Leave-Multiple-Out Informal Benchmarking

A Simulation Study on the Impact of Sample Size

Maja Czerwińska[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: Maja Czerwińska
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Matej Havelka, Avishek Anand

Abstract

In observational studies, establishing robust causal relationships is frequently challenged by the presence of hidden variables that distort the true relationships within the data. While sensitivity analysis offers a formal framework to assess how vulnerable a study's conclusions are to such hidden biases, its widespread adoption is often hindered by complex mathematical assumptions and interpretation hurdles. To make these tools more intuitive, researchers frequently employ informal benchmarking—a technique that uses the measured explanatory power of observed features as a baseline to calibrate the hypothetical strength a hidden variable would need to overturn a conclusion. While widely used, the behavior of these benchmarks across varying sample scales lacks a systematic evaluation under different structural conditions. This thesis addresses this gap by investigating the sample size dynamics within the Leave-Multiple-Out (LMO) informal benchmarking framework. Leveraging an extended simulation grid across 18 structural scenarios and sample sizes ranging from $N = 20$ to $N = 50,000$, we demonstrate that increasing data volume does not produce a uniform trajectory. Instead, the benchmark follows two entirely different paths depending on the predictive power of the available features. In weak or uncorrelated settings, increasing the sample size acts as a variance reducer, eliminating small-sample overfitting ($N \leq 200$) and stabilizing scores at a conservative baseline. In contrast, in strong or highly correlated settings where true propensity scores span a wide range, scaling the sample size grants well-specified models the statistical power to accurately map the extreme tails of the distribution. Because the non-linear mechanics of odds ratio calculations are highly sensitive to probabilities near 0 or 1, these captured tail cases drive the benchmark scores steadily upward. While the framework successfully reflects model quality as the sample size increases, these dynamics prompt a theoretical warning: the benchmark inherently rewards the generation of extreme probabilities, rendering it potentially vulnerable to overparameterized or uncalibrated estimators.

1 Introduction

In science, one often wants to study how one phenomenon affects another. The best currently available method of conducting such a study is a Randomized Clinical Trial [1], in which subjects are randomly assigned to a control or a treatment group. Although very insightful, such studies come at a high cost and could have insufficient statistical power or even be unethical [2]. This is why scientists often turn to observational data. These data are readily available in large quantities; however, since participants self-assign, the two groups - one receiving a treatment or participating in an activity and another that is not - could differ substantially and be no longer comparable. In the presence of significant differences, it is difficult to tell whether it was the treatment that caused an effect for one group but not for the other or whether it was some unmeasured variable that put people in their groups in the first place.

Sensitivity analysis measures the robustness of a study's findings by determining how strong a hidden confounder, a variable affecting both treatment and outcome, would have to be to alter the conclusion [3]. Despite its importance, sensitivity analysis is underutilized in empirical research. Complex assumptions, the lack of conventions on reporting the results and the difficulty of translating them into the real-world are named as the reasons [4]. To bridge this interpretability gap, researchers frequently employ informal benchmarking (e.g., [4], [5], [6]). This practice involves estimating the confounding strength of observed variables already present in the dataset and using them as real-world baselines to calibrate the hypothetical strength an unmeasured confounder would need to invalidate the study.

While advanced formal frameworks explicitly incorporate mathematical properties to handle finite sample uncertainty (e.g., [7], [8]), to the best of our knowledge, it remains unexamined how varying the number of samples impacts the results of the popular informal benchmarks. This seems like a critical gap, given that practitioners frequently rely on large sample sizes as a safeguard against unreliable statistics [9]. This motivates our research question: How does the sample size affect the behavior of informal benchmarking? By studying how benchmark values change across different sample sizes and feature profiles, this work aims to improve the understanding of informal benchmarking, provide empirical evidence about its behavior in large-sample settings, and warn against common misconceptions regarding its use.

To address this question, the framework proposed by Baitairian et al. [10] was adapted and extended to allow for a wider range of testing scenarios and to investigate how sample size affects them. The simulated environments varied the true predictive power of observed features, their correlation with the hidden confounder, and the true strength of the hidden confounder. Because the data-generating process incorporates non-linear transformations, calculating whether the benchmark mathematically recovers an exact analytical truth on the combined features' strength remains an open question. Instead, this work focuses on evaluating the tool's empirical behavior and consistency across different settings as the sample size scales.

Our results demonstrate that an increasing sample size acts primarily as a variance reducer, tightening the wide and chaotic bounds observed in small-sample settings ($N \leq 200$). However, the exact behavior of the benchmark as the sample size grows is entirely dependent on the strength of the feature profile. When observed features are relatively weak and uncorrelated with the hidden confounder (but the confounder is significant for making predictions), a larger sample size stabilizes the benchmark outputs at a low value, reflecting the informational limitations of a weak model. In contrast, when the model is strong, either because the observed fea-

tures are highly predictive or directly correlated with an important hidden confounder, the benchmark's scores exhibit a clear, steady increase with sample size once the extreme sampling variability of low-sample settings subsides. Across our tested scenarios, better-specified models consistently output higher scores, demonstrating that the benchmark effectively differentiates between predictive and non-predictive specifications.

While no simulation can exhaustively capture every real-world misspecification, within the defined scope (detailed in the Methodology and Appendix A), our work demonstrates that the framework's large-sample behavior is highly predictable, offering practitioners a baseline for how the tool responds to data volume.

The remainder of this paper is organized as follows. Section 2 provides a formal overview of informal benchmarking, followed by Section 3, which details the methodology used for data generation. Section 4 describes the design of the experiments and their results, while Section 5 addresses the ethical considerations of this project. The experimental results are analyzed and discussed in Section 6. Finally, Section 7 presents our conclusions and outlines directions for future work.

2 Background

In this section, the mathematical framework and notation necessary for implementing and evaluating informal benchmarking is described.

2.1 Notation and the Marginal Sensitivity Model

Throughout the remainder of this paper, $\mathbf{X} \in \mathbb{R}^{p_\mathbf{X}}$ represents a vector of measured covariates, and $\mathbf{U} \in \mathbb{R}^{p_\mathbf{U}}$ represents a vector of hidden confounders. The treatment assignment is defined as a binary variable $T \in \{0, 1\}$ and Y denotes the observed outcome. While the term "treatment" is historically derived from clinical trial literature, it is used here in a generalized sense to represent any binary exposure, action, or intervention, such as participation in a program or exposure to a specific condition. To maintain a focused analysis, continuous treatments are omitted from this study.

The primary objective of causal inference models is to find a true causal effect of T on Y rather than mere correlations. However, these models operate under the strong ignorability assumption.

Definition 1 (Strong Ignorability)**.** Potential outcomes under treatment ($Y(1)$) and control ($Y(0)$) are conditionally independent of treatment assignment given the observed covariates:

$$(Y(1), Y(0)) \perp T \mid \mathbf{X} \tag{1}$$

This means that all variables affecting treatment and outcome are captured within $\mathbf{X}$. In real-world observational environments, this strict requirement is rarely satisfied because there is almost always a risk of unmeasured confounding variables. Consequently, researchers rely on sensitivity analysis frameworks to relax the strong ignorability assumption, allowing for a specified degree of hidden confounding.

