When the Propensity Model Is Wrong

Informal Benchmarking and a False Sense of Robustness in Causal Sensitivity Analysis

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

Abstract

Causal effect estimates from observational data rely on the assumption that all confounders, variables that influence both treatment and outcome, are observed. Sensitivity analysis with the Marginal Sensitivity Model (MSM) relaxes this assumption through a parameter Γ that bounds how strongly a hidden confounder may distort an individual's probability of treatment, but choosing a realistic value for Γ is difficult. A common solution, Informal Benchmarking (IB), estimates Γ by removing observed covariates from the propensity model (the model of treatment probability) and measuring the resulting shift. Because IB depends entirely on this model, this paper investigates how IB and the resulting sensitivity bounds behave when the propensity model is misspecified. A controlled simulation study isolates a single functional-form error: a non-linear term that is part of the true treatment mechanism is omitted from the fitted model. Even though the benchmark is computed only on covariates that are individually well specified, the omitted term shrinks every fitted coefficient toward zero, and this leakage deflates the benchmark below the value a correctly specified model reports. The result is falsely robust bounds that understate the true risk of hidden confounding, the more dangerous direction of error, and the effect grows with the strength of the omitted term while standard diagnostics give no warning. A simple safeguard is proposed: refit the propensity model with a richer specification and rerun the benchmark, treating any rise in the estimate as evidence that the original was deflated.

1 Introduction

Causal inference moves beyond correlation to measure true cause-and-effect relationships, forming the foundation for reliable conclusions from observational data. In real-world observational studies, causal methods depend heavily on the assumption of ignorability (also known as unconfoundedness): every variable that influences both the treatment (the action whose effect is being measured) and the outcome is assumed to be observed. A variable that influences both is called a confounder. This assumption is rarely realistic, because hidden confounders often go unmeasured (Baitairian et al., 2025). For example, when measuring the effectiveness of a vaccine by comparing average outcomes of vaccinated and unvaccinated patients, results can be heavily biased if an unobserved confounder is present. If people with stronger underlying immune health are both less likely to seek vaccination and naturally more resilient to disease, this unrecorded advantage skews the comparison, producing a biased and unreliable estimate of the vaccine's true effect.

To evaluate the robustness of causal estimates against such unobserved confounders, sensitivity analysis is widely employed. Frameworks like the Marginal Sensitivity Model (MSM) relax the strict ignorability assumption by introducing a sensitivity parameter, Γ, which quantifies the maximum degree to which an unobserved confounder could alter an individual's propensity score (probability of being treated) (Tan, 2006). Rather than a single number, the analysis then reports a range of treatment effects consistent with that level of hidden confounding, and Γ controls how wide that range is. Choosing a realistic value for Γ is difficult: if it is too high, the range becomes impractically wide, and if it is too low, the causal conclusions appear falsely robust (Baitairian et al., 2025; Cinelli and Hazlett, 2020). To resolve this, one of the options researchers frequently use is Informal Benchmarking (IB) (Baitairian et al., 2025). This method treats subsets of observed covariates (the recorded variables describing each individual) as if each were a hypothetical unobserved confounder, estimating how much their removal from the propensity model shifts the estimated odds of treatment, and using the maximum such shift as a benchmark for a plausible strength of hidden confounding. While Informal Benchmarking is widely adopted, the value it returns depends on the propensity model the practitioner chooses to fit (Baitairian et al., 2025). This raises the question of how it behaves when that model is misspecified, that is, when its assumed form does not match how treatment is truly assigned.

This concern is practically significant because misspecification is the norm rather than the exception. Estimating a propensity score requires choosing a model for how covariates map to the probability of treatment, and in practice the dominant choice is logistic regression with main effects only (each covariate entered on its own, with no non-linear terms or interactions between covariates), valued for its simplicity, speed, and interpretability (Austin, 2011). Such a model assumes the covariates combine in a simple, additive way to set the treatment probability. This assumption fails whenever the true mechanism is more complex, for example when a covariate's effect is curved or when two covariates must act together (Setoguchi et al., 2008). More flexible alternatives, such as random forests, gradient boosting, neural networks, and ensemble Super Learners, can absorb this structure and demonstrably reduce misspecification bias, but at the cost of additional tuning, a greater risk of overfitting, and reduced interpretability (Lee et al., 2010; Pirracchio et al., 2015). Critically, the true treatment mechanism is never observed, so practitioners cannot directly verify that their model is correct. Standard balance diagnostics, which check whether treated and untreated groups look similar after adjustment, can flag some failures but cannot certify correct specification (Zhang et al., 2019). Misspecification is therefore common, hard to rule out, and directly relevant to anyone

applying Informal Benchmarking, which motivates this study of how it distorts the benchmark.

This paper addresses this gap by investigating the following primary research question: *What is the interaction between the parameter from informal benchmarking and the bound from sensitivity analysis if the propensity model is incorrect?* The central concern is the dangerous direction of error: whether a misspecified propensity model can lead Informal Benchmarking to *under*-estimate confounding strength, reporting falsely robust bounds that make a causal conclusion look more certain than the evidence supports.

To answer this question, the main contributions of this work are as follows:

- We show that omitting a non-linear (quadratic) term from a parametric propensity model biases $\hat{\Gamma}_{IB}$ downward, producing falsely robust sensitivity bounds that understate the true confounding risk. This is the more dangerous failure mode, since it makes hidden confounding appear less threatening than it is.
- We trace the mechanism behind this bias. Even when Informal Benchmarking is applied only to covariates that are individually well specified, the omitted term shrinks their fitted coefficients toward zero through the non-collapsibility of logistic regression, and the same misspecified model contaminates every odds ratio the benchmark reads. The premise that Informal Benchmarking requires can hold for each benchmarked covariate, and the estimate can still be deflated.
- We provide a practical, single-check procedure for practitioners who suspect misspecification: refit the propensity model with a richer cross-fitted specification and rerun the benchmark, treating a rise in $\hat{\Gamma}_{IB}$ as a sign that the original estimate was deflated.

