{red}} \leftarrow$ Fit and predict reduced propensity scores using $\hat{E}(X_{(-S)}, T)$
    $\mathbf{O}_{\text{red}} \leftarrow \hat{e}_{\text{red}}/(1 - \hat{e}_{\text{red}})$

    $\mathbf{OR} \leftarrow \mathbf{O}_{\text{full}}/\mathbf{O}_{\text{red}}$ ▷ Vector of odds ratios for all N individuals
    $\mathbf{\Gamma}_{\text{dev}} \leftarrow \max(\mathbf{OR}, 1/\mathbf{OR})$ ▷ Element-wise maximum deviation
    $\hat{\Gamma}^{(S)} \leftarrow \max(\mathbf{\Gamma}_{\text{dev}})$ ▷ Maximum bound for subset S
    Store $(S, \hat{\Gamma}^{(S)})$
end for

return Stored subsets sorted by $\hat{\Gamma}^{(S)}$ in descending order

Covariate Distributions To isolate how often these boundary regions are represented in the sample, the covariate matrix $X \in \mathbb{R}^{N \times p}$ was generated from three independent marginal distributions. Each feature X_{ij} was drawn independently based on one of the following distributions:

1. **Uniform (Flat Density):** $X_{ij} \sim \mathcal{U}(-1, 1)$. Probability mass is distributed evenly across the domain. This serves as the baseline setting; the full hypercube is possible, but finite samples may still contain few observations near the aligned corners.
2. **Beta U-Shaped (Boundary Mass):** To increase the probability of observing boundary-aligned individuals, latent variables were drawn from a Beta distribution with shape parameters $\alpha, \beta < 1$, specifically $\text{Beta}(0.2, 0.2)$, and linearly scaled:
$$X_{ij} = 2 \cdot B_{ij} - 1, \quad B_{ij} \sim \text{Beta}(0.2, 0.2) \tag{6}$$
3. **Beta Bell-Shaped (Centered Mass):** To create the opposite geometry, with most observations concentrated near zero and few near the extremes, latent variables were drawn using shape parameters $\alpha, \beta > 1$, specifically $\text{Beta}(5.0, 5.0)$, scaled identically:
$$X_{ij} = 2 \cdot B_{ij} - 1, \quad B_{ij} \sim \text{Beta}(5.0, 5.0) \tag{7}$$

Treatment Assignment Mechanism Following the generation of the covariate matrix, treatment assignment was governed by the fixed structural weight vector defined above. The true propensity score $e(X_i)$ for each simulated patient was calculated using the logistic cumulative distribution function:
$$e(X_i) = \frac{1}{1 + \exp(-X_i^\top \mathbf{w})} \tag{8}$$
Finally, the observed binary treatment vector $T \in \{0,1\}^N$ was generated by drawing from a Bernoulli distribution parameterized by these exact structural probabilities:
$$T_i \sim \text{Bernoulli}(e(X_i)) \tag{9}$$

Thus, the conditional treatment assignment rule $e(x)$ is fixed across experiments. When the marginal distribution of X is changed, observations are placed in different regions of the same bounded covariate space, which may also change the realized distribution of $e(X)$. This variation is therefore interpreted as a change in covariate geometry, not as a different treatment mechanism.

Changing p would also change the dimension of the hypercube, the number of possible omitted subsets, and the probability of observing near-corner points. We want to isolate the effect of covariate geometry, so we keep $p = 10$ constant and only change the marginal distribution of the covariates.

3.3 Evaluation Metrics and Analytical Benchmarks

To evaluate the behavior of the LMO procedure, we compare the empirical bound with two reference values: the theoretical ceiling in Equation 5 and an Oracle structural benchmark. The theoretical ceiling maximizes over the entire bounded covariate domain and asks what confounding strength is possible in principle. The Oracle benchmark instead uses the known DGP to ask a different question: if the treatment-assignment model is known exactly, how much of the log-odds shift would be visible in the observed sample?