While several different sensitivity frameworks exist, this analysis utilizes the definition of confounding from the Marginal Sensitivity Model (MSM) proposed by Tan [11]. This definition focuses specifically on the mechanism of selection into treatment, evaluating how unobserved factors alter the odds of receiving treatment (Definition 2). This framework ignores outcome mechanisms entirely, and this study follows the same restriction.

Definition 2 (Nominal Treatment Odds)**.** Let $e(\mathbf{X}) = P(T = 1 \mid \mathbf{X})$ denote the nominal propensity score, representing the probability of receiving treatment conditional on the observed covariates $\mathbf{X}$. The baseline nominal treatment odds are defined as:

$$\mathbf{\Omega}(\mathbf{X}) = \frac{e(\mathbf{X})}{1 - e(\mathbf{X})} \tag{2}$$

Definition 3 (Marginal Sensitivity Model Bound)**.** The MSM bounds the maximum degree of hidden confounding by introducing a sensitivity parameter $\Gamma \geq 1$, formulated as:

$$\frac{1}{\Gamma} \leq \frac{\mathbf{\Omega}(\mathbf{X}, \mathbf{U})}{\mathbf{\Omega}(\mathbf{X})} \leq \Gamma \tag{3}$$

This mathematical bound dictates how much an individual's true treatment odds, which depend on both observed and unobserved variables ($e(\mathbf{X}, \mathbf{U}) = P(T = 1 \mid \mathbf{X}, \mathbf{U})$), can deviate from their nominal treatment odds defined in Equation 2. $\Gamma = 1$ denotes no hidden confounding, which satisfies the strong ignorability assumption (Definition 1). As Γ increases, the model allows for progressively stronger hidden confounders. The ultimate goal of sensitivity analysis is to find the specific parameter Γ at which the estimated causal relationship loses statistical significance or changes direction.

2.2 Informal Benchmarking

Informal benchmarking allows researchers to evaluate the threat of a potential unobserved confounder by comparing it directly to the known, observed variables in the dataset. With various implementations existing, most often using Leave-One-Out or Leave-Multiple-Out procedures [12] [13] [4], the core mechanism relies on systematically removing observed covariates to isolate and measure their impact.

2.2.1 Implementation

In this paper, the version found in Baitairian et al. [10] was used. Their paper synthesizes a body of work on implementations of informal benchmarking into one standard and easy-to-understand pseudo-algorithm, which is non-parametric and model-agnostic. The execution of the algorithm is as follows; the nominal propensity score, the baseline probability of a subject receiving treatment given all observed data, is first estimated. The algorithm then systematically obstructs data by deleting every possible combination of measured features (excluding the removal of all of them). For each subset of remaining features, the propensity scores are re-calculated. The odds ratio between the baseline and the propensity scores obtained from the reduced model is recorded for each data point, and the maximum observed ratio across all combinations is selected as the benchmark score. This value represents the maximum shift in treatment odds caused by omitting these observed variables, serving as

a reference point for sensitivity analysis. This procedure is formalized in Algorithm 1.

An important clarification must be made regarding this framework. While the text of Baitairian et al. [10] primarily details a Leave-One-Out (LOO) procedure, their official open-source repository implements a Leave-Multiple-Out (LMO) routine. To ensure exact alignment with the actual software tool provided to practitioners, this study utilizes the repository's native LMO implementation. Consequently, our analysis isolates how sample size dynamics affect these multi-variable benchmark bounds as they are computed in practice.

2.2.2 Criticism

This way of analysis is not exempt from critiques. One of its issues is the untestable positivity assumption, which states that every propensity score must be strictly higher than 0 and strictly lower than 1. In plain English, every data point needs a nonzero chance of both receiving and not receiving the treatment. This is not always the case, for example, when testing a drug that pregnant women cannot take. Because the underlying math uses fractions of the form $x/(1-x)$, these edge case situations result in a division-by-zero error. Another criticism is the sensitivity to the method used to estimate the propensity scores. Because different models, such as a logistic regression versus a random forest, handle feature interactions and correlation differently, switching the underlying estimator can drastically alter the final benchmark bounds. Finally, a fundamental limitation of this approach is that removing a feature or a subset of them does not necessarily measure the isolated strength of that specific combination. In real-world datasets, variables rarely exist in isolation. Consequently, when the algorithm deletes a subset of data, the resulting change in the model could be distorted by correlations between features, hidden and measured. This could cause the informal benchmark to over- or underestimate the true strength of features.

Despite all its issues, informal benchmarking remains the only method known today that does not require specific data types or restrictions on them. Because it is so accessible, many researchers rely on it. This makes studying its performance worthwhile.

3 Methodology

In order to examine the behavior of the Leave-Multiple-Out informal benchmark as sample size changes, a synthetic data generation process inspired by Baitairian et al. [10] was implemented. We construct controlled simulation scenarios where a fixed level of hidden confounding (Γ), in accordance with the definition of the Marginal Sensitivity Model (MSM), is explicitly injected into the data. The following subsections detail the data generation process and the implementation of the LMO benchmark.

3.1 Data Generation

First, the observed covariates $\mathbf{X} \in \mathbb{R}^{p_{\mathbf{X}}}$ are generated from a uniform multivariate distribution $\mathcal{U}(-1,1)^{p_{\mathbf{X}}}$. Next, to introduce a controllable dependency between the observed and unobserved variables, an intermediate raw confounding signal U_{raw} is generated. The linear combination of features $\beta^T\mathbf{X}$ is standardized to unit variance, and the confounder is drawn conditionally as:

$$U_{\text{raw}} \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left((1-\lambda)\frac{\beta^T\mathbf{x}}{\sigma_{\beta^T\mathbf{X}}},\ 1-(1-\lambda)^2\right)$$

Here, $\beta \in \mathbb{R}^{p_{\mathbf{X}}}$ assigns weights to each feature, and the $\lambda \in [0,1]$ parameter dictates the degree of correlation. $\lambda = 1$ results in a standard normal distribution $\mathcal{N}(0,1)$ and represents the uncorrelated case. The conditional variance is defined as $1-(1-\lambda)^2$ to ensure that the total variance of U_{raw} remains 1.0. This raw variable is then passed through a standard normal cumulative distribution function (CDF) to map it cleanly to a $(0,1)$ interval, yielding the final unobserved confounder U.

It is worth noting that this setup consciously departs from the original framework in Baitairian et al. [10], where $\mathbf{U}$ is multidimensional. Reducing the hidden confounder to a single dimension allows the confounding signal to propagate directly into the system, preventing it from being distorted through aggregations. When Baitairian et al. [10] take a mean along the dimensions to pass that further into the system, the values cluster around the middle, preventing them from taking on extreme values and investigating scenarios of full-strength confounder. By keeping U single-dimensional, we preserve more of its impact.

It is worth noting that this setup consciously departs from the framework in Baitairian et al. [10], where $\mathbf{U}$ is multidimensional. Reducing the hidden confounder to a single dimension allows the confounding signal to propagate directly into the system, preventing it from being diluted through aggregation. When Baitairian et al. take a mean across multiple dimensions, the values naturally cluster around the middle. This compression prevents the confounder from taking on extreme values, making it impossible to simulate scenarios in which it exerts its full strength. By keeping U single-dimensional, we better preserve its true impact. This gives us a much more valid setup when evaluating how a confounder of a specific strength actually affects the results.