The remainder of this paper is organized as follows. Section 2 contextualizes the research by reviewing necessary background concepts and related literature. Section 3 formulates the methodology and the simulated data-generating processes used to evaluate model misspecification. Section 4 presents the core experimental results. Section 5 discusses the implications and limitations of these findings. Section 6 concludes the paper. Finally, Section 7 reflects on the responsible research aspects of this work, including reproducibility.

2 Background and Related Work

Causal inference aims to estimate the true effect of a treatment on an outcome from observational data, rather than relying on statistical associations alone. A standard approach for this is Inverse Probability Weighting (IPW), which re-weights observations by the inverse of the propensity score. The propensity score is defined as the probability of receiving treatment given observed covariates, $e(X) := \mathbb{P}(T = 1 \mid X)$ (Rosenbaum and Rubin, 1983). The validity of IPW rests on the *ignorability* assumption: conditionally on observed covariates X, treatment assignment is independent of potential outcomes (Rosenbaum and Rubin, 1983). In observational studies, this assumption is frequently violated because unobserved confounders are absent from the data (Baitairian et al., 2025). When ignorability fails, IPW estimates of the Average Treatment Effect (ATE) are biased, and the direction and magnitude of that bias are generally unknown.

To handle potential ignorability violations, the Marginal Sensitivity Model (MSM) (Tan, 2006) provides a formal framework. The MSM assumes that, for any two individuals with identical observed covariates, the odds of treatment can differ by at most a factor of $\Gamma \geq 1$ due to unobserved confounding. Formally, letting $e(X, U) := \mathbb{P}(T = 1 \mid X, U)$ denote the true propensity score that also depends on the unobserved confounder U, the MSM imposes the constraint

$$\Gamma^{-1} \;\leq\; \frac{e(X,U)\,/\,(1-e(X,U))}{e(X)\,/\,(1-e(X))} \;\leq\; \Gamma. \tag{1}$$

Under this constraint, rather than a single point estimate of the ATE, a sensitivity interval $[\hat{\theta}^{-}(\Gamma),\, \hat{\theta}^{+}(\Gamma)]$ is obtained, whose width grows monotonically with Γ (Dorn and Guo, 2023). A value of $\Gamma = 1$ recovers the standard ignorability assumption, while larger values allow for progressively stronger hidden confounding.

Selecting an appropriate Γ is challenging without domain knowledge. A value that is too large yields uninformatively wide bounds, while a value that is too small produces falsely robust conclusions (Baitairian et al., 2025; Cinelli and Hazlett, 2020). Informal Benchmarking addresses this by treating each observed covariate as a hypothetical unobserved confounder (Baitairian et al., 2025). In the leave-one-out variant, each covariate X^i is treated as if it were unobserved. The odds ratio between the full-covariate propensity score $\hat{e}(X)$ and the reduced-covariate score $\hat{e}(X^{(-i)})$ is computed for every individual, and the maximum such ratio defines a per-covariate sensitivity estimate $\hat{\Gamma}_i$ (Baitairian et al., 2025). The overall benchmarking estimate is then $\hat{\Gamma}_{IB} := \max_i \hat{\Gamma}_i$. A leave-multiple-out extension considers subsets of covariates simultaneously, providing sensitivity estimates for combinations that may jointly represent a multi-dimensional hidden confounder. To avoid overfitting, propensity scores in both variants are estimated via K-fold cross-fitting (Baitairian et al., 2025).

Each of these steps (IPW, the MSM's nominal propensity score $e(X)$, and the leave-one-out scores behind Informal Benchmarking) relies on a propensity score that is never known in practice and must be estimated by fitting a model to the data. This makes the choice of model consequential: when the fitted model does not match the true treatment mechanism, whether by assuming the wrong functional form or by leaving out a relevant covariate, it is misspecified and the estimated propensity scores are systematically wrong. Such misspecification is a known threat to causal estimation: a

model that omits a relevant term, be it a non-linear effect or a correlated covariate, produces biased weights and, in turn, a biased IPW point estimate of the ATE (Setoguchi et al., 2008). Because Informal Benchmarking reads $\hat{\Gamma}_{IB}$ from these same estimated propensity scores, a model error can show up in the benchmark and be mistaken for the signal of a hidden confounder.

Despite this, the robustness of Informal Benchmarking to propensity model misspecification has not been systematically studied. Baitairian et al. (2025) note that $\hat{\Gamma}_{IB}$ is sensitive to the choice of propensity estimator and recommend a well-specified model as a prerequisite, but do not characterize the direction or magnitude of bias that arises when specification fails. Work on omitted variable bias has addressed the risks of informal benchmarking in adjacent settings, demonstrating that it can yield misleading and falsely robust conclusions (Cinelli and Hazlett, 2020). That critique, however, does not target the specific interaction between propensity model misspecification and the MSM sensitivity parameter. This gap motivates the simulation study presented in the following sections.

3 Methodology

Evaluating the robustness of Informal Benchmarking under propensity model misspecification requires a controlled environment in which the relationship between covariates, confounders, and treatment assignment is known exactly. A simulation study built on synthetic Data Generating Processes (DGPs) is therefore conducted. Synthetic data make it possible to prescribe the true propensity score, inject a single isolated form of model error, and measure how the resulting estimate $\hat{\Gamma}_{IB}$ deviates from the estimate obtained under a correctly specified reference model, without the noise of real-world data.