This distinction matters because the Oracle is not a maximum over the full covariate space. It is "oracle" only with respect to the treatment-assignment rule. In the implementation, the empirical LMO bound is obtained by fitting a full propensity model, refitting a reduced propensity model after dropping each subset S, and measuring the largest absolute difference in their fitted odds. The Oracle removes the estimation step: it does not use the sampled treatment labels or refit any reduced model, but directly computes the known dropped structural contribution for each observed individual. Thus, the Oracle separates propensity-model estimation error from sample-geometry limitations. If the Oracle benchmark

is still far below the theoretical value, then the gap comes from the geometry of the observed sample, not from estimation error.

For a given omitted subset S, the true dropped log-odds contribution for individual i is

$$\sum_{j \in S} X_{ij} w_j.$$

The corresponding Oracle subset benchmark is

$$\Gamma^{(S)}_{\text{Oracle}} = \max_i \exp\left(\left|\sum_{j \in S} X_{ij} w_j\right|\right), \tag{10}$$

and the Oracle benchmark for subset size m is

$$\Gamma^{(m)}_{\text{Oracle}} = \max_{S:|S|=m} \Gamma^{(S)}_{\text{Oracle}}. \tag{11}$$

Using the theoretical ceiling as the target, we report recovery ratios to measure how much of it is realized in the observed data:

$$\text{Recovery}^{(m)}_{\text{emp}} = \frac{\hat{\Gamma}^{(m)}}{\Gamma^{(m)}_{\text{theory}}}, \qquad \text{Recovery}^{(m)}_{\text{Oracle}} = \frac{\Gamma^{(m)}_{\text{Oracle}}}{\Gamma^{(m)}_{\text{theory}}}. \tag{12}$$

Values below one indicate under-recovery of the theoretical maximum. Values above one can occur when the empirical procedure misspecifies the propensity model, leading to an overestimation of the confounding strength. We also summarize the location and size of the observed peak:

$$m^* = \arg\max_m \hat{\Gamma}^{(m)}, \qquad \text{Drop} = 1 - \frac{\hat{\Gamma}^{(p)}}{\hat{\Gamma}^{(m^*)}}. \tag{13}$$

These quantities help us understand how the empirical bound behaves as more features are dropped, and how much it declines after reaching its maximum, which is a key aspect of the observed behavior that we seek to explain.

4 Experimental Results

This section reports four experiments on the behavior of LMO informal benchmarking. We first compare empirical fitted bounds and Oracle structural benchmarks against the theoretical ceiling, then examine the individual trajectories behind the sample-level pattern. We then test the proposed mechanism by injecting corner observations and, finally, vary the covariate distribution to study the role of sample geometry.

4.1 Experiment 1: Empirical LMO Behavior as Omitted Subset Size Increases

Experiment 1 investigates how the LMO benchmark evolves as more covariates are dropped. We expect the theoretical Γ to give a monotone curve: the confounding effect grows as more information is hidden. We examine whether this reference pattern is also observed in the other benchmarks.

Using the uniform DGP, we run the experiment for $N \in \{1{,}000, 10{,}000, 100{,}000\}$ over 10 Monte Carlo repetitions. For each sample size, we evaluate all subset sizes $m \in \{1, \ldots, p\}$ and compare the mean empirical benchmark with the Oracle benchmark and the theoretical Γ.

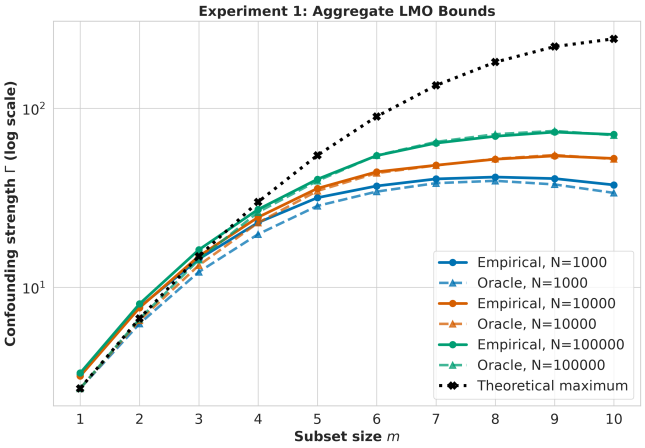

Figure 1: Empirical fitted and Oracle LMO bounds compared with the theoretical Γ. Observed bounds plateau while the theoretical Γ keeps increasing.

