# Benchmarking the Unobserved

## Coverage Failure in Omitted-Variable Sensitivity Bounds

**Viktor Popdonchev**

**Supervisor(s): Jesse Krijthe, Matej Havelka**

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

## Abstract

Researchers often use observational data to estimate the causal effect of a treatment on an outcome. The central threat to such estimates is an unobserved confounder: a variable that affects both the treatment and the outcome but is not measured. An omitted confounder biases the estimated effect, and this bias does not shrink as the sample grows. Sensitivity analysis addresses this threat by asking how strong a hidden confounder would need to be to overturn a result. A widely used method for linear regression, introduced by Cinelli and Hazlett [6] and extended by Chernozhukov et al. [5] and implemented in the `sensemakr` software, reports an upper bound on the possible bias together with a small set of summary statistics. The bound is valid for whatever confounder strengths the analyst specifies; it cannot supply those strengths, which are unknown. In practice they are supplied by *benchmarking*, which compares the confounder to an observed covariate. This makes two claims at once: that the confounder is no stronger than the covariate, and that the covariate is an appropriate reference for it. The second claim cannot be checked from the data. We study the formal leave-one-out version of this procedure and ask a question its validity proof leaves open: when this assumption is false, does the reported bound still contain the true bias? We answer with Monte Carlo simulations in which the confounder is known, so that the bias and the bound can be compared directly. The bound covers the true bias until the confounder reaches roughly the strength of the covariate it is benchmarked against, and then fails sharply rather than gradually. The strength at which it fails depends on the covariate set in raw terms, so to locate it we derive a relative-strength coordinate that expresses the confounder's strength in the bound's own units. In these units the failure sits at a single point across structurally different covariate sets, the point at which the confounder overtakes the benchmark, and it shifts predictably when the worst-case assumption behind the bound is relaxed. None of the summary statistics warn of any of this: as the bound begins to fail, every one moves in the direction that appears more reassuring. We conclude that these summary statistics, read on their own, do not establish that an estimate is robust to unobserved confounding; they establish robustness only when the benchmarking assumption holds. We recommend that analysts either defend this assumption explicitly or set the confounder's strengths directly from subject-matter knowledge, rather than treating the reported statistics as sufficient.

# 1 Introduction

Observational studies estimate causal effects by comparing units that received a treatment with units that did not, after adjusting for measured covariates. This adjustment is only valid if every variable that affects both the treatment and the outcome has been measured and included. A variable that affects both but is omitted is a *confounder*, and its omission biases the estimate. In linear regression this bias is called *omitted variable bias* (OVB). Unlike sampling error, it does not shrink as the sample grows: a larger biased dataset is just as wrong as a smaller one [13; 2].

The assumption that all confounders have been measured is called *unconfoundedness*, or ignorability [15; 13]. It underlies most methods for causal inference from observational data [4; 10], and it cannot be verified from the data it concerns: the data contain no information about variables that were never measured. A regression fit on the wrong set of controls can be precise and still biased.

Because unconfoundedness cannot be verified, researchers instead ask how far it would have to fail to change a conclusion. This is the goal of *sensitivity analysis*. The approach dates to the study of smoking and lung cancer, where Cornfield et al. [7] showed that any confounder able to explain away the observed association would have to be implausibly strong. Modern sensitivity analysis reports, alongside an estimate, how strong an unobserved confounder would need to be to overturn it.

Two families of methods formalise this question, differing in how they measure the strength of a confounder. The marginal sensitivity model, introduced by Tan [16] and building on the earlier sensitivity model of Rosenbaum [14], bounds how much an unobserved variable could shift treatment assignment; the E-value of VanderWeele and Ding [17] is a related measure used in epidemiology. The OVB / partial-$R^2$ family, developed for linear regression by Cinelli and Hazlett [6] and extended to nonparametric and machine-learning estimators by Chernozhukov et al. [5], measures strength as a share of explained variance, using the same $R^2$ quantities researchers already report; recent work reviews how the different sensitivity models quantify confounding strength [3]. This paper concerns the second family, whose reference implementation is the `sensemakr` package.