The danger studied here is deflation. If misspecification deflates $\hat{\Gamma}_{IB}$, the practitioner is handed a value below the truth and reports bounds that are too narrow, the more dangerous case, since the result then looks more robust to unobserved confounding than it is. A smaller $\hat{\Gamma}_{IB}$ yields a narrower sensitivity interval, so this study isolates whether and how a single functional-form error pushes the benchmark down.

The experiment introduces *functional-form* misspecification, where a non-linear term present in the true propensity score is omitted by the fitted model. The reason to expect deflation is specific to logistic regression: omitting a term that drives treatment leaves its variation unexplained, and this leftover variation does not stay local to the affected covariate but shrinks every fitted coefficient toward zero, a consequence of the odds ratio being non-collapsible (Greenland et al., 1999; Mood, 2010). This same misspecified model is used at every step of Informal Benchmarking, both the full propensity fit and each leave-one-out refit, so the odds-ratio computation at the heart of the method is contaminated even when the benchmarked covariates are individually well specified. Because Informal Benchmarking reads confounding strength off the size of these coefficients, this shrinkage should lower $\hat{\Gamma}_{IB}$, producing falsely robust bounds. The experiment that follows tests whether this hypothesis holds.

3.1 Shared Data Generating Process

The simulation draws $N = 5000$ individuals, and every configuration is repeated over $M = 100$ independent Monte Carlo trials to quantify sampling variability. All results below report the mean and standard deviation across these trials, and all randomness is seeded for reproducibility. Each individual is assigned $p_X = 5$ observed covariates $X = [X_1, \ldots, X_5]^\top$ and $p_U = 2$ unobserved confounders $U = [U_1, U_2]^\top$. The observed covariates are drawn independently as $X_i \sim \mathcal{U}(-1, 1)$. Bounding the covariates in this way keeps the linear predictor inside the logistic function in a moderate range, so the propensity scores stay away from 0 and 1. This preserves overlap (the positivity assumption): every individual keeps a non-trivial chance of being either treated or untreated, which stops the inverse-propensity weights, which divide by the propensity, from becoming unstable.

The unobserved confounders are generated conditionally on X so that they are correlated with the observed covariates, reproducing realistic, interdependent confounding. Following Baitairian et al. (2025), each confounder is drawn as

$$U_j \mid X \;\sim\; \mathcal{N}\left((1-\lambda)\,\rho \sum_{k=1}^{p_X} X_k,\; \lambda^2 \right), \qquad j \in \{1, 2\}, \tag{2}$$

Here each confounder is a noisy linear function of the observed covariates. The strength of the link between X and U is set by how the signal coefficient $(1-\lambda)\,\rho$ on the mean compares with the noise variance λ^2: as this ratio grows the confounder becomes more determined by X, and as it shrinks the confounder approaches pure noise. The chosen values $\lambda = 0.6$ and $\rho = 1.75$ give each confounder a correlation of about 0.37 with each observed covariate, a moderate dependence that is neither negligible nor close to deterministic. This design deviates from Baitairian et al. (2025) in two ways: the coupling is fixed at a constant ρ rather than sampled per confounder, and U enters the propensity logit directly and additively, rather than through their more elaborate construction that fixes the true confounding strength in advance. Given a true propensity score $e(X, U)$, the binary treatment is drawn as

$$T \mid X, U \;\sim\; \text{Bernoulli}\big(e(X, U)\big). \tag{3}$$

Each experiment specifies its own true propensity score $e(X, U)$ below, in each case a logistic function of the covariates and confounders.

3.2 Estimation and Outcome Measures

Throughout, $\hat{\Gamma}_{IB}$ is estimated using the leave-one-out procedure of Algorithm 1, with all propensity scores obtained by 5-fold cross-fitted logistic regression to avoid

Algorithm 1 Informal Benchmarking (leave-one-out, binary treatment).

Require: Observed confounders $\{X_i\}_{i=1}^n = \{(X_i^{(1)}, \ldots, X_i^{(p_X)})\}_{i=1}^n$ and treatments $\{T_i\}_{i=1}^n$, with $T_i \in \{0,1\}$; a cross-fitted estimate $\hat{e}$ of the propensity score.

for $i \in \{1, \ldots, p_X\}$ **do**
  for $j \in \{1, \ldots, n\}$ **do**
    Compute $\hat{r}_{i,j} := \mathrm{OR}\big(\hat{e}(X_j),\, \hat{e}(X_j^{(-i)})\big)$
  end for
  Compute $\hat{\Gamma}_i^{+} := \max_j(\hat{r}_{i,j})$ and $\hat{\Gamma}_i^{-} := 1/\min_j(\hat{r}_{i,j})$
  Compute $\hat{\Gamma}_i := \max(\hat{\Gamma}_i^{+}, \hat{\Gamma}_i^{-})$
end for
Compute $\hat{\Gamma}_{\text{low}} := \min_i(\hat{\Gamma}_i)$ and $\hat{\Gamma}_{\text{high}} := \max_i(\hat{\Gamma}_i)$
return $\hat{\Gamma}_{IB} := \hat{\Gamma}_{\text{high}}$

overfitting bias (Baitairian et al., 2025). Logistic regression is used rather than a non-parametric estimator because the object of study is parametric model misspecification. Informal Benchmarking is run twice on the same data: once with a *correctly specified* estimator and once with a *misspecified* estimator. When the propensity model is correct, Informal Benchmarking is intended to return a conservative stand-in for Γ, so the correctly specified run represents the method working as designed. The misspecified run then tests whether that intended behaviour survives a single functional-form error. The correctly specified run serves as the reference, and the quantity of interest is the gap between the two $\hat{\Gamma}_{IB}$ curves as the misspecification strength α grows. This gap is isolated deliberately. Misspecification can distort two separate things: the nominal effect estimate that the bounds are centered on, and the sensitivity parameter $\hat{\Gamma}_{IB}$ that sets how far the bounds extend. Only the latter is studied here. A smaller $\hat{\Gamma}_{IB}$ yields narrower bounds and a larger one wider bounds, so the sign of the gap decides whether the bounds become falsely narrow and over-confident or merely over-conservative.

