When the Propensity Model Is Wrong

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

Roland Vízner

1. The Problem: Inferring Causation

Example: Smoking and Lung Cancer

- Smokers show higher lung cancer rates.
- But hazardous jobs could raise both smoking and cancer risk: a **confounder**.
- Comparing smokers to non-smokers misleads unless we adjust for it.

Adjusting for observed covariates X:

- **Propensity score** $e(X) = P(T{=}1 \mid X)$: probability of treatment.
- **IPW:** re-weight by $1/e(X)$ to mimic a randomized trial.
- Relies on **ignorability**: all confounders are observed.

Key point: causal estimates are trustworthy only if every confounder is observed.

2. The Catch: Unobserved Confounders

Ignorability rarely holds in real data.

- A hidden confounder U distorts the cause-effect link.
- The truth is $P(T{=}1 \mid X, U)$, so $e(X)$ is biased and IPW fails.

3. Sensitivity Analysis: Bounding the Unknown

Relax ignorability and bound the worst case with a parameter $\Gamma \geq 1$.

Marginal Sensitivity Model (MSM): with the odds ratio $OR(X) = e(X)/(1-e(X))$, the true odds (with hidden U) stay within a factor Γ of the observed ones:

$$\Gamma^{-1} \cdot OR(X) \leq OR(X,U) \leq \Gamma \cdot OR(X)$$

- Gives **bounds** on the effect, not one biased point estimate.
- Larger Γ allows stronger hidden confounding (wider bounds).

Key point: the study's robustness now hinges on choosing a credible Γ.

4. Informal Benchmarking

Choosing Γ is hard. Informal Benchmarking gives a data-driven guess:

1. Take an observed covariate.
2. Drop it from the propensity model.
3. Read the shift in predictions as a hidden confounder of similar strength.

The largest shift across covariates is $\hat{\Gamma}_{IB}$.

5. Knowledge Gap & Research Question

Informal benchmarking assumes the propensity model is correctly specified. What if it is not?

Main Question: how does the informal benchmarking parameter interact with the sensitivity bound when the propensity model is incorrect?

6. Methodology & Simulation Pipeline

A synthetic data-generating process (DGP) isolates a single non-linear error.

- **Variables:** 5 observed covariates X and 2 hidden confounders U coupled to X.
- **True propensity:** logistic in X with coefficients β_X, β_U, plus a quadratic term in X_1 of strength α:

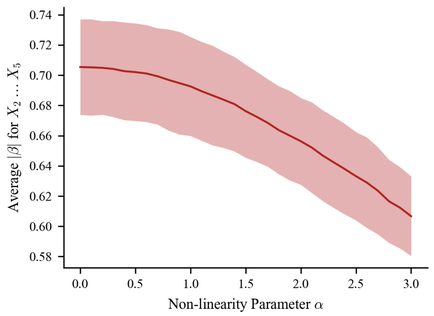
$$e(X,U) = \text{logistic}(\beta_X^\top X + \alpha(X_1^2 - 1/3) + \beta_U U)$$

- **Two models, same data:** *Correct* (includes X_1^2) vs *Misspecified* (strictly linear).
- **Sweep** α and run IB ($M = 100$ trials, $N = 5000$) on $X_2 \dots X_5$ to estimate $\hat{\Gamma}_{IB}$.

Output: two $\hat{\Gamma}_{IB}$ curves, correct vs misspecified, compared as α grows.

7. The Shrinkage (Coefficient Attenuation)

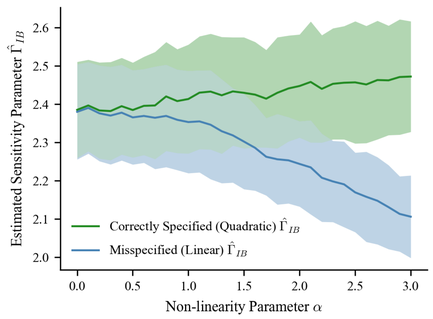
- IB reads confounding from how much dropping a covariate moves predictions.
- That move is set by the covariate's coefficient.
- Unable to fit the U-shape, the linear model shrinks every coefficient toward zero.



Coefficients shrink as the error α grows.

8. The Illusion (False Robustness)

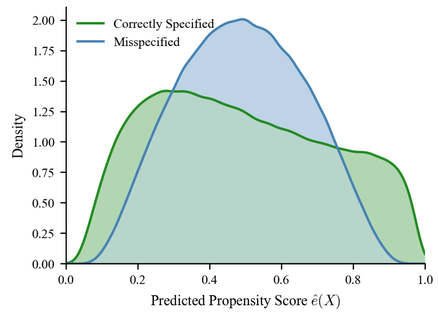
- Shrunk coefficients barely move the odds ratio.
- So $\hat{\Gamma}_{IB}$ deflates while the correct model holds.
- A smaller benchmark means narrower, falsely robust bounds.



The misspecified benchmark deflates, the correct one holds.

9. Seen Directly (Probability Squish)

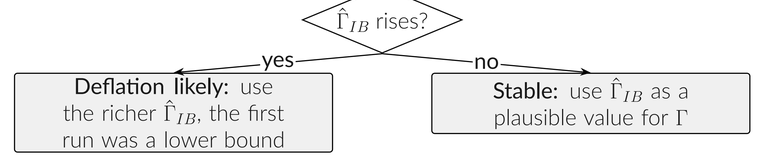
The misspecified model collapses its predictions toward 0.5.



Misspecified predictions pile up near 0.5.

10. The Answer & The Safeguard

- **The answer:** a misspecified model deflates $\hat{\Gamma}_{IB}$ and reports *falsely robust* bounds.
- This is the dangerous direction: hidden confounding looks weaker than it is, with no warning from standard diagnostics.
- **Safeguard:** refit with a richer cross-fitted model and rerun. A rise in $\hat{\Gamma}_{IB}$ exposes the deflation.



11. Limitations & Future Work

- Evidence is from synthetic data, not a real observational study.
- Only one form of error is tested: an omitted quadratic term.
- Only logistic propensity models are compared.
- Future work: nonparametric models and validation on real data.

12. Key Takeaways

- A misspecified propensity model deflates $\hat{\Gamma}_{IB}$, making the bounds look falsely robust.
- The cause is mechanical: shrunk coefficients and predictions squished toward 0.5.
- Cheap safeguard: refit with a richer model, rerun IB, and check if $\hat{\Gamma}_{IB}$ rises.

R.Vizner@student.tudelft.nl

github.com/rolandVi/causal-sensitivity-benchmarking

TU Delft