Figure 1 shows that the empirical bound increases at first, but then reaches a plateau and can slightly decrease for larger omitted subsets. Table 1 shows the same pattern numerically: for all three sample sizes, the empirical peak occurs before $m = 10$, although the final value remains close to the peak.

Table 1: Empirical peak and final LMO bound in Experiment 1.

N	Peak m	Peak $\hat{\Gamma}$	$\hat{\Gamma}_{m=10}$	Rec.$_{m=10}$
1k	8	41.28	37.30	0.152
10k	9	54.20	52.45	0.214
100k	9	73.63	71.40	0.292

The effect of sample size is also clear. Larger samples produce larger empirical Γ values and recover a greater fraction of the theoretical Γ. This supports the geometric interpretation: as N increases, the sample is more likely to contain individuals near the corner regions that generate large bounds. However, this improvement is limited. Even when $N = 100{,}000$, the empirical recovery at $m = 10$ is only 0.292, meaning that the observed sample realizes less than one third of the theoretical maximum.

Figure 2 shows this under-recovery more directly. For small m, the empirical benchmark can slightly exceed the theoretical Γ because it is based on ratios between the estimated propensity models, which can be wrong. However, both the empirical and Oracle recovery ratios decline as m increases. This indicates that the main limitation is not simply estimation error from the propensity model. Even when the benchmark is computed directly from the known DGP, the observed sample does not contain individuals that perfectly match the extreme corner configurations assumed by the theoretical ceiling. The empirical LMO benchmark is therefore constrained by the geometry of the finite sample, especially as more covariates are dropped.

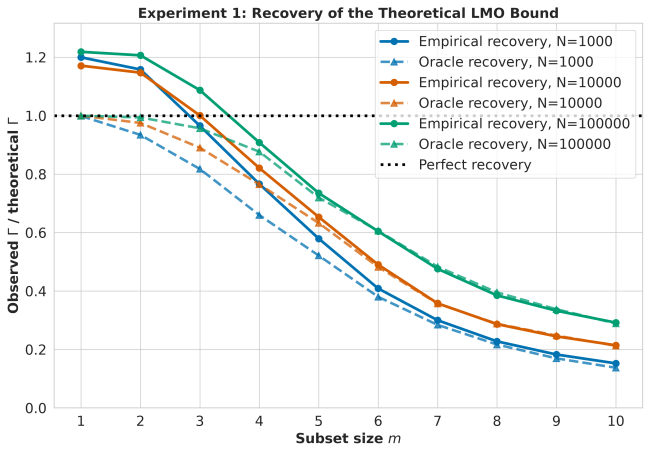

Figure 2: Recovery ratio of empirical fitted and Oracle LMO bounds relative to the theoretical Γ. Recovery declines as m increases, including for the Oracle benchmark.

4.2 Experiment 2: Individual Trajectories Behind the LMO Bound

Experiment 1 showed that the LMO benchmark can plateau and under-recover the theoretical maximum. Experiment 2 examines the individual-level structure behind this pattern. Using the same uniform DGP as in Experiment 1, with $N = 10{,}000$, we tracked the largest $\hat{\Gamma}$ value achieved by each individual across all possible subsets of dropped covariates.

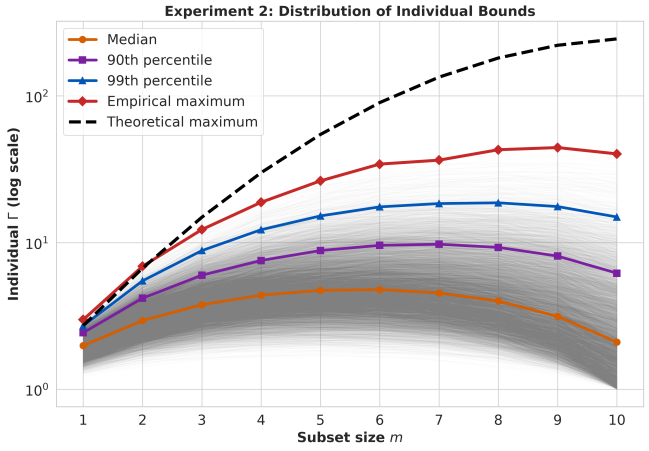

Figure 3: Distribution of individual bounds across omitted subset sizes. Semi-transparent gray lines show the trajectory of each individual. The median, 90th percentile, and 99th percentile individual bounds remain far below the theoretical maximum as m increases.