## 1.1 The bound and its reporting template

The Cinelli–Hazlett method describes a hypothetical confounder using two quantities: the share of residual outcome variance it would explain, and the share of residual treatment variance it would explain [6]. From these it computes an upper bound on the OVB. The bound is valid for whatever strengths it is given; if the analyst can argue that no realistic confounder exceeds them, the result is robust. To support this argument the method reports a fixed set of summary statistics: robustness values, the partial $R^2$ between treatment and outcome, and an adjusted estimate with its $t$-statistic, summarised in a contour plot. We refer to these collectively as the reporting template, and they appear in this form across many applied papers.

The difficulty is that the analyst does not know the confounder's two strength quantities, since the confounder is unobserved. They are supplied indirectly through *benchmarking*: the analyst sets the confounder's strength equal to a multiple of a measured covariate's strength, most often a

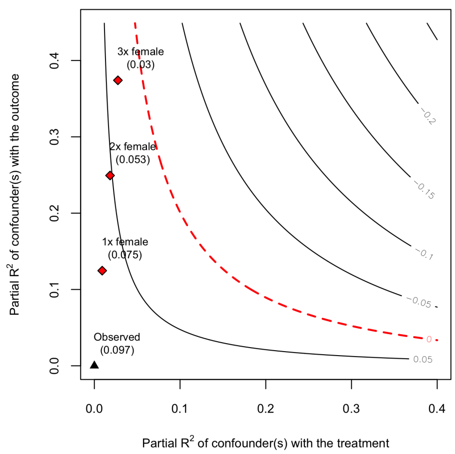


Figure 1: The reporting template, as produced by `sensemakr`. The axes are the confounder's two hypothetical strengths (its partial $R^2$ with the treatment and with the outcome). The solid curve is the boundary at which the bias would equal the estimate; benchmark points mark a confounder $k$ times as strong as a covariate $X_j$. The analyst argues robustness by checking that plausible confounders fall below the boundary.

multiple of one - "the confounder is no stronger than this covariate." Cinelli and Hazlett showed that an earlier, informal form of benchmarking can mislead even when the multiple is correct, because the observed statistics it uses are themselves distorted by the omitted confounder [1; 8; 9; 12]. They proposed a corrected, leave-one-out (LOO) procedure that removes this distortion, derived in Appendix B of their paper [6], and it is this corrected procedure, as implemented in `sensemakr`, that we study.

### 1.2 The benchmarking assumption

Benchmarking makes a claim about strength: that the unobserved confounder is no stronger than a chosen measured covariate, in the partial-$R^2$ units the bound uses. This claim is meaningful only if the covariate is an appropriate reference, that is, if the confounder resembles it closely enough that "no stronger than $X_j$" compares like with like. We call this second requirement the *benchmarking assumption*. Benchmarking therefore rests on two conditions that are easy to conflate: a strength condition, which the analyst controls through the multiplier, and a resemblance condition, which the analyst cannot control and rarely states.

The resemblance condition is needed because benchmarking learns about the confounder only by comparing it to a chosen covariate. If the confounder is unlike any measured covariate, the comparison is to an inappropriate reference, but the procedure still runs and still returns a finite bound. A clear case is an unobserved genetic factor in a study that adjusts only for demographic variables: no measured covariate is a suitable reference, yet nothing in the output indicates this.

This resemblance condition involves an unobserved relationship and therefore cannot be checked from the data; it is a subject-matter judgement. The corrected LOO procedure does not remove the requirement. It still treats each measured covariate, in turn, as a stand-in for the confounder, so its output is only as meaningful as that stand-in is appropriate.

### 1.3 Contribution

The validity proof for the bound assumes the confounder's two strength quantities are known [6; 5]. It does not address the step the analyst actually performs: filling in those quantities by benchmarking. We ask what happens to the bound's coverage of the true bias when the benchmarking assumption is false, that is, when the chosen covariate is not an appropriate reference for the confounder.

This question matters because the benchmarking assumption is rarely stated and almost never checked. Applied papers report the robustness statistics and stop there, so a reader cannot tell whether the reported robustness rests on an assumption that happens to fail. The answer is not settled by the theory alone, because the bound is computed from data that already carry the confounder's influence, so it might widen enough to keep covering.

We answer with simulations in which the confounder is known. We find that the bound covers until the confounder grows strong enough, then fails over a narrow range rather than gradually. The strength at which it fails moves with the covariate set, so we derive a coordinate that measures the confounder in the bound's own units; in these units the failure points coincide at a single threshold, the point where the confounder overtakes the benchmark. No reported statistic warns of the failure: every one moves toward greater apparent robustness as coverage collapses.