3.3 Functional-Form Misspecification: Quadratic Omission

This experiment evaluates functional-form misspecification. Intuitively, the error is easy to picture: the true probability of treatment bends with X_1, tracing a U-shape, but a strictly linear model can only fit a straight line through that bend, so it flattens the curve and treats the leftover variation as noise. Because Informal Benchmarking reads its odds ratios off this flattened fit, the treatment-probability swings it measures are compressed before any covariate is set aside. The treatment probability depends on a non-linear (quadratic) transformation of X_1, so that the true propensity score is

$$e(X,U) = \mathrm{logistic}\Big(\beta_X^\top X + \alpha\big(X_1^2 - \tfrac{1}{3}\big) + \beta_U^\top U\Big), \qquad \beta_X = [0.3]_5^\top, \beta_U = [0.3]_2^\top. \tag{4}$$

The constant $\frac{1}{3} = \mathbb{E}[X_1^2]$ centers the quadratic term so that it has mean zero, adding curvature to the propensity without shifting its average level. The observed-covariate and confounder coefficients are set equal ($\beta_X = \beta_U = 0.3$) so that each observed covariate is as strong a driver of treatment as the hidden confounder it serves to benchmark. The precise value is not important, only that it stays moderate enough to preserve overlap. The parameter α is swept over 31 equally spaced values in $[0.0, 3.0]$ to scale the severity of the misspecification, the special case $\alpha = 0$ recovers a perfectly specified DGP and provides the baseline. Two estimators are compared:

- **Correctly specified:** logistic regression augmented with the squared term,
$$\hat{e}(X) = \mathrm{logistic}(\hat{\beta}_X^\top X + \hat{\alpha} X_1^2).$$
- **Misspecified:** strictly linear logistic regression that omits the squared term,
$$\hat{e}(X) = \mathrm{logistic}(\hat{\beta}_X^\top X).$$

Normally, dropping a term biases the coefficients of the variables correlated with it (omitted-variable bias). That does not happen here: because X_1 is centered at zero, X_1 and X_1^2 are uncorrelated, so omitting X_1^2 does not bias X_1's coefficient directly. The unmodelled curvature instead acts as unexplained variation in the treatment model. In a linear model such leftover variation would leave the other coefficients untouched, but in a logistic model it shrinks all of them toward zero, the non-collapsibility property noted above. This shrinkage is how a local functional-form error contaminates the whole fitted model.

The IB leave-one-out set is restricted to $\{X_2, \ldots, X_5\}$. Benchmarking removes one covariate at a time and records how much the fitted propensity moves, and reads that movement as how influential a hidden confounder of comparable strength could be. It never references the true U. The covariate X_1 is kept out of this set because it carries the functional-form error, so benchmarking on it would mix the misspecification into the measurement itself. The remaining covariates enter the truth linearly and are mutually independent, so each leave-one-out swing reflects that covariate's own influence, and any change between the two runs is due to

the shrinkage alone, not to a change in those covariates' functional form. The misspecified run thus violates Informal Benchmarking's premise of a correctly specified propensity globally, while keeping the benchmarked covariates individually well specified. The experiment asks whether this drives $\hat{\Gamma}_{IB}$ downward consistently, meaning a gap that persists in the Monte Carlo mean and grows with α rather than mere sampling noise.

4 Experimental Results

This section evaluates how propensity model misspecification distorts the Informal Benchmarking estimate $\hat{\Gamma}_{IB}$. The leave-one-out procedure of Algorithm 1 is run twice on the same simulated data, once with a correctly specified estimator and once with a misspecified one, and both estimates are tracked as the misspecification strength α increases. The correctly specified run is the reference, and the gap between the two curves is the quantity of interest. Every curve reports the mean over $M = 100$ Monte Carlo trials, and the shaded bands show ± 1 standard deviation. At $\alpha = 0$ the two estimators coincide by construction, so any divergence that appears for $\alpha > 0$ is caused by the misspecification alone.

4.1 Functional-Form Misspecification

Omitting a quadratic term deflates the benchmark. Figure 1 plots the two $\hat{\Gamma}_{IB}$ curves as the strength α of the quadratic term in the true propensity score (Equation (4)) grows. At $\alpha = 0$ both estimators agree at $\hat{\Gamma}_{IB} \approx 2.38$, confirming that they behave identically when the linear model is correct. As α increases the curves separate: the correctly specified estimator stays essentially flat, rising slightly to 2.47, while the misspecified estimator falls steadily to 2.11 at $\alpha = 3$. A smaller $\hat{\Gamma}_{IB}$ means a narrower sensitivity interval (Section 2), so this downward bias makes the result look more robust to hidden confounding than it really is. Table 1 summarizes the benchmark and the coefficient attenuation behind it at representative misspecification strengths α.

Table 1: Benchmark $\hat{\Gamma}_{IB}$ under the misspecified (linear) and correctly specified (quadratic) models, and the mean coefficient magnitude $|\hat{\beta}|$ of the benchmark covariates under the misspecified model, at representative α. Values are mean (s.d.) over $M = 100$ trials.