Nominal (observed) propensity score generation is based on a logistic function, with a 0.5 constant added to introduce class imbalance:

$$e^{\text{obs}} = \mathbb{P}(T=1|\mathbf{X}) = \text{logistic}(\delta^T\mathbf{X} + 0.5)$$

where $\delta \in \mathbb{R}^{p_{\mathbf{X}}}$ represents the true feature importance weights.

Our treatment assignment mechanism also diverges from the original implementation. Baitairian et al. introduce strict mathematical constraints forcing the true propensity scores $e(\mathbf{X},\mathbf{U})$ to marginalize exactly to the nominal propensity scores $e(\mathbf{X})$. As they note in their own remarks, this effectively ensures that altering the correlation between the observed covariates $\mathbf{X}$ and the unobserved confounder U has no impact on the ultimate treatment assignment T. While this represents a perfectly valid theoretical edge case to explore, in real-world observational data, treatment assignment is rarely so isolated; rather, it is jointly driven by multiple baseline characteristics and the complex interactions between

Algorithm 1: Informal Benchmarking (Leave-Multiple-Out, which includes Leave-One-Out, for Binary Treatment)

Require: Dataset containing binary treatment vector $\mathbf{T}$ and a matrix of observed confounders $\mathbf{X} \in p_{\mathbf{X}}$.

Estimate the full propensity score $\hat{e}(\mathbf{X})$ using all observed confounders;
Compute the full baseline odds: $\mathbf{\Omega} := \hat{e}(\mathbf{X}) / \big(1 - \hat{e}(\mathbf{X})\big)$;
for *subset size* $k \in \{1, \ldots, p_{\mathbf{X}} - 1\}$ **do**
    Identify all possible combinations $\mathcal{C}_k$ of k features to drop;
    for *each combination* $\mathbf{c} \in \mathcal{C}_k$ **do**
        Let $\mathbf{X}^{(-\mathbf{c})}$ be the data matrix excluding the features in $\mathbf{c}$;
        Estimate the reduced propensity score $\hat{e}\big(\mathbf{X}^{(-\mathbf{c})}\big)$;
        Compute the reduced odds: $\mathbf{\Omega}^{(-\mathbf{c})} := \hat{e}\big(\mathbf{X}^{(-\mathbf{c})}\big) / \big(1 - \hat{e}\big(\mathbf{X}^{(-\mathbf{c})}\big)\big)$;
        for *each individual observation* $j \in \{1, \ldots, n\}$ **do**
            Calculate the odds ratio: $r_{\mathbf{c},j} := \mathbf{\Omega}_j / \mathbf{\Omega}_j^{(-\mathbf{c})}$;
        Compute the subset-specific bound: $\hat{\Gamma}_{\mathbf{c}} := \max\left(\max_j(r_{\mathbf{c},j}),\ \frac{1}{\min_j(r_{\mathbf{c},j})}\right)$;
Identify the worst-case benchmark bound: $\hat{\Gamma}_{\text{high}} := \max_{\mathbf{c}}(\hat{\Gamma}_{\mathbf{c}})$;

them [14]. This is why this study departs from their approach, making it possible to explore a much wider and more realistic range of environments. We work directly with the bounds defined by the Marginal Structural Model (MSM). We calculate the nominal odds $\Omega(\mathbf{X}) = e(\mathbf{X})/(1 - e(\mathbf{X}))$ and scale them to obtain the true odds:

$$\Omega(\mathbf{X}, U) = \Omega(\mathbf{X}) \cdot \Gamma^{h(U)}$$

where $h(U) = 2(U - 0.5)$ transforms the confounder into the range $[-1, 1]$. This enforces the MSM bounds:

$$\frac{1}{\Gamma} \leq \frac{\Omega_{\text{true}}}{\Omega_{\text{obs}}} \leq \Gamma$$

By leveraging the mathematical definition of odds (Definition 2), we extract the true probability of treatment $e(\mathbf{X}, U)$ to generate the final binary assignment via a Bernoulli trial, $T \sim \text{Bernoulli}(e(\mathbf{X}, U))$. Exact algebraic procedures and coefficient matrices (β, δ) are detailed in Appendix A.

3.2 Benchmark Implementation

To execute the informal benchmark, a Leave-Multiple-Out (LMO) routine was implemented in accordance with Algorithm 1. Following Baitairian et al. [10], we utilized 5-fold cross-validation to estimate the baseline and reduced propensity scores. However, while their framework relies on standard, unregularized logistic regression, this approach frequently suffers from perfect separation in small-sample settings, causing probabilities extremely close to 0 and 1, which result in severe numerical explosions when calculating the odds ratios. To resolve this, we modified their setup to utilize L2-penalized logistic regression. Finally, to prevent remaining edge cases from causing division-by-zero errors, all estimated probabilities are clipped within the interval $[10^{-15}, 1 - 10^{-15}]$.

4 Experimental Results

In this section, we present the empirical findings of our simulation study, beginning with the parameters of our experimental setup before analyzing the performance of the benchmark across different feature settings.

4.1 Experimental Design

To investigate whether different data environments alter the behavior of the informal benchmark as sample size increases, we systematically varied key parameters across a simulation grid. As detailed in Table 1, we tested 18 distinct structural scenarios by altering the hidden confounding strength ($\Gamma \in \{4, 8, 16\}$), the feature importance profiles (uniform vs. skewed), and the covariate-confounder correlation structure (uncorrelated, low, or high).

The feature importance profiles are governed by the coefficient vector $\delta \in \mathbb{R}^{p_{\mathbf{X}}}$, which dictates each covariate's weight in the treatment assignment mechanism. For the uniform configuration, the coefficients are drawn such that $\delta_i \sim \mathcal{U}(0.1, 0.5)$. In contrast, the skewed configuration concentrates predictive power into a few dominant variables using the fixed vector $\delta = [1.5, 0.8, 0.2, 0.05, 0.00]^T$.

Each scenario was evaluated across 13 distinct sample sizes ($N \in \{20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 15000, 20000, 25000, 50000\}$). For every parameter combination, we executed 50 independent Monte Carlo iterations to ensure statistical stability and mitigate sampling error.

4.2 Features of Equal Importance

The equal-importance configuration adapts the baseline setup from Baitairian et al. [10] to ensure our simulation parameters remain grounded in their original framework. Because our data generation process allows the correlation between $\mathbf{X}$ and U to directly propagate into the treatment assignment mechanism, we also examine how these dependencies affect the empirical behavior of the benchmark. Each scenario is evaluated across multiple true Γ strengths of a hidden confounder.

The results of this part of the analysis are summarized in Figure 1(a). The figure is organized as a 3×3 grid, with

Table 1: Systematic Overview of Experimental Configurations and Parameter Settings.