The remainder of the paper is organised as follows. Section 2 reviews the OVB bound and its reporting template. Section 3 describes the simulation design. Section 4 presents the results. Sections 5 to 7 discuss the implications and give a reporting recommendation.

## 2 Background

### 2.1 Omitted variable bias

Let $Y$ be an outcome, $D$ a treatment, $X$ a vector of observed covariates, and $A$ an unobserved confounder. The causal effect $\tau$ of $D$ on $Y$ is identified from observational data only if every confounder is observed and included in the regression [15; 13]. Without $A$, the analyst can only fit the *short regression* of $Y$ on $D$ and $X$. The resulting estimator $\hat{\tau}_s$ is biased for $\tau$, and the bias comes entirely from the omitted $A$.

This bias is the product of two quantities: how strongly $A$ predicts the outcome, and how unevenly $A$ is distributed across treatment levels once $X$ is accounted for. If the omitted variable barely predicts the outcome, or is already balanced across treatment levels after conditioning on $X$, its omission costs little; if it does both, the short estimate is badly biased. Unlike sampling noise, this bias does not shrink with more data, since a larger sample simply estimates the biased quantity more precisely. This is what makes confounding different from the uncertainty that standard errors describe.

### 2.2 The bound in partial-$R^2$ form

Cinelli and Hazlett [6] summarise a hypothetical confounder by two numbers: how much of the residual outcome variance it would explain, and how much of the residual treatment variance. Given these two numbers, the method returns the largest bias any confounder that strong could produce. The squared bias is bounded by

$$(\hat{\tau}_s - \tau)^2 \;\le\; C_Y^2 \cdot C_D^2 \cdot S^2, \tag{1}$$

where $C_Y^2 = R^2_{Y\sim A|D,X}$ is the share of residual outcome variance the confounder would explain, $C_D^2 = R^2_{D\sim A|X}/(1-R^2_{D\sim A|X})$ is the corresponding term on the treatment side, and $S^2$ is a scale factor computed from the short regression alone. Chernozhukov et al. [5] show that this is the partially linear special case of a general bound $|\theta_s - \theta|^2 = \rho^2_{\text{align}} B^2 \le B^2$, where $\rho_{\text{align}} \in [0,1]$ measures how aligned the confounder's effects on the treatment and the outcome are; it is distinct from the confounder–covariate correlation $\rho_{\text{proxy}}$ used in Section 3. Equation (1) is the worst case $\rho_{\text{align}} = 1$. We return to this quantity in Section 4.6, because it explains where the failure we observe is located.

### 2.3 The reporting template

The plot in Figure 1 is what an analyst actually reads, and the statistics behind it have plain readings. The robustness value $RV_q$ is the confounder strength, on the diagonal $C_Y^2 = C_D^2$, needed to reduce the estimate by a fraction $q$. Its significance counterpart $RV_\alpha$ is the strength needed to make the estimate no longer statistically significant. The partial $R^2$ of treatment with outcome, $R^2_{Y\sim D|X}$, is a worst-case ceiling: a confounder can overturn the result only if it is at least this strongly associated with the treatment. Finally, the adjusted estimate and adjusted $t$-statistic report what the estimate becomes under a confounder of a stated strength. These five quantities are the reporting template, and the contour plot of Figure 1 is their graphical summary.

The contour plot is worth reading in detail, because our main result concerns it. The horizontal axis is the confounder's partial $R^2$ with the treatment, the vertical axis its partial $R^2$ with the outcome. The study sits at the origin, where no confounder is assumed. The solid *danger boundary* is the curve along which the adjusted estimate reaches zero: a confounder beyond it would explain the effect away entirely. The benchmark points mark where a confounder $1\times$, $2\times$, or $3\times$ as strong as a covariate $X_j$ would lie. The analyst argues robustness by checking that every plausible confounder lies below the danger boundary.