α	$\hat{\Gamma}_{IB}$ Miss.	$\hat{\Gamma}_{IB}$ Corr.	$\lvert\hat{\beta}\rvert$
0	2.38 (0.12)	2.39 (0.13)	0.71 (0.03)
1	2.35 (0.13)	2.41 (0.14)	0.69 (0.03)
2	2.24 (0.11)	2.45 (0.13)	0.66 (0.03)
3	2.11 (0.11)	2.47 (0.14)	0.61 (0.03)

This deflation is driven by coefficient attenuation. A strictly linear model cannot represent the U-shaped contribution of X_1, so it treats that curvature as unexplained noise and lowers its predictions toward 0.5 to avoid confident but wrong probabilities. Figure 2 tracks the average magnitude $|\hat{\beta}|$ of the linear coefficients for the benchmark covariates $X_2, \ldots, X_5$. It shrinks steadily from 0.71 at $\alpha = 0$ to 0.61 at $\alpha = 3$. Informal Benchmarking judges a covariate's confounding strength by how much the predictions move when that covariate is removed. When every coefficient is artificially shrunk, removing a covariate moves the predictions less, and the benchmark reads this as weaker confounding. The global model error thus leaks into $\hat{\Gamma}_{IB}$ even though it is measured only on the correctly specified covariates $X_2, \ldots, X_5$.

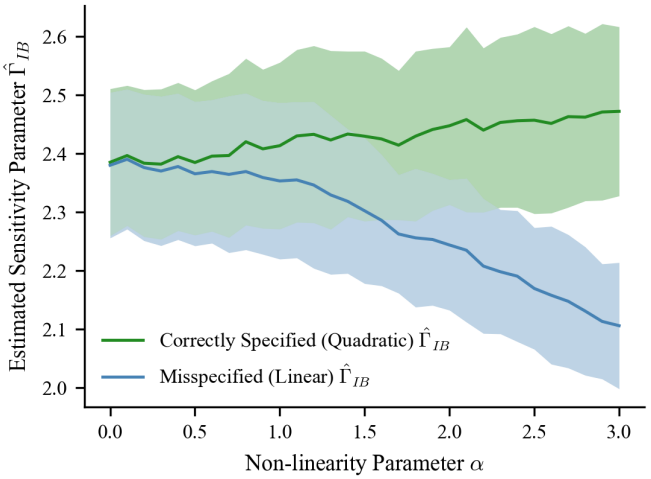

Figure 1: Functional-form misspecification. Estimated sensitivity parameter $\hat{\Gamma}_{IB}$ as the strength α of the omitted quadratic term grows, for the correctly specified (quadratic) and misspecified (linear) propensity models. Bands: ± 1 s.d. The estimators agree at $\alpha = 0$. As the misspecification grows, the misspecified benchmark deflates while the reference stays flat.

The same attenuation is visible directly in the predicted propensity scores. Figure 3 shows their density at $\alpha = 3$ under both estimators. The misspecified model concentrates its predictions near 0.5, while the correctly specified model spreads them toward 0 and 1. Because Informal Benchmarking reads confounding strength off how far removing a covariate moves these scores, predictions packed around 0.5 leave less room to move, confirming how the deflation of $\hat{\Gamma}_{IB}$ arises.

5 Discussion

The experiment answers the primary research question. When the propensity model is functionally misspecified, $\hat{\Gamma}_{IB}$ no longer measures confounding strength alone: it absorbs part of the model error and carries it into the sensitivity bounds. A single omitted non-linear term is enough to deflate the benchmark, and because a smaller $\hat{\Gamma}_{IB}$ narrows the sensitivity interval, the distortion runs in the dangerous direction, making a causal conclusion look more robust to hidden confounding than it is.

The deflation is dangerous above all because it is

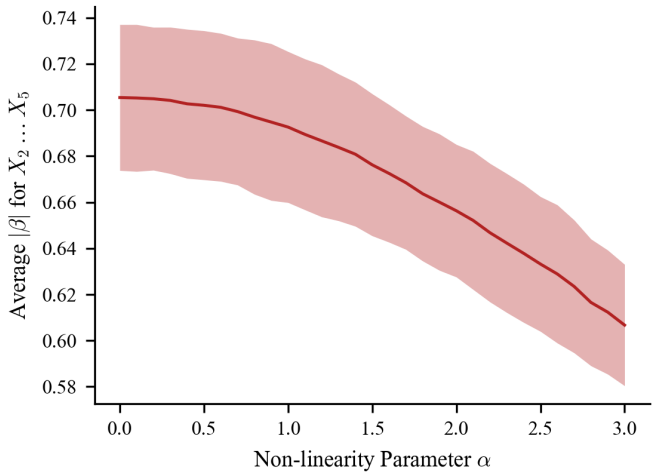

Figure 2: Functional-form mechanism. Average magnitude $|\hat{\beta}|$ of the linear coefficients of the benchmark covariates $X_2, \ldots, X_5$ under the misspecified model, against the omitted quadratic strength α. The coefficients shrink as α grows. This attenuation is what deflates $\hat{\Gamma}_{IB}$.