Exp. ID	$p_{\mathbf{X}}$	Correlation Type	true Γ	Feature Importance
#0	5	Uncorrelated	8	Uniform / Default
#1	5	Uncorrelated	16	Uniform / Default
#2	5	Uncorrelated	4	Uniform / Default
#3	5	Low	8	Uniform / Default
#4	5	Low	16	Uniform / Default
#5	5	Low	4	Uniform / Default
#6	5	High	8	Uniform / Default
#7	5	High	16	Uniform / Default
#8	5	High	4	Uniform / Default
#9	5	Uncorrelated	8	Skewed Importance
#10	5	Uncorrelated	16	Skewed Importance
#11	5	Uncorrelated	4	Skewed Importance
#12	5	Low	8	Skewed Importance
#13	5	Low	16	Skewed Importance
#14	5	Low	4	Skewed Importance
#15	5	High	8	Skewed Importance
#16	5	High	16	Skewed Importance
#17	5	High	4	Skewed Importance

rows corresponding to the covariate-confounder correlation strength (uncorrelated, low, high) and columns to the strength of a hidden confounder ($\Gamma \in \{4, 8, 16\}$). To maintain a uniform scale across all subplots and prevent varying axis ranges from distorting cross-scenario comparisons, the y-axis on these plots is clipped at 20. The corresponding graphs with full axes are available in Appendix B.

A clear divergence emerges between the high-correlation setting and the other two environments. In the uncorrelated case (top row), the estimates stabilize around a low value as the sample size increases. In the low-correlation case (middle row), the overall trajectory remains similar, although the final baseline settles at a slightly higher value than in the uncorrelated setting. Across both rows, the estimates exhibit a minor inverse relationship with the hidden confounding strength: a higher true Γ leads to marginally lower benchmark values. In both settings, the estimates quickly stabilize and compress to a tight variance by $N = 500$.

The high-correlation setting (bottom row) completely reverses these trends. Rather than flattening out, the estimated benchmark values keep climbing across the entire evaluated sample range, with higher true Γ values accelerating this growth. Furthermore, the variance fails to collapse; past $N = 500$, the boxplots maintain a wide, steady spread all the way through $N = 50,000$.

4.3 Features of Varying Importance

To evaluate the behavior of the benchmark under more realistic data structures, a skewed assignment of feature weights (through δ) was introduced. In practical applications, observed covariates rarely contribute equally to treatment assignment. By concentrating predictive power within a subset of features, this configuration tests whether the omission of very influential variables inflates the estimated bounds or distorts convergence behavior as sample size grows.

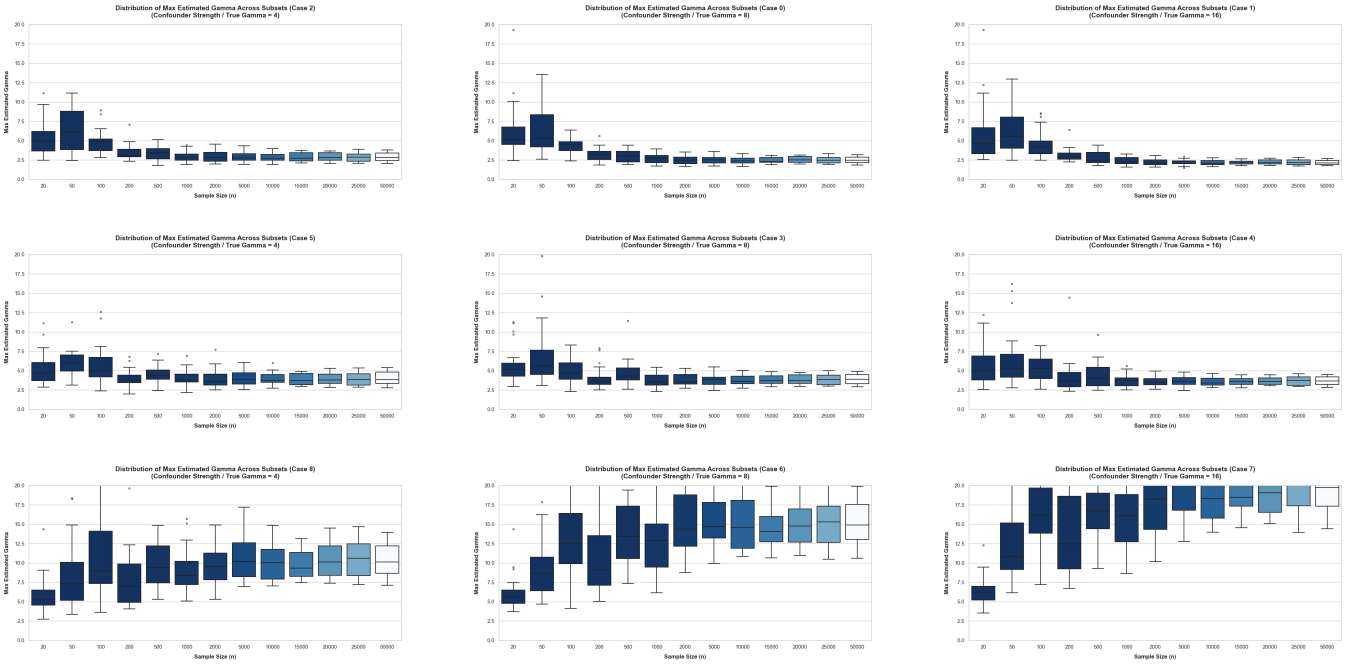
The results for this asymmetric configuration are presented in Figure 1(b). When evaluating this environment, a distinct pattern emerges: the skewed feature profile raises the benchmark values across all settings while significantly increasing the number of samples required for the estimates to settle.

In both the uncorrelated and low-correlation settings (top and middle rows), the benchmark values follow a similar path that departs noticeably from the uniform case. Instead of decreasing immediately, the estimates in both environments initially rise to a peak at $N = 100$ before gradually decreasing and flattening out. Stabilization happens much later, around $N = 5,000$ samples rather than $N = 500$, and settles at a noticeably higher value than under the uniform configuration. While the trajectories across these two rows are highly similar, the low-correlation case settles at a slightly higher level than the uncorrelated one. Both rows maintain the minor inverse relationship with the hidden confounder, where a higher true Γ yields marginally lower benchmark values.