### 2.4 Benchmarking and the benchmarking assumption

Because the confounder's partial-$R^2$ values are unobserved, the analyst benchmarks them against an observed covariate $X_j$. Two multipliers $k_D$ and $k_Y$ set the confounder's strength relative to $X_j$; this is a statement about strength, not about the confounder being interchangeable with $X_j$. In the weak-confounding regime studied here, they reduce to the partial-$R^2$ ratios

$$k_D \approx \frac{R^2_{D\sim A|X_{-j}}}{R^2_{D\sim X_j|X_{-j}}}, \qquad k_Y \approx \frac{R^2_{Y\sim A|X_{-j},D}}{R^2_{Y\sim X_j|X_{-j},D}}, \tag{2}$$

a simplification of the exact multipliers of Cinelli and Hazlett [6]. The default $k_D = k_Y = 1$ encodes "the confounder is no stronger than $X_j$." That statement is only as good as the covariate it is made against, so we state the separate resemblance requirement, the benchmarking assumption proper, as

$$\exists\, j : R^2(A \mid X_j) > 0, \tag{3}$$

where $R^2(A \mid X_j)$ is the share of the confounder's variance that $X_j$ would explain. It asks only that at least one covariate carry some information about $A$, and it cannot be checked from the data, because $A$ is unobserved.

The benchmark is therefore not a pass-or-fail test but the basis of an argument: "a confounder no stronger than $X_j$ cannot overturn the result." That argument rests on two inputs: the chosen multiplier $k$, which sets strength, and the assumption that $X_j$ is an appropriate reference. Passing the check is meaningful only if both hold. Cinelli and Hazlett state explicitly that sensitivity analysis is a tool for structured argument rather than an automatic verdict [6]; our results give that point an empirical basis.

## 3 Methodology

We test whether the LOO bound contains the true bias when the benchmarking assumption is false. This cannot be checked on real data, because it requires knowing the confounder. We therefore use simulations in which we set the confounder ourselves and measure the bias directly. The study is five experiments, each isolating one question: whether the bound keeps containing the bias as the confounder's strength grows (Experiment 1); whether the strength at which it stops depends on the covariate set (Experiment 2); whether the failure is caused by the benchmarking violation or simply by a strong confounder (Experiments 3 and 4); and whether the location of the failure is a real property of the bound or an artefact of how we measure strength (Experiment 5).

### 3.1 Data-generating process

We use a partially linear data-generating process with one treatment, five observed covariates, and one unobserved confounder:

$$X \sim \mathcal{N}_5(0, I_5), \tag{4}$$

$$A = \rho_{\text{proxy}}\, X_1 + \sqrt{1-\rho^2_{\text{proxy}}}\,\eta, \quad \eta \sim \mathcal{N}(0,1), \tag{5}$$

$$D^* = X\beta_{DX} + A\beta_{DA} + \varepsilon_D, \qquad D = \mathbf{1}[D^* > 0], \tag{6}$$

$$Y = \tau D + X\beta_{YX} + A\beta_{YA} + \varepsilon_Y, \tag{7}$$

with $\varepsilon_D, \varepsilon_Y \sim \mathcal{N}(0,1)$ and $\tau = 1$. The treatment is binary, obtained by thresholding the latent index $D^*$ at zero; this is a common setting in applied work and gives a balanced treatment. The method does not depend on the form of the treatment [6; 5]; we report the continuous-treatment case in Appendix D. Each trial draws $n = 2000$ observations.

The parameter $\rho_{\text{proxy}}$ in Equation (5) is how well the covariates proxy the confounder, and it controls the resemblance condition directly. At $\rho_{\text{proxy}} = 0$ the confounder is orthogonal to the covariates, so $R^2(A \mid X) \approx 0$ and no covariate is an appropriate reference: the benchmarking assumption (3) is fully violated. At $\rho_{\text{proxy}} = 1$ the confounder coincides with $X_1$ and is effectively observed. We vary the confounder's strength with a multiplier $m$, setting $\beta_{DA} = \beta_{YA} = 0.5\,m$ and sweeping $m$ over 50 values in $[0.5, 2.0]$, where $m = 1$ makes the confounder as strong as a uniform covariate. Each grid point is repeated over 250 Monte Carlo trials. By default $\rho_{\text{proxy}} = 0$; Experiments 3 and 4 raise it to restore the assumption, stopping short of the degenerate endpoint, and Experiment 5 changes the internal structure of the confounder (Appendix C).

### 3.2 Experimental configurations

Each experiment fixes a *panel*: the covariates and the coefficients $\beta_{DX}, \beta_{YX}$ that set how strongly each covariate relates to the treatment and the outcome. The confounder $A$ is held structurally identical across panels, so that any difference in the results is due to the panel and not to the confounder.