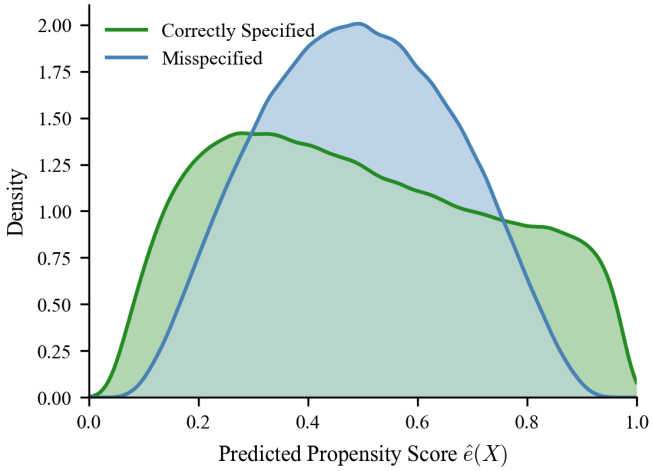

Figure 3: Functional-form mechanism, confirmation. Density of the predicted propensity scores $\hat{e}(X)$ under the correctly specified (quadratic) and misspecified (linear) estimators, at omitted-quadratic strength $\alpha = 3.0$. The misspecified model concentrates its predictions near 0.5, while the correctly specified model spreads toward 0 and 1. This concentration is the visible counterpart of the coefficient attenuation that deflates $\hat{\Gamma}_{IB}$.

silent. The benchmark was computed only on covariates whose functional form was correct, yet the global model error still leaked into their odds ratios through coefficient attenuation. Attenuation under unexplained variation is a known property of logistic regression (Mood, 2010), but here it has a consequence specific to sensitivity benchmarking: restricting Informal Benchmarking to seemingly well-modelled covariates does not protect it. Worse, the procedure gives no warning. The misspecified model still converges and produces plausible-looking propensity scores, and balance diagnostics cannot certify that the specification is correct (Zhang et al., 2019). A practitioner reading the deflated benchmark would conclude the causal estimate is robust precisely when it is not.

In a real setting a practitioner would benchmark on every covariate including X_1. A sanity check shows this changes little: at $\alpha = 3$, benchmarking on X_1 gives $\hat{\Gamma}_1 \approx 1.91$, close to the other covariates, and $\hat{\Gamma}_{IB}$ is essentially unchanged whether X_1 is included (2.12) or excluded (2.11). Excluding X_1 here is therefore a choice to keep the measurement clean, not a condition the deflation depends on.

Beyond the specifics of this experiment, the results speak to a general feature of sensitivity analysis. Informal Benchmarking is one instance of a broader strategy: calibrating the unknown strength of unobserved confounding from the observed covariates. That same strategy underlies omitted-variable-bias sensitivity analysis in linear regression, the setting in which Cinelli and Hazlett (2020) formalized it and cautioned that such benchmarks can mislead when an observed covariate is used to estimate the strength of an unobserved one. The present results show that the caution carries into the Marginal Sensitivity Model: because the benchmark is read off a fitted propensity model, it inherits that model's misspecification, now expressed directly in the sensitivity parameter Γ. The lesson is that $\hat{\Gamma}_{IB}$ should be treated as a model-dependent quantity rather than a robust summary of the data, whatever sensitivity framework it feeds.

Functional-form error is not the only way a propensity model can be misspecified. To probe a structurally different error, Appendix A reports a second experiment that drops a relevant covariate correlated with a retained one. It is kept out of the main text because its mechanism differs, the retained covariate acts as a proxy for the dropped one (the classic omitted-variable bias), and its effect on the benchmark is non-monotone, over-estimating confounding relative to a correctly specified model at weak-to-moderate strength before under-reporting it at high strength. This complementary case reinforces the central lesson that $\hat{\Gamma}_{IB}$ inherits whatever the fitted model gets wrong, while showing that the direction of the distortion is not fixed.

5.1 Practical Implications

These findings sharpen the recommendation of Baitairian et al. (2025) that a well-specified propensity model is a prerequisite for Informal Benchmarking, since a misspecified model fails quietly and toward false robustness. Correct specification cannot be verified from data alone, but misspecification still leaves a measurable trace: a richer model that fits the held-out data better reveals that the simpler one left structure unexplained, and Figure 4 turns this comparison into a short procedure. Relying on a held-out fit comparison is deliberate: of the diagnostics a practitioner might reach for, it is the one that exposes this failure. Checks

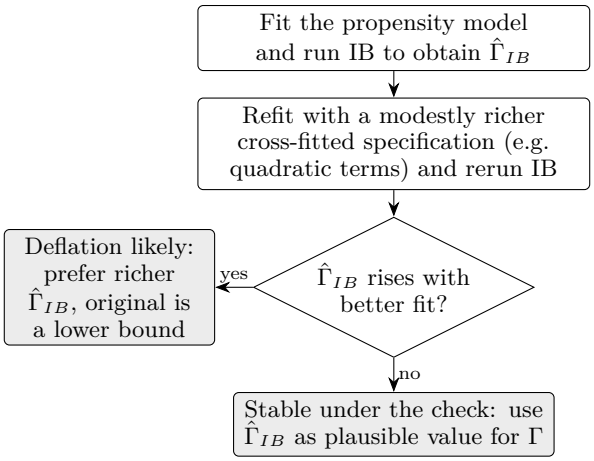

Figure 4: Decision flowchart for applying Informal Benchmarking when propensity model misspecification is suspected. The single check probes the attenuation failure mode: refit with a richer specification and see whether $\hat{\Gamma}_{IB}$ rises. Gray boxes are terminal recommendations.

that look for instability instead, such as the variance of estimates across folds or resamples, do not flag it, because a misspecified linear model fails through bias rather than variance and so stays stable. The check is cheap: refit the propensity model with a modestly richer, still cross-fitted specification and rerun the benchmark. If $\hat{\Gamma}_{IB}$ rises while the held-out fit improves, the original was likely deflated and should be treated as a lower bound, with cross-fitting guarding against mistaking overfitting for recovered structure (Baitairian et al., 2025). Only when the benchmark is stable under a richer specification should it serve as a plausible value for Γ. More broadly, because the benchmark inherits whatever the propensity model gets wrong, a practitioner must be confident that the model reflects the true treatment mechanism before trusting $\hat{\Gamma}_{IB}$.