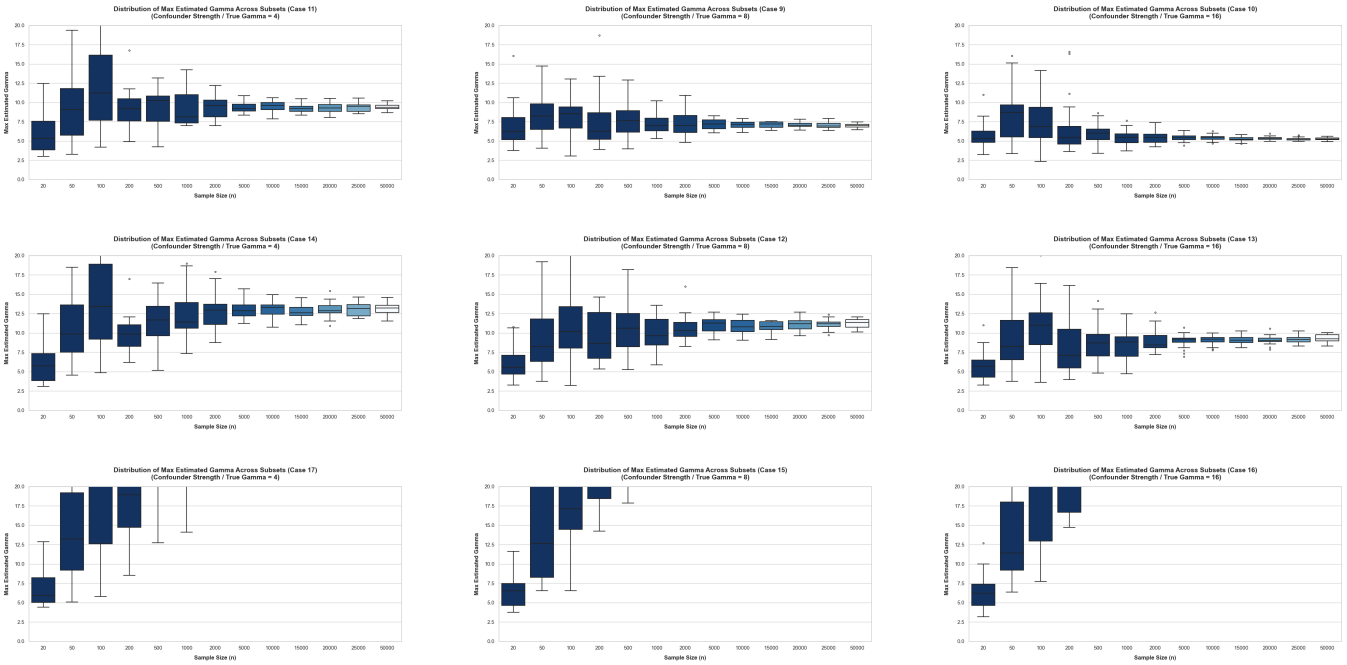
Under high correlation (bottom row), the benchmark exhibits a steep upward growth trend that far surpasses its uniform counterpart. However, just like in the uniform case, the values continue to climb steadily across the entire evaluated sample range, showing no signs of stabilizing even at $N = 50,000$. This row also aligns with the uniform configuration regarding the column trends, where a higher hidden confounder strength (Γ) produces higher benchmark scores. However, unlike the uniform case where the variance stays wide, the boxplot width here narrows significantly at larger sample sizes, becoming tightly bounded around the continuously climbing median from $N = 5,000$ onward.

5 Responsible Research

This section addresses the ethical, data-related, and practical considerations of the project, ensuring adherence to institu-



(a) Uniform feature importance. Rows: correlation strength (uncorrelated, low, high); columns: $\Gamma \in \{4, 8, 16\}$.



(b) Skewed feature importance. Rows: correlation strength (uncorrelated, low, high); columns: $\Gamma \in \{4, 8, 16\}$.

Figure 1: Benchmark bounds across correlation structures and true confounding strengths Γ, under (a) uniform feature importance; (b) skewed feature importance. Each graph shows results across sample sizes $N \in \{20, 50, 100, 200, 500, 1000, 2000, 5000\}$ from 50 Monte Carlo iterations. The dashed diagonal lines represent the true Γ. Rows vary correlation strength (uncorrelated, low, high); columns vary $\Gamma \in \{4, 8, 16\}$. Y-axis is clipped at 20 to allow for a fair side-to-side comparison.

tional standards and transparent scientific reporting.

5.1 Code of Conduct

This project was designed and executed in strict accordance with the TU Delft Code of Conduct. By prioritizing open science, full algorithmic transparency, and honest reporting of model limitations, the research upholds the university's foundational principles for responsible academic practice.

In line with standards of academic integrity, all literature used to contextualize and support this study was obtained from peer-reviewed, credible sources and is accurately cited. The simulation pipeline and benchmark computations were implemented entirely in Python using standard, open-source libraries. No closed-source or proprietary statistical software was used, ensuring that all analytical steps remain fully transparent and auditable.

5.2 Data and Privacy

Because this study relies entirely on synthetically generated data, considerations regarding human subjects and privacy protections do not arise. No real-world participants were involved, meaning no approval from TU Delft's Human Research Ethics Committee was required. The datasets contain no personal data, identifiable metrics, or proxy variables for real people, ensuring zero risk of data leaks or GDPR violations.

5.3 Reproducibility

To ensure all empirical findings are fully verifiable and reproducible, the data-generating mechanisms and experimental setups are detailed across Section 3 and Appendix A. The complete and annotated source code repository is publicly available at https://github.com/majaczerwinska/lmo-informal-benchmark. To guarantee the statistical stability of the reported metrics, all evaluated test scenarios were executed across 50 independent Monte Carlo iterations.

5.4 Bias

Because this project relies entirely on synthetically generated data, it does not inherit nor propagate the historical, systemic, or human-centric biases often embedded in real-world observational datasets. Instead, the primary risks of bias in this study are methodological: specifically, confirmation bias in the simulation design and representational bias in the data generation process.

Confirmation bias would occur if the parameter grid was deliberately engineered to force the benchmark to fail or spike, producing a cleaner critique of the framework. To minimize this risk, the baseline configurations directly replicate the established data-generating processes from the reference literature, and the extensions (the skewed and higher-dimensional regimes) reflect realistic variations in observational data. Naturally, while we aimed for a comprehensive grid, a simulation study can never completely exhaust the entire parameter space, meaning the possibility remains that specific, untested data configurations could deviate from our findings.

Additionally, to counteract representational bias, where conclusions are limited to sterile simulation environments, non-ideal conditions were introduced, such as high covariate-confounder correlation structures and highly asymmetric feature distributions. These inclusions ensure that the evaluation of the benchmark is not skewed toward unrepresentatively cooperative data settings.

5.5 Impact and Risk of Misuse

While the primary objective of this work is to encourage methodological caution, we recognize that by mapping out how changing feature dimensions and importance profiles alter benchmark estimates, our findings could be exploited to strategically shift or inflate results. Specifically, a researcher could misuse these insights to engineer the claimed strength or resilience of their study against hidden confounding, misleading peers or policymakers about its true robustness. To mitigate this risk, we explicitly emphasize: a benchmark value alone should never be taken at face value as definitive proof of a study's resilience. Instead, we advocate for treating informal benchmarks as a comparative tool within a broader sensitivity analysis, rather than a standalone metric to validate a study's conclusions.

5.6 Use of Large Language Models (LLMs)

Large Language Models (LLMs) were utilized as assistive tools during this project. Regarding the report preparation, their application was restricted to polishing text for flow and clarity, as well as LaTeX boilerplate code generation. Within the code base, LLMs were leveraged to assist with secondary programming tasks, including the generation of utility functions for visualization, the enhancement of code documentation, and brainstorming algorithmic refinements for execution speed and numerical stability in edge-case settings.

6 Discussion

In this section, the experimental results from Section 4 are thoroughly analyzed to give insights on what behavior and why to expect from the informal benchmark as the number of samples increases.

6.1 Interpretation of Sample Size Dynamics

Because the Leave-Multiple-Out (LMO) framework computes the informal benchmark by selecting the maximum value across various feature subsets, a standard expectation would be that the benchmark score ($\hat{\Gamma}_{\text{high}}$) should drift upward as the sample size (N) increases. Intuitively, larger datasets increase the probability of capturing outliers from the tails of a distribution. However, our empirical findings reveal that this expected sample-size inflation does not automatically occur. Instead, the trajectory of the benchmark is governed by the interaction between the non-linear mechanics of the odds ratio and the optimization behavior of the underlying propensity score model.