**Experiment 1 (Uniform panel).** $\beta_{DX,j} = \beta_{YX,j} = 0.5$ for all $j$. The five covariates are interchangeable.

**Experiment 2 (Dominant covariate).** $\beta_{DX,1} = \beta_{YX,1} = 0.9$; the remaining four have $\beta \in \{0.2, 0.2, 0.1, 0.1\}$.

**Experiment 3 (Restoring the assumption).** Uniform panel, with $m$ fixed at 1.0 and $\rho_{\text{proxy}}$ swept from 0 to 0.9.

**Experiment 4 (Strength and proxy quality jointly).** Uniform panel, with $m$ and $\rho_{\text{proxy}}$ swept jointly over the full grid.

**Experiment 5 (Alignment).** Uniform panel, $\rho_{\text{proxy}} = 0$, with the confounder split into two channels correlated at $\rho_{\text{align}}$ (Appendix C), swept over $\{1.0, 0.8, 0.6, 0.4\}$ against $m$.

### 3.3 Coverage

The quantity we measure is whether the reported bound contains the true bias. For a single trial, the bound *covers* the truth when $|\text{OVB}| \le B_{\max}$, where $B_{\max}$ is the largest bound the procedure can report at $k = 1$: for each covariate we form the bound of Equation (1) that a confounder exactly as strong as that covariate would imply, and take the largest across the five. *Coverage* is the fraction of trials in which $|\text{OVB}| \le B_{\max}$ holds. A working bound should cover close to 100% of trials; coverage below this means the reported robustness is not supported by the truth. Coverage is the main outcome throughout Section 4.

### 3.4 Quantities recorded

For each trial we record two kinds of quantity. The first kind can only be computed because we built the simulation, and we use it to judge whether the bound covers: the absolute bias $|\text{OVB}| = |\hat{\tau}_s - \tau|$, the coverage indicator, the benchmark bound $B_{\max}$ above, and the oracle bound $B_{\text{oracle}}$ computed from the true $A$. The second kind is what a real analyst would see, the five reporting statistics of Section 2, and we record it to test whether any of it would flag the failure. We define the *failure point* as the strength at which mean coverage crosses 50%, found by linear interpolation. We use the 50% crossing, rather than a higher level, because it marks where the bound the procedure reports and the bound implied by the true confounder cross: at that point the two are nearly equal, so on each trial sampling noise decides whether the bias is contained, and about half the trials fall each way. Coverage degrades before this crossing; in Experiment 1 it falls below 95% already at $m \approx 0.80$, so the strength at which the bound becomes practically unreliable is lower still. The estimation details, and a check that the leave-one-out partial $R^2$ matches the closed-form Frisch–Waugh–Lovell identity to machine precision, are in Appendix B.

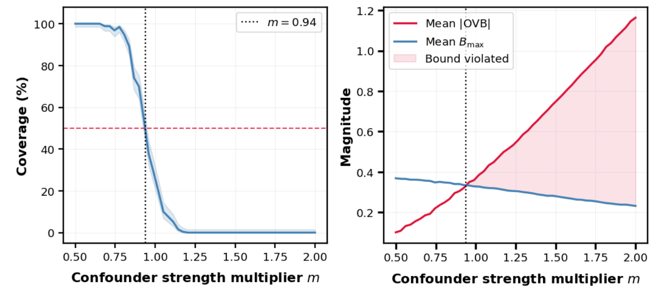


Figure 2: Experiment 1. *Left:* coverage of the LOO bound against $m$; it holds near 100%, then drops through 50% at $m \approx 0.94$. *Right:* the mechanism. The slowly declining bound $B_{\max}$ and the rising bias $|\text{OVB}|$ cross at the failure point.

## 4 Experiments and Results

We present the five experiments as a single argument: each answers the question raised by the one before it.

### 4.1 Experiment 1: the bound fails sharply

We start with the simplest setting, five interchangeable covariates, and ask whether the bound keeps containing the bias as the confounder's strength grows. It does, until it stops abruptly. Coverage stays near 100% while $m$ is small, then drops through 50% at $m \approx 0.94$ over a narrow range (Figure 2, left). The drop is sharp, not gradual.