Figure 3 shows that most individuals remain far below the theoretical maximum. The maximum $\hat{\Gamma}$ found is not representative of a typical sample member. Instead, it is determined by a small number of observations in the extreme upper tail.

The mechanism can be seen from the individual dropped log-odds contribution,

$$\Delta_i(S) = \sum_{j \in S} X_{ij} w_j, \qquad \Gamma_i(S) = \exp(|\Delta_i(S)|).$$

Because the structural weights are positive in this DGP, an individual's Γ value is large when the omitted covariates have large magnitudes and the same sign. In that case, their weighted contributions reinforce one another. When the omitted covariates have mixed signs, the terms partially cancel inside the sum, reducing $|\Delta_i(S)|$ even if more covariates are omitted.

This explains the shape of the individual trajectories. For small and intermediate values of m, the best subset for a boundary individual can often select covariates that point in the same direction, so the trajectory increases. As m becomes larger, however, the subset must include more of that individual's remaining covariates. These additional covariates may be weaker or may point in the opposite direction, so they no longer add to the existing contribution. The resulting trajectory can then plateau or decline.

Experiment 2 therefore provides an individual-level explanation for the pattern observed in Experiment 1. The LMO benchmark is controlled by rare boundary individuals rather than by the average sample member. If the sample does not contain observations close to the required corner configurations, the empirical bound cannot reach the theoretical maximum.

4.3 Experiment 3: Corner Injection

If the plateau in the empirical curve is caused by missing or rare corner-aligned individuals, then adding such individuals should increase the observed LMO bound. Experiment 3 tests this prediction using the same uniform DGP as before, with $N = 10{,}000$, $p = 10$, and 10 Monte Carlo repetitions to generate the mean results. For each repetition, we compare two samples: a standard uniform sample and a corner-injected version. The corner-injected sample is the same as the standard sample, except that two observations are replaced by the positive and negative corners:

$$X_+ = (1, 1, \ldots, 1), \qquad X_- = (-1, -1, \ldots, -1).$$

They are the configurations most favorable to producing large odds-ratio shifts when multiple covariates are omitted, because their values align across all dimensions. Treatment assignments for these injected observations are generated using the same treatment-assignment rule as in the original DGP.

Figure 4 shows that the samples with injected corner observations produce larger empirical LMO bounds than the standard uniform samples. This supports the interpretation from Experiment 2: the empirical bound is not determined only by the structural weights w, but also by whether the observed sample contains individuals whose covariate profiles align with those weights.

The theoretical maximum is therefore attainable in principle, but only when the sample contains observations near the relevant geometric extremes. In ordinary finite samples, especially in higher dimensions, such observations may be rare or absent.

4.4 Experiment 4: Covariate Geometry

Experiment 4 examines whether the fitted LMO benchmark changes with the geometry of the covariate distribution. As

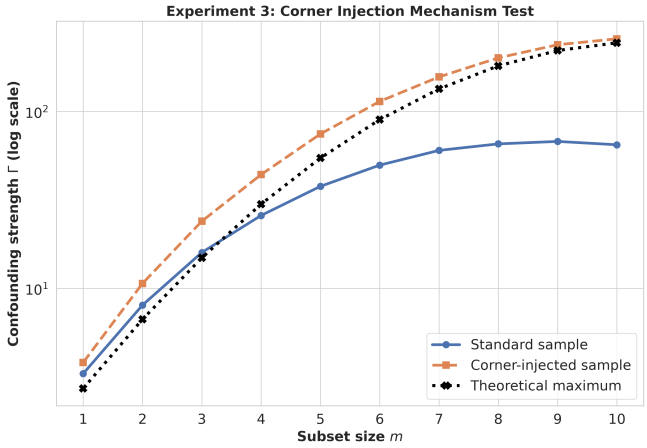

Figure 4: Comparison between standard uniform samples and samples with injected corner observations. Adding extreme corner individuals increases the empirical LMO bound and moves it closer to the theoretical maximum.

in previous experiments, we use $N = 10{,}000$, $p = 10$, and 10 Monte Carlo repetitions. We compare the three covariate distributions defined in the methodology: a boundary-heavy Beta distribution, a uniform distribution, and a center-heavy Beta distribution. Across these settings, the treatment-assignment mechanism is kept fixed.