6.1.1 The Uncorrelated and Weak Feature Regime

In settings where the observed covariates $\mathbf{X}$ are uncorrelated with the unobserved confounder U, or where the features possess weak predictive power and the correlation is low, the benchmark exhibits a decreasing trajectory before flattening out as the sample size increases.

This downward trajectory is directly explained by the propensity score model behavior at different sample scales. At small sample sizes, logistic regression is highly susceptible to overfitting due to high sample variance, causing it to erroneously fit random noise and generate extreme propensity scores purely by chance. Because the LMO framework optimizes for the maximum ratio across feature subsets, it actively selects these overfitted anomalies, resulting in artificially inflated benchmark scores when N is low.

However, as the sample size increases, this overfitting behavior vanishes because the propensity score model is fed more data points that it cannot fully explain with the available covariates. Because logistic regression optimizes for global generalization and minimizes its loss function across the entire sample, it responds to this growing uncertainty by assigning conservative, centralized propensity scores. When features are systematically dropped during the LMO routine, the model is forced to retrain under an obstructed view of the world. It adjusts its weights to remain stable, keeping the propensity scores even closer to the sample mean, due to the increase in the amount of uncertainty it needs to account for.

Because the confounder we introduced is strong, it is likely to push the true treatment probabilities toward the extremes to meet the bound introduced through the Marginal Structural Model (Definition 3). However, since $\mathbf{X}$ has no correlation with U in the uncorrelated setting, the observed model cannot recover this pattern. A subtle variation of this constraint occurs in the low-correlation case for weak features: although these covariates benefit slightly from capturing fragments of the hidden confounder, their individual signal is simply too weak to predict the true treatment well and produce higher odds.

Consequently, in both scenarios, the estimator primarily receives an unexplainable variance rather than a clean signal. The higher the true confounding strength, the more prominent this informational gap becomes, forcing the model to remain highly conservative to generalize well. Adding more samples does not drive the score upward; it merely increases the precision of the estimation, tightening the results around a low baseline.

6.1.2 The High Correlation and Strong Feature Regime

The behavior changes fundamentally when features are strong or highly correlated with U. In these settings, the observed covariates serve as an effective proxy for the true treatment assignment mechanism. As the sample size grows, the model does not face unexplained noise; instead, it is given enough data to accurately reconstruct the reality, including its extreme tails. If the true underlying distribution contains individuals with highly lopsided treatment assignments, a larger sample size allows the model to find them, producing higher odds.

In stark contrast to the uncorrelated regime, this creates a direct trend between confounding strength and the benchmark score. Because $\mathbf{X}$ provides a clear glimpse into U, any increase in the confounder's strength means there are genuine, extreme probabilities being observed through $\mathbf{X}$. The model successfully picks up on these signals as N grows, mapping the true distribution more accurately. When the full model correctly assigns an extreme probability, such as 0.98, the non-linear scaling of the odds ratio ($\frac{0.98}{1-0.98} = 49$) completely dominates the calculation by providing a massive numerator. Here, the upward scaling of the benchmark with sample size is a positive attribute: it awards high scores for a model that reflects the highly confounded reality well.

6.2 A Theoretical Warning

While the informal benchmark functions effectively within our simulations, these dynamics expose a critical question regarding what the framework actually rewards. In our data generation process, we assumed the true probabilities spanned a wide range, and the benchmark performed well because it rewarded models that could accurately recover those extremes. However, this could also mean: the framework does not necessarily reward models that mirror reality well; it rewards models that are capable of generating extreme probabilities.

Although testing the exact behavior of the LMO informal benchmark under model misspecification and atypical data distributions falls outside the scope of this study, we raise a theoretical warning for practitioners. Imagine a scenario where the true treatment assignment mechanism is highly conservative, such as a uniform distribution $\mathcal{U}(0.4, 0.6)$. A perfectly specified model operating in this world will correctly output propensity scores clustered tightly around 0.5. Consequently, the full model's odds (Definition 2) will hover near 1.0, providing a very low numerator for the final bound (Definition 3) and probably yielding a low benchmark score. In contrast, if a researcher utilizes a severely misspecified, overparameterized, or uncalibrated model on the same data, the model might erroneously hallucinate extreme propensity scores (e.g., 0.99). Because the LMO framework evaluates study robustness based on the maximum ratio achieved across feature subsets, it could reward this bad model with an inflated score. While the behavior of the denominator introduces an additional layer of complexity that could possibly counteract this, the exact interaction requires further empirical testing. The above example is presented as a word of caution and a potential direction for future research.

6.3 The Comparative Necessity of the Benchmark

The realization that the benchmark scales with extreme probabilities raises a fundamental question: should practitioners inherently desire a high informal benchmark score ($\hat{\Gamma}_{\text{high}}$)? From one perspective, a large benchmark value is traditionally viewed as an indicator of robustness, implying that an unobserved confounder would have to be exceptionally massive to overturn the estimated effect. However, it can be argued that an exceptionally high score simply signals a highly volatile and sensitive data environment where minor adjustments yield drastic shifts in propensity scores.

Ultimately, this conceptual divide reveals a deeper practical truth: an informal benchmark score is entirely uninterpretable in isolation and functions strictly as a comparative baseline. If a study yields a high informal benchmark ($\hat{\Gamma}_{\text{high}}$) but the corresponding sensitivity analysis reveals a low survival threshold (Γ), the causal claim is highly vulnerable.

This specific mismatch ($\hat{\Gamma}_{\text{high}} > \Gamma$) demonstrates that the effect can be neutralized by a hidden confounder that is weaker than the covariates already observed in the model. Because the empirical results have proven that relationships of that magnitude exist within the data, the presence of an equally powerful unobserved confounder becomes highly plausible, rendering the study's conclusions fragile.

7 Conclusions and Future Work

This thesis provided s systematic evaluation of how sample size scaling impacts the behavior of Leave-Multiple-Out (LMO) informal benchmarking. By testing the framework across a diverse simulation grid, this work answers our primary research question: data volume does not inherently inflate or distort benchmark scores. Instead, the trajectory of the benchmark as the sample size grows is entirely determined by whether the observed covariates can successfully mirror the true treatment assignment mechanism.

The empirical results reveal two distinct operational regimes that practitioners must consider. In weak or uncorrelated settings, scaling the sample size acts purely as a variance reducer. It eliminates the severe small-sample overfitting ($N \leq 200$) that causes the LMO routine to capture random noise and output artificially inflated bounds. Once this sampling instability subsides, the lack of predictive signal forces the model to produce conservative probabilities, causing the benchmark scores to stabilize at a low baseline.

In contrast, in strong or highly correlated settings, the benchmark score scales steadily upward alongside the sample size. Crucially, this upward trajectory holds under the condition that the true, underlying propensity scores span a wide range of values. As data volume increases, a well-specified model gains the statistical power necessary to accurately map the extreme tails of the distribution. Because the non-linear mechanics of the odds ratio calculation are exceptionally sensitive to probabilities close to 0 or 1, it is precisely these captured tail cases that drive the final benchmark score upward. Ultimately, the framework functions as intended under scaled conditions: it rewards better-specified models with higher robustness scores, provided the underlying reality features these driving extremes.