**Why this happens.** The mechanism is visible in Figure 2 (right): the two terms in the coverage condition move in opposite directions. The bias $|\text{OVB}|$ grows roughly linearly in $m$. The bound $B_{\max}$ moves the other way: as the confounder inflates the residual variance of the treatment and the outcome, each covariate explains a smaller share, so the benchmark it supplies shrinks. The bound does respond to the confounder, because it is computed from $(Y, D, X)$, which carry the confounder's influence, but it responds far too weakly, declining slightly while the bias it must contain rises severalfold. Because coverage is the event $|\text{OVB}| \le B_{\max}$, the failure point is simply where the two curves cross. The location of that crossing in $m$ is not universal: $m$ is defined relative to a covariate of strength 0.5, so its meaning changes when the covariates change. To test whether the failure point is stable we change the covariate set.

## 4.2 Experiment 2: the failure point depends on the covariate set

We now make one covariate dominant ($\beta_1 = 0.9$, the rest weak) and repeat the sweep. The failure point moves to $m \approx 1.43$, about $0.5$ units of $m$ later than in Experiment 1 (Figure 3, left). The bias grows at the same rate, because the confounder is unchanged, but $B_{\max}$ is larger throughout: the strongest benchmark is now the dominant covariate $X_1$, which supplies a wider bound than any covariate in the uniform set. Starting higher, the bound takes longer for the rising bias to overtake, so the failure point moves to the right.

The consequence for an analyst is that a set with one dominant covariate appears to allow a wider safe zone than a set of five moderate covariates, even though the true sensitivity to a given confounder is the same. The apparent safety is set by the strongest covariate the analyst happened to include, which need not have any relation to the confounder that actually threatens the study. So the failure point is not universal in $m$. But $m$ measures the confounder against a covariate whose strength differs between the two experiments, and in coefficient units, while the bound itself is in partial-$R^2$ units. The next step is to measure the confounder's strength the way the bound does.

## 4.3 Synthesis: relative strength is the right coordinate

Because the failure point on the $m$ axis moves with the covariate set, raw coefficient strength cannot predict where the bound breaks. To find a coordinate that can, we measure the confounder's strength in the bound's own units, as the bound it implies divided by the strongest benchmark bound,

$$r := \frac{B_{\text{oracle}}}{B_{\max}} = \sqrt{\frac{C^2_{Y,A}\, C^2_{D,A}}{C^2_{Y,j^*}\, C^2_{D,j^*}}} \approx \sqrt{k_D \cdot k_Y}. \tag{8}$$

The scale factor $S^2$ cancels, so $r$ is in partial-$R^2$ units, and the final approximation, accurate at small partial $R^2$, identifies $r$ as the geometric mean of the true benchmarking multipliers of Equation (2): $r$ is the confounder's true strength expressed in the analyst's own benchmarking units. Re-plotted against $r$, the two failure points coincide, at $r \approx 1$ (Experiment 1 at $r \approx 0.99$, Experiment 2 at $r \approx 1.00$; the bootstrap interval for Experiment 1, $[0.96, 1.00]$, contains 1; Figure 3). The gap on the $m$ axis was an artefact of the axis.

**Why the failure sits at $r = 1$.** Having defined $r$, we can see why the coincidence is forced. The method guarantees that the true bias never exceeds the oracle bound, $|\text{OVB}| \le B_{\text{oracle}}$ [6; 5]. So whenever $r < 1$ we have, per trial, $|\text{OVB}| \le B_{\text{oracle}} < B_{\max}$ and coverage holds; once $r > 1$ the oracle bound exceeds the benchmark bound, and the bias may slip past it. Two covariate sets of different structure must therefore fail at the same $r$. At $r \approx 1$ the two bounds nearly coincide, which is why the transition is sharp and why mean coverage passes through $50\%$ there: a small increase in strength flips many trials at once, and at the crossing itself sampling noise sends about half each way. This settles one side of the result, that $r \le 1$ guarantees coverage; how far past $r = 1$ the bound continues to hold is taken up in Section 4.6.