5.2 Limitations and Future Work

Several limitations bound the scope of these conclusions. All results come from synthetic data with five observed covariates, two confounders, a binary treatment, and logistic treatment mechanisms, so the reported effect sizes are specific to these designs. The result is an existence proof of the failure mode, not an estimate of its typical magnitude in practice. Only logistic estimators were studied, a deliberate choice that isolates parametric misspecification but leaves the behaviour of flexible estimators unexamined. The experiment also injects a single, isolated error, whereas real propensity models plausibly combine functional-form and structural errors whose interaction was not studied. Finally, the reference is a correctly specified model rather than the ground truth. The true Γ of these DGPs is not available in closed form and was not computed here, so the experiment quantifies the relative distortion caused by misspecification rather than the absolute accuracy of Informal Benchmarking itself.

These limitations map directly onto future work. The most pressing extension is to flexible estimators such as random forests. There misspecification takes a different form: instead of omitting a known term, a flexible model can under-fit by smoothing real structure away, for example a forest too shallow to capture the curvature studied here. Comparing an under-capacity estimator against a well-tuned one would test whether the same attenuation arises when curvature is smoothed rather than dropped. It would also generalize the practical check, since enriching the model then means adding capacity such as greater depth, so the instability signal extends naturally to non-parametric estimators. Omitted interaction terms, and combinations of functional-form and structural error, would show whether the two mechanisms compound or partially cancel. The leave-multiple-out variant, where covariate subsets are omitted jointly, is a natural extension. Because the covariates benchmarked here are independent, jointly omitting them mainly enlarges the measured swings without changing the underlying attenuation, so the deflation is expected to persist. Whether correlated subsets behave differently remains open. Computing the true Γ of these DGPs numerically, or adopting a construction in which it is known in closed form, would upgrade the comparison from relative distortion to absolute calibration. Finally, semi-synthetic studies built on real covariate distributions would indicate how large these biases are in realistic settings.

6 Conclusion

This paper asked whether the sensitivity parameter estimated by Informal Benchmarking can still be trusted when the propensity model is misspecified. The answer is that it cannot be taken at face value. Under a functional-form error $\hat{\Gamma}_{IB}$ stops measuring confounding strength alone: part of the model error leaks into it and into the Marginal Sensitivity Model bounds it feeds, deflating the estimate. The same misspecified model drives every fit behind the benchmark, so the leakage persists even when the benchmark is restricted to covariates that are individually well specified.

The error is dangerous because it is silent. A model that looks correctly specified, and a benchmark restricted to covariates whose form looks correct, can still deflate $\hat{\Gamma}_{IB}$ and make the bounds look robust exactly when they should not be trusted. We therefore conclude that $\hat{\Gamma}_{IB}$ should be reported as a model-dependent quantity rather than a robust property of the data, that a practitioner must be confident the propensity model reflects the true treatment mechanism before relying on the benchmark, and that the one safeguard they can realistically apply is to test its stability under a richer specification (Figure 4). This result is an existence proof of the failure mode on synthetic, logistic designs. Its typical magnitude, and whether flexible estimators behave the same way, is left to future work.

7 Responsible Research

Reproducibility is the central concern for a simulation study, and every result here is made exactly repeatable. Every parameter of the data generating processes is stated explicitly in Section 3: the sample size, the number of Monte Carlo trials, the confounder parameters, the coefficient vectors, the correlation structure of the supplementary experiment (Appendix A), and the α sweeps of both experiments. All randomness is controlled by fixed seeds, so every figure and number in Section 4 can be regenerated exactly. The full implementation, covering the data generating processes, the estimators, the benchmarking procedure of Algorithm 1, and the scripts that produce each figure, is publicly available.[1]

The study involves no human subjects and no personal data. All data are synthetic, so no privacy or consent concerns arise.

Care has been taken not to overstate the originality of this work. The data generating process adapts Baitairian et al. (2025) with the deviations flagged in Section 3, and Informal Benchmarking is a widely used practice rather than a contribution of this paper. The contribution is the controlled study of its behaviour under misspecification, not the method itself.

The ethical motivation of this work is the harm that misleading sensitivity estimates can cause. Sensitivity analysis is used to defend causal conclusions in high-stakes domains such as medicine, and a deflated benchmark gives false confidence in an estimate that hidden confounding may have distorted (Section 1). Honestly characterizing the failure modes of Informal Benchmarking is therefore the responsible aim of this study, not an argument against using it, and the flowchart of Figure 4 turns the findings into constructive guidance. For the same reason, Section 5 states the limitations of the design explicitly, so that the findings are not generalised beyond what the experiments support.

7.1 The Use of Generative AI

Two generative AI tools were used in this project. Claude Code (Anthropic) assisted with the implementation of the codebase and with the writing of this report, where it was used to rephrase text, restructure sections, and assist with LaTeX formatting. Gemini (Google) was used to help with the understanding of technical literature and to brainstorm ideas. All AI-assisted code and text was reviewed and verified by the author, who retains full responsibility for the content of this work.

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.

Baitairian, J.-B., Sebastien, B., Jreich, R., Katsahian, S., and Guilloux, A. (2025). Sensitivity analysis to unobserved confounders: A comparative review to estimate confounding strength in sensitivity models. *arXiv preprint arXiv:2510.16560*.