7.1 Future Work

Future research can expand on these findings in three specific areas. First, evaluating the LMO routine under deliberate model misspecification and overparameterization will verify whether uncalibrated models that hallucinate extreme probabilities distort the framework to yield artificially high robustness scores. Second, testing data distributions where true treatment probabilities are strictly bounded away from extreme values will isolate how the odds ratio behaves when tail-end outliers are genuinely absent from the population. Finally, expanding this sample-size analysis beyond logistic regression to non-linear machine learning estimators, such as random forests or even neural networks, can establish clearer baseline behaviors for modern machine learning workflows in causal sensitivity analyses.

References

[1] Eduardo Hariton and Joseph J. Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics and Gynaecology*, 125(13):1716, 2018.

[2] Kjell Benson and Arthur J. Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342(25):1878–1886, 2000.

[3] Paul R. Rosenbaum. Sensitivity analysis in observational studies. In Brian S. Everitt and David C. Howell, editors, *Encyclopedia of Statistics in Behavioral Science*, volume 4, pages 1809–1814. John Wiley & Sons, Ltd, 2005.

[4] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 02 2020.

[5] AlexanderM. 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.

[6] Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10999–11009. Curran Associates, Inc., 2020.

[7] Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding, 2022.

[8] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding, 2022.

[9] Xiao-Li Meng. Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

[10] Jean-Baptiste Baitairian, Bernard Sebastien, Rana Jreich, Sandrine Katsahian, and Agathe Guilloux. Sensitivity analysis to unobserved confounders: A comparative review to estimate confounding strength in sensitivity models. *arXiv*, 2025.

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

[12] Guido W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, 93(2):126–132, 2003.

[13] A. McClean, Z. Branson, and E. H. Kennedy. Calibrated sensitivity models. *Biometrika*, 113(2):asag001, 2026.

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

A Mathematical Derivations and Simulation Mechanics

To ensure exact reproducibility, all variables and parameters in the data-generating process are initialized using a pseudo-random seed stream. Each simulation setting uses a fixed base seed, and every subsequent Monte Carlo iteration $m \in \{1, \dots, M\}$ increments this base seed sequentially (i.e., $\text{seed}_m = \text{base\_seed} + m$). This approach guarantees non-overlapping, reproducible pseudo-random streams across all experimental conditions for both covariate generation and parameter sampling, even when iterations are discarded due to numerical safeguards.

A.1 Derivation of Confounder Conditional Variance

To ensure that the unobserved confounder U remains evenly distributed and the propensity scores generated from it are well-behaved, the linear dot product of the observed features must be standardized, and the total variance of the resulting raw confounder U_{raw} must be constant at 1.0. If the variance of U_{raw} shifts with the changes to the correlation parameter λ, passing it through a cumulative distribution function (CDF) distorts the resulting probabilities. A variance that is too high forces the confounder values to saturate the extreme 0 and 1 boundaries, while a variance that is too low compresses all values into the center.

By ensuring the linear combination of observed features is standardized such that $\text{Var}\left(\beta^T \mathbf{X} / \sigma_{\beta^T \mathbf{X}}\right) = 1$, the total variance of the conditional formulation can be established analytically as the sum of the variance of its deterministic mean and its random error component:

$$\begin{aligned} \text{Var}(U_{\text{raw}}) &= \text{Var}\left((1-\lambda)\frac{\beta^T \mathbf{X}}{\sigma_{\beta^T \mathbf{X}}}\right) + \text{Var}(\epsilon) \\ &= (1-\lambda)^2 \cdot \text{Var}\left(\frac{\beta^T \mathbf{X}}{\sigma_{\beta^T \mathbf{X}}}\right) + \left(1 - (1-\lambda)^2\right) \\ &= (1-\lambda)^2 \cdot (1) + 1 - (1-\lambda)^2 \\ &= 1 \end{aligned} \tag{4}$$

Crucially, adjusting the variance of the random error component to $1 - (1-\lambda)^2$ stabilizes the total variance of U_{raw} at exactly 1.0 across all choices of λ. Because the distribution maintains a unit variance and a controlled mean, mapping U_{raw} through a standard normal cumulative distribution function ensures that the final bounded confounder $U \in (0, 1)$ is a perfectly uniform distribution, preserving the integrity of the subsequent true propensity score generation process.

A.2 Simulation Parameterizations and Feature Profiles

The true underlying parameters governing the data-generating process are initialized using a pseudo-random seed stream to ensure perfect reproducibility. Each simulation batch is generated from a sequentially tracked seed.

We map the correlation parameter λ to three distinct target environments:

$$\lambda = \begin{cases} 1.00 & \text{if Uncorrelated} \\ 0.85 & \text{if Low Correlation} \\ 0.45 & \text{if High Correlation} \end{cases} \tag{5}$$

In this framework, λ directly controls the balance between the feature signal and independent noise. Because the total pre-CDF variance is fixed at 1.0, the baseline theoretical Pearson correlation simplifies to $r = 1 - \lambda$, creating three specific testing setups:

Uncorrelated ($\lambda = 1.00, r = 0.00$): The confounder resembles independent random noise.

Low Correlation ($\lambda = 0.85, r = 0.15$): Features contribute a minor fraction (2.25%) of the variance, introducing a subtle confounding signal.

High Correlation ($\lambda = 0.45, r = 0.55$): Features drive a substantial portion (30.25%) of the variance, creating a realistic moderate-to-high confounding environment.

Transforming U_{raw} through the standard normal CDF guarantees that the final bounded confounder $U \in (0, 1)$ is perfectly uniformly distributed. This uniform distribution prevents numerical saturation at the boundaries and ensures stable propensity score generation across all three target environments.

For a baseline environment with $p_{\mathbf{X}}$ features, the parameter vectors are initialized according to the following configurations:

- **Confounder Generation Weights:** The coefficients of the vector $\beta \in \mathbb{R}^{p_{\mathbf{X}}}$ are sampled uniformly such that $\beta_i \sim \mathcal{U}(1.5, 2.0)$.
- **Nominal Propensity Weights:** The coefficients of the feature importance vector $\delta \in \mathbb{R}^{p_{\mathbf{X}}}$ are sampled uniformly such that $\delta_i \sim \mathcal{U}(0.1, 0.5)$.

For the 9 structural scenarios designated as having a "skewed" importance profile (Scenarios 9 through 17), the baseline arrays are overwritten to isolate the effects of feature dominance and uninformative noise:

$$\delta = \begin{bmatrix} 1.5 \\ 0.8 \\ 0.2 \\ 0.05 \\ 0.00 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ \beta_{31} \\ \beta_{41} \\ 0.00 \end{bmatrix} \tag{6}$$

This modification ensures that the fifth covariate (X_4) acts as pure, uninformative noise relative to both nominal treatment assignment and hidden confounding generation.