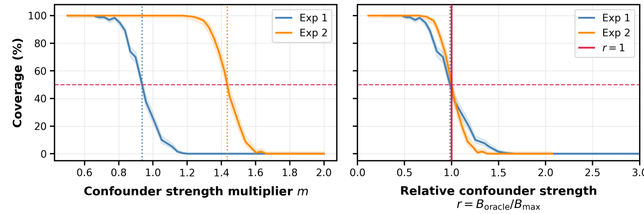


Figure 3: The main result. *Left:* coverage against the multiplier $m$, with the two failure points apart. *Right:* coverage against relative strength $r = B_{\text{oracle}}/B_{\max}$; both sets fail at the same point, $r \approx 1$. Shaded bands show the $p_{10}$–$p_{90}$ spread of $r$.

The bound fails exactly when the confounder overtakes the benchmark. But $r$ is computed from the true $A$, which the analyst never sees; the analyst has only the reporting template. We now ask whether any reported statistic tracks $r$.

## 4.4 The reported statistics do not show the failure

We track every statistic an analyst would report across the sweep (Figure 4). Each changes monotonically as the confounder grows, and each moves in the direction that appears more robust, straight through the failure point. The robustness value $RV_q$ rises; its significance counterpart $RV_\alpha$ rises; the partial $R^2_{Y\sim D|X}$ rises; the adjusted estimate rises without changing sign; and the adjusted $t$-statistic rises. None turns, reverses, or signals an alarm where coverage collapses.

**No threshold rule helps.** Because the statistics move away from danger exactly as the bound fails, no fixed cutoff on any of them can separate the covered case from the uncovered one. Under the rule "declare robust if $RV_\alpha$ exceeds a threshold," $RV_\alpha$ exceeds any reasonable threshold both before and after the failure point, so the rule reports robustness even where the bound has failed, independently of the threshold chosen.

**Why this is unavoidable.** The reason is the same non-identifiability that makes sensitivity analysis necessary in the first place. Unconfoundedness cannot be tested from the observed data [13; 15], and the benchmarking assumption, which concerns the unobserved relationship between $A$ and $X$, inherits this: coverage depends on the bias, which is set by $A$, while every reported statistic is a function of $(Y, D, X)$ alone, and distinct confounding structures can produce the same distribution of $(Y, D, X)$ with different coverage. Detecting the violation therefore requires information from outside the regression. A negative control, for example, can flag residual confounding that no function of $(Y, D, X)$ would reveal [11], but the reporting template contains only such functions.

If the violation is invisible to every observed statistic, is the failure actually caused by the benchmarking violation, or would any bound fail once the confounder is strong enough? We test this by restoring the assumption.

## 4.5 Restoring the assumption restores coverage

We fix the strength at $m = 1.0$, just past the Experiment 1 failure point where coverage has fallen to about $29\%$, and