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

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

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, 26(1):67–82.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Pirracchio, R., Petersen, M. L., and Van Der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology*, 181(2):108–119.

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.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.

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

Zhang, Z., Kim, H. J., Lonjon, G., Zhu, Y., et al. (2019). Balance diagnostics after propensity score matching. *Annals of translational medicine*, 7(1):16.

A Structural Misspecification: Correlated Omission

As a supplement to the functional-form study of the main text, this experiment probes a different error, dropping a covariate correlated with a retained one, and shows that the benchmark can also be biased upward. It reuses the data-generating process of Section 3.1 with two changes. First, X_1 and X_2 are coupled through a Gaussian copula (Nelsen, 2006) with latent correlation $\rho = 0.5$ (rank correlation ≈ 0.48), so X_1 can act as a

[1] https://github.com/rolandVi/causal-sensitivity-benchmarking

proxy for X_2. The other covariates and the confounders stay independent. Second, the true propensity score is

$$e(X,U) = \text{logistic}\Big(\beta_X^\top X + \alpha X_2 + \beta_U^\top U\Big), \quad \beta_X = [1.0,\, 0,\, 0.3,\, 0.3,\, 0.3]^\top, \quad \beta_U = [0.3]_2^\top, \tag{5}$$

so X_1 is a dominant predictor while X_2's entire effect is carried by the swept term αX_2. The parameter α is swept over $[0.0, 2.0]$, and a correctly specified model (all five covariates) is compared with a misspecified one that drops X_2 entirely, each running IB over its own covariates. Because X_1 and X_2 are correlated, dropping X_2 forces X_1 to absorb part of its effect, the classic omitted-variable-bias mechanism (Cinelli and Hazlett, 2020).

Dropping the correlated covariate biases the benchmark in two phases (Figure 5). At weak-to-moderate strength the misspecified estimator sits above the reference, for example $\hat{\Gamma}_{IB} \approx 9.2$ against 6.6 at $\alpha = 1$, an over-estimate that yields over-conservative bounds. The curves cross near $\alpha \approx 1.4$. Beyond it the correctly specified estimator climbs to 20.1 at $\alpha = 2$ while the misspecified one plateaus near 10.2, under-reporting the reference and producing falsely robust bounds again.

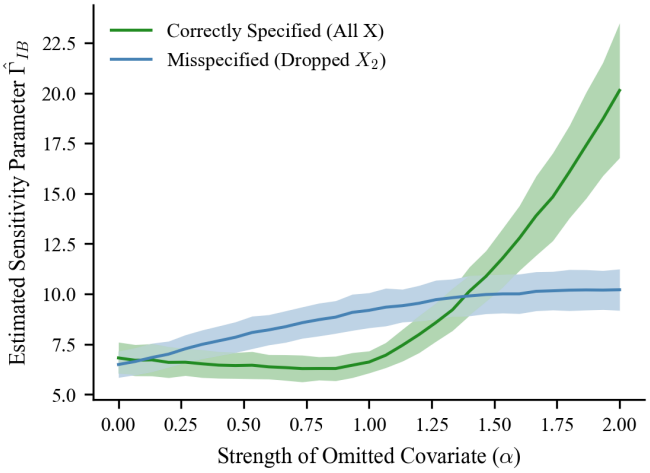

Figure 5: Structural misspecification. Estimated sensitivity parameter $\hat{\Gamma}_{IB}$ against the strength α of the omitted covariate, for the correctly specified (all covariates) and misspecified (drops X_2) models. The misspecified benchmark over-estimates the reference at low α, the curves cross near $\alpha \approx 1.4$, and at high α it plateaus while the reference keeps rising.

The driver is proxy inflation (Figure 6). The correctly specified coefficient $\hat{\beta}_1$ stays flat near 1.40, whereas under the misspecified model X_1 absorbs the dropped covariate's influence and $\hat{\beta}_1$ inflates from 1.58 at $\alpha = 0$ to 2.00 at $\alpha = 2$: the retained covariate stands in as a proxy for the removed one.

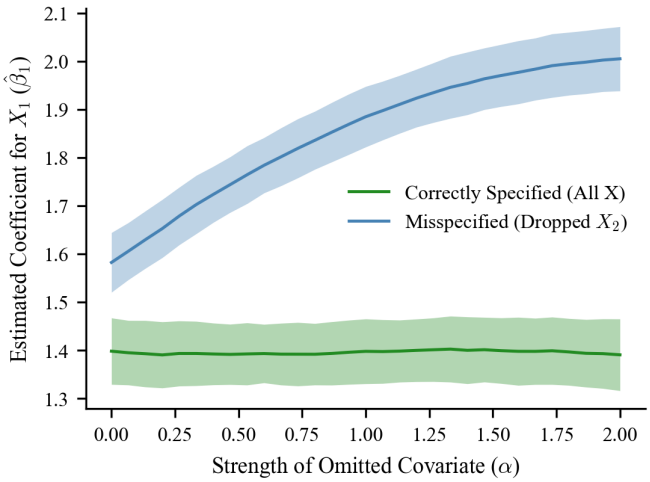

Figure 6: Structural mechanism. Estimated coefficient $\hat{\beta}_1$ of the observed covariate X_1 against the strength α of the omitted, correlated covariate X_2. Dropping X_2 (misspecified model) inflates $\hat{\beta}_1$, whereas the correctly specified model keeps it stable.

The plateau arises because this proxy effect has a ceiling set by the correlation ρ, not by the strength α: X_1 can only substitute for X_2 as far as the two move together, so once it has absorbed what the correlation allows, further increases in α cannot lift the benchmark.