If our previous conclusions hold, then the U-shaped distribution should produce larger fitted bounds because it places more probability mass near the edges of the covariate space. By contrast, the bell-shaped distribution should produce smaller fitted bounds because most observations are concentrated near the center.

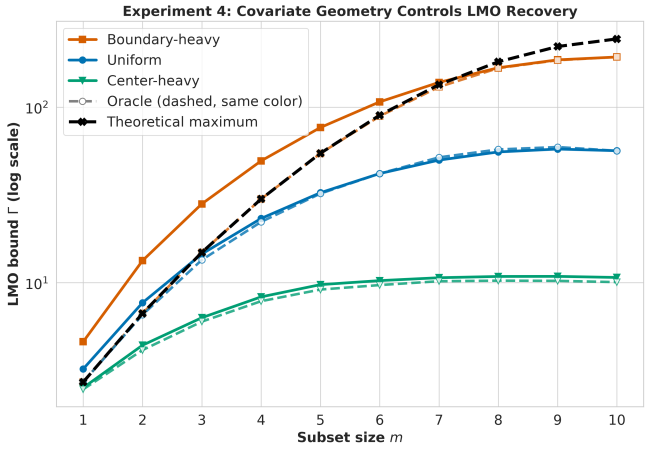

Figure 5: Fitted and Oracle LMO bounds across three covariate geometries. Solid lines show fitted benchmarks, dashed colored lines show Oracle structural benchmarks, and the black dashed line shows the theoretical maximum. Boundary-heavy covariates yield the largest bounds, while center-heavy covariates produce the smallest bounds.

Figure 5 shows the expected ordering for both fitted and Oracle bounds. The U-shaped Beta distribution produces the largest bounds, consistent with its greater concentration of observations near the boundary of the covariate space. The fitted boundary-heavy curve can exceed the theoretical ceiling for smaller subset sizes, reflecting the same estimation error discussed in Experiment 1, as the more extreme values hurt the logistic regression fit. The Oracle curve instead almost perfectly tracks the structural contribution represented in the sample, up to $m = 7$, and then slightly under-recovers it for larger subset sizes.

The uniform distribution is as expected in the middle, as seen in the previous experiments.

The bell-shaped Beta distribution produces the smallest bounds, since most observations lie near zero and omitted covariates therefore generate weaker log-odds shifts.

This experiment shows that the LMO benchmark is sensitive to sample space captured. When the distribution places sufficient probability mass near the relevant extremes, the fitted benchmark can recover a larger fraction of the theoretical bound. When the distribution concentrates mass near the center, the sample-realized bound can severely under-recover the theoretical confounding strength.

5 Discussion

The experiments show that LMO informal benchmarking depends not only on the structural strength of the omitted covariates, but also on which covariate patterns are present in the observed sample. The key mechanism is directional alignment: confounding is strongest when the omitted covariates reinforce each other in the treatment-assignment model. In our DGP, all weights are positive, so this happens when omitted covariates have the same sign and large magnitude. As m increases, this becomes harder to realize in the sample, which can cause the empirical benchmark to plateau or decline.

The Oracle benchmark separates this effect from propensity-model estimation error. Computed directly from the known DGP, it reflects whether the strongest confounding configurations are represented in the observed sample. Declining Oracle recovery shows that the plateau is not an artifact of propensity-model estimation error, but rather a consequence of the sample.

The practical implication is that informal benchmarking should be interpreted as a sample-realized benchmark, not as a measure of the maximum confounding strength possible over the full covariate space. There is a difference between the theoretical Γ that covariates could produce in principle and the empirical $\hat{\Gamma}$ recovered by dropping them in the observed sample. Informal benchmarking treats the empirical value as a proxy for the theoretical. This matters because under-recovery can make robustness appear stronger than it is. If the maximum possible Γ is underestimated, informal benchmarking may give false reassurance about the stability of the causal conclusion under MSM.