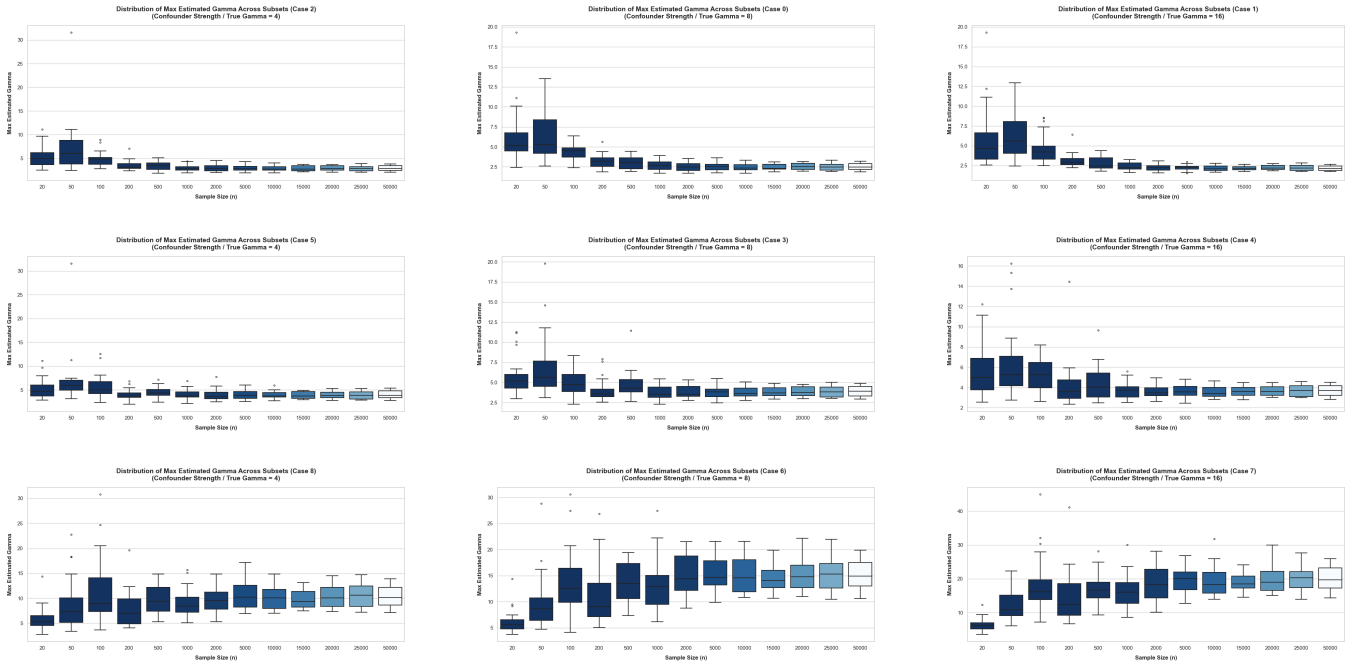
A.3 Numerical Safeguards and Cross-Validation Edge Cases

For small sample sizes ($N \in \{20, 50\}$), stochastic sampling can produce datasets with zero treatment variance, which causes standard logistic regression models to fail. To ensure numerical stability across the Monte Carlo simulation, we implement two programmatic safeguards:

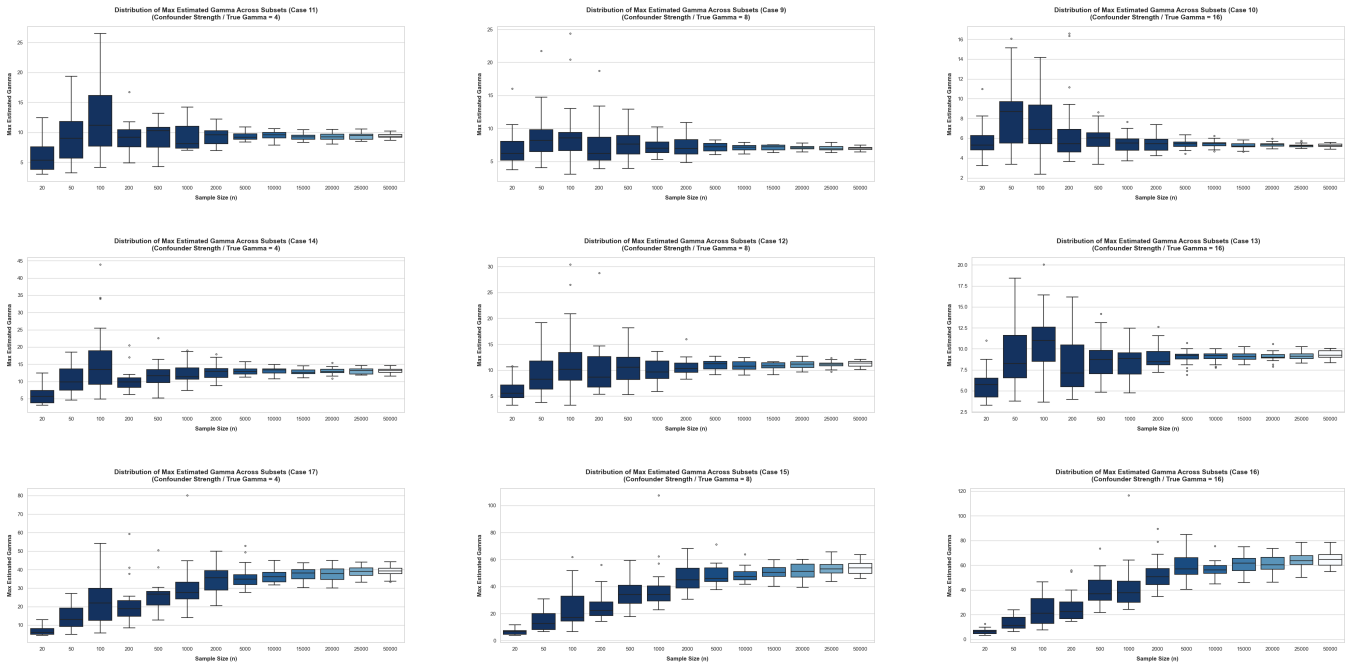
1. **Global Separation Filter:** If a generated sample contains only treatment ($T = 1$) or only control ($T = 0$) units globally, the iteration is discarded. The simulation then advances to a new seed until a valid sample is registered.
2. **Localized Fold Fallback:** During the 5-fold cross-fitting routine, if a specific training fold exhibits zero treatment variance, the propensity score estimator cannot converge. In these edge cases, the model defaults to assigning the global empirical treatment mean as the predicted probability for the validation fold:

$$\hat{e}(\mathbf{X}_{\text{test}}) = \frac{1}{N} \sum_{i=1}^{N} T_i \tag{7}$$

B Graphs with Complete Y-axis



(a) Uniform feature importance. Rows: correlation strength (uncorrelated, low, high); columns: $\Gamma \in \{4, 8, 16\}$.



(b) Skewed feature importance. Rows: correlation strength (uncorrelated, low, high); columns: $\Gamma \in \{4, 8, 16\}$.

Figure 2: Benchmark bounds across correlation structures and true confounding strengths Γ, under (a) uniform feature importance; (b) skewed feature importance. Each graph shows results across sample sizes $N \in \{20, 50, 100, 200, 500, 1000, 2000, 5000\}$ from 50 Monte Carlo iterations. The dashed diagonal lines represent the true Γ. Rows vary correlation strength (uncorrelated, low, high); columns vary $\Gamma \in \{4, 8, 16\}$. Y-axis is not clipped to allow for an analysis of curves.