More generally, IB carries two layers of uncertainty. First, it relies on the assumption that the confounding implied by observed covariates is informative about possible hidden confounding. Second, even for the observed covariates, the finite sample may not contain the strongest confounding configuration present in the population. The second uncertainty

becomes increasingly important as the number of covariates grows. For small values of m, only a few covariates need to align, so the empirical benchmark is more likely to capture the maximum confounding. As m grows, however, the theoretical maximum depends on increasingly specific combinations of covariates aligning at the same time, which becomes less likely in a finite sample.

However, the problem with interpreting the plateau as an indication of underestimation is that the plateau may also reflect a genuine decline in confounding strength as more covariates are dropped. This is why greater understanding of the underlying problem is needed to interpret the benchmark. For example, knowing whether the covariates are bell-shaped or U-shaped can help determine whether the benchmark underestimates the true confounding strength or reflects the lower importance of certain covariates.

Several limitations follow from the controlled simulation design. The covariates are independent, bounded, and continuous, the treatment mechanism is logistic and the structural weights are fixed and positive. Real observational data may include correlated features, categorical variables, nonlinear assignment rules, unknown covariate bounds, and proxy variables. Correlations may make some corner configurations impossible rather than merely rare, while proxies may create other forms of alignment not captured here.

6 Responsible Research

This study uses only synthetic data generated from explicitly defined data-generating processes. It does not involve personal data, real patient records, consent issues, or direct privacy and data-security risks.

The main ethical risk lies in interpretation. Sensitivity analysis is used in medical, sociological, and policy-oriented observational research, where causal conclusions may influence treatment decisions, institutional choices, or claims about social interventions. If an LMO benchmark is interpreted as the maximum possible hidden confounding strength, rather than as the strongest confounding pattern represented in the observed sample, it may give false reassurance about the robustness of such conclusions. This risk can persist beyond the original study. Even if the original assumptions are stated clearly, later work may cite the result without them, making the claim appear more general than it is.

Real observational data may include correlated features, categorical variables, nonlinear assignment, measurement error, and limited overlap. Therefore, the results of this study should be interpreted with caution and not treated as universally applicable. Instead, they should be seen as evidence that IB can be sensitive to the observed sample, and that its reliability should be examined carefully in each applied setting.

To support reproducibility, the data-generating processes, structural weights, covariate distributions, sample sizes, and evaluation metrics are specified in the methodology. The experimental scripts use fixed and deterministic seeds, and the repository contains the code used to run the experiments: https://github.com/NaidenBoro/LMO-Informal-Benchmarking.

AI-assisted tools and LLMs were used during the project for writing support, LaTeX editing, code generation, debugging, and clarification of explanations. They were not used as a substitute for methodological decisions or interpretation, and all code changes, numerical outputs, and scientific claims were reviewed by the author.

7 Conclusion

This paper addressed a practical problem in informal benchmarking for sensitivity analysis: when multiple observed covariates are removed at once, it is unclear whether the resulting LMO benchmark reliably captures the strength of confounding implied by those covariates. The research questions were how the empirical LMO benchmark behaves as the number of dropped features increases, and how this behavior depends on the structure of the data.

The experiments answer these questions by showing that the empirical LMO benchmark does not necessarily increase with the number of omitted covariates, even though the theoretical maximum does. Instead, it can plateau and eventually decline. This means that dropping more covariates does not necessarily produce a stronger empirical benchmark. Larger samples improve recovery, but do not remove the gap completely. As m grows, the theoretical ceiling depends on increasingly specific combinations of covariates, making the relevant high-shift configurations less likely to appear in a finite sample.

The main reason is not only propensity-model estimation error. Even the Oracle benchmark, computed directly from the known DGP, under-recovers the theoretical Γ when the observed sample lacks the specific individuals producing that ceiling. Individual trajectories, corner injection, and alternative covariate distributions all point to the same mechanism: the LMO informal benchmark is governed by the extremes present in the sample.

The main conclusion is therefore interpretive. LMO informal benchmarking for larger numbers of omitted covariates should be read as a sample-realized benchmark rather than the maximum hidden confounding strength truly possible. If the calibrated Γ comes from a sample-realized benchmark, then robustness claims under MSM based on that value are most defensible for the datapoints in the sample. They should not automatically be read as guarantees for the entire population. Failing to make this distinction can lead to false reassurance about the robustness of causal conclusions.

Future work should test this mechanism in less idealized settings, including correlated covariates, mixed continuous and categorical features, nonlinear treatment-assignment rules, and unknown covariate bounds. Such work could help distinguish genuinely weak hidden-confounding benchmarks from benchmarks that appear weak because the observed data do not cover the relevant parts of the covariate space.

Overall, the results suggest that LMO informal benchmarking should be interpreted with increasing caution as larger subsets of covariates are dropped. It may understate the true strength of hidden confounding, which could lead to overly optimistic assessments of robustness. It is not a failure of the method, but rather a reflection of its finite-sample nature.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424. PMID: 21818162.

Baitairian, J.-B., Sebastien, B., Jreich, R., Katsahian, S., and Guilloux, A. (2026). Calibrating confounding strength in sensitivity models for weighting estimators: a comparative review and a new method.

Braga, L. H., Farrokhyar, F., Dönmez, M. İ., Nelson, C. P., Haid, B., Herbst, K., Garriboli, M., Cascio, S., Nieuwhof-Leppink, A., Kaefer, M., Bägli, D. J., Kalfa, N., Ching, C., Fossum, M., and Harper, L. (2025). Randomized controlled trials – the what, when, how and why. *Journal of Pediatric Urology*, 21(2):397–404.

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

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In Everitt, B. S. and Howell, D. C., editors, *Encyclopedia of Statistics in Behavioral Science*, pages 1809–1814. John Wiley & Sons.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

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

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5):1–46.

Zabor, E. C., Kaizer, A. M., and Hobbs, B. P. (2020). Randomized controlled trials. *CHEST*, 158(1):S79–S87.

A Appendix: Analytical Derivation of the Theoretical Maximum Bound

We derive the theoretical maximum confounding bound, $\Gamma^{(m)}_{\text{theory}}$, under a logistic Data Generating Process where covariates are bounded $X \in [-1, 1]^p$ and $\mathbf{w} \in \mathbb{R}^p$ represents the true structural weights.

Step 1: Full Odds and Dropped Structural Contribution

The true propensity score is $e(X) = 1/(1 + \exp(-X^\top \mathbf{w}))$. The full-information odds of treatment are:

$$\text{O}_{\text{full}}(X) = \exp(X^\top \mathbf{w}) = \exp\left(\sum_{j=1}^{p} X_j w_j\right) \tag{14}$$

Let $S \subset \{1, \ldots, p\}$ be a subset of omitted features of size m. The structural log-odds contribution of these omitted features is:

$$\Delta_S(X) = \sum_{j \in S} X_j w_j. \tag{15}$$

This is not the logit of the true reduced propensity $P(T = 1 \mid X_{-S})$, which would require marginalizing over the omitted variables. It is the structural contribution removed from the full linear predictor when the covariates in S are treated as hidden, and is the quantity used by the Oracle benchmark in this paper.

Step 2: Individual and Subset Bounds

The Oracle dropped-logit benchmark for individual i is the odds-ratio shift implied by this removed contribution:

$$\begin{aligned} \Gamma_i^{(S)} &= \exp\left(|\Delta_S(X_i)|\right) \\ &= \exp\left(\left|\sum_{j \in S} X_{ij} w_j\right|\right). \end{aligned} \tag{16}$$

The global bound $\Gamma^{(S)}$ is the maximum over the domain $X \in [-1, 1]^p$. This is mathematically maximized when each X_j takes its extreme value matching the sign of w_j (i.e., $X_j = \text{sgn}(w_j)$):

$$\begin{aligned} \Gamma^{(S)} &= \max_{X \in [-1,1]^p} \exp\left(\left|\sum_{j \in S} X_j w_j\right|\right) \\ &= \exp\left(\sum_{j \in S} |w_j|\right) \end{aligned} \tag{17}$$

Step 3: The Multidimensional Maximum

To find the absolute worst-case bound for any subset of size m, we select the m features with the largest absolute weights. Sorting the weights such that $|w|_{(1)} \geq \cdots \geq |w|_{(p)}$ yields the theoretical analytical ceiling:

$$\Gamma^{(m)}_{\text{theory}} = \exp\left(\sum_{k=1}^{m} |w|_{(k)}\right) \tag{18}$$