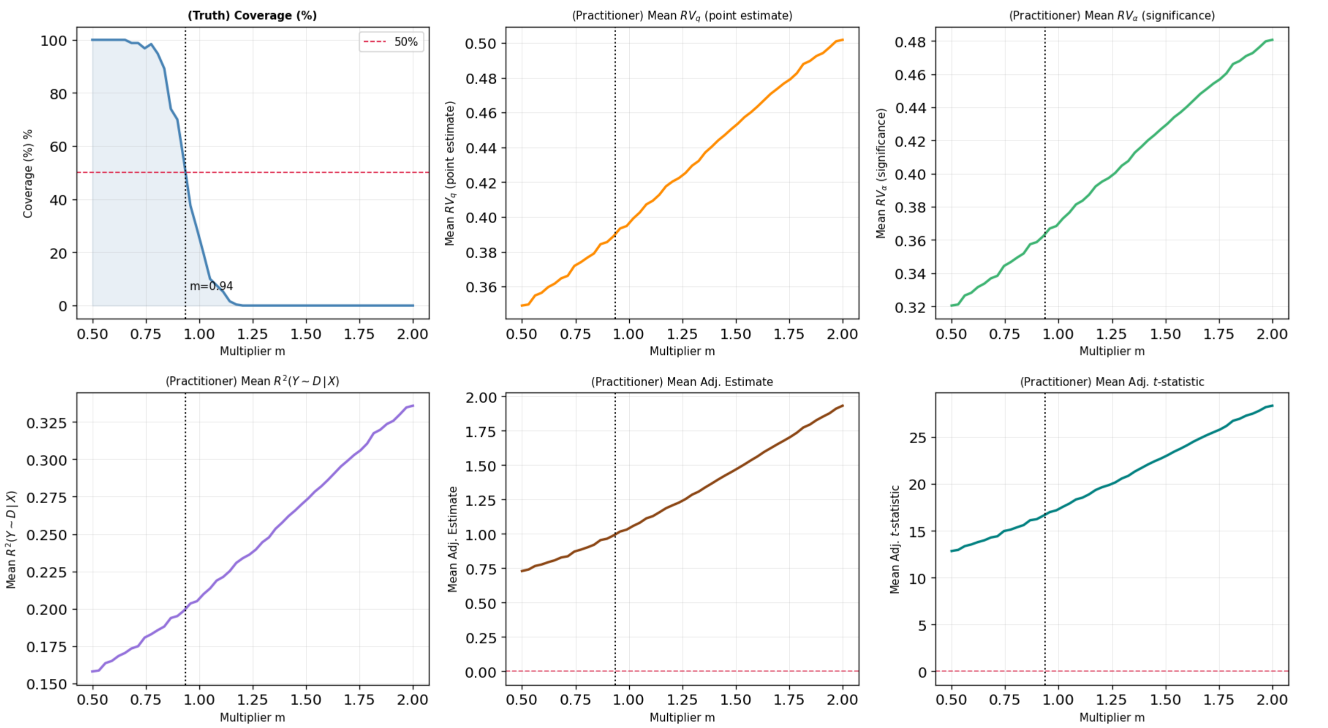


Figure 4: The five reported statistics (and coverage, for reference) across the sweep. The vertical line marks the failure point. Every statistic moves toward greater apparent robustness as coverage collapses; none signals the failure.

raise the correlation $\rho_{\text{proxy}}$ from 0 to 0.9, turning the benchmarking assumption from false back to true. Coverage returns to about 100% by $\rho_{\text{proxy}} \approx 0.3$. Two things happen together: the bias shrinks, because $X_1$ now absorbs part of $A$ in the short regression, and $B_{\max}$ rises, because $X_1$ becomes a better leave-one-out reference for $A$. Here $X_1$ is the proxy by construction; in a real study the analyst would not know which covariate, or combination of covariates, plays that role.

Benchmarking at $k = 1$ bundles two assumptions, and sweeping $m$ and $\rho_{\text{proxy}}$ jointly separates them (Figure 5). The horizontal axis, $m$, carries the strength assumption, that the confounder is no stronger than the benchmark, which holds while $m \leq 1$. The vertical axis, $\rho_{\text{proxy}}$, carries the resemblance assumption, that the covariate is an appropriate reference, restored as $\rho_{\text{proxy}}$ rises. Coverage holds wherever either assumption holds, and fails only in the corner where both are violated, where the confounder is both stronger than the benchmark and unlike it. The contour $r = 1$, computed from the bound's definition rather than fitted, follows the empirical 50% failure line across the whole plane. A benchmarking violation is thus one lever that pushes $r$ past 1, but it is neither necessary nor sufficient: a strong enough confounder crosses $r = 1$ with the assumption intact, and a weak one stays below it with the assumption fully violated.

The single condition is $r > 1$, reached through two levers. One objection remains: $r = 1$ is a definition, since we divided one bound by another and named the crossing. Is the failure a real property of the bound, or only this definition?

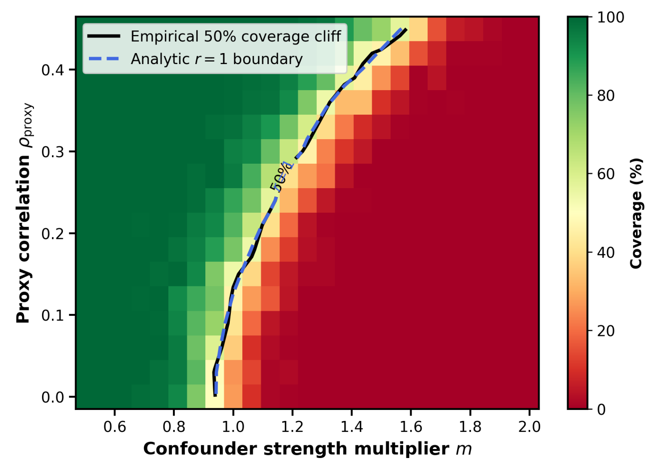


Figure 5: Coverage over the $(m, \rho_{\text{proxy}})$ plane (uniform panel). Green: the bound covers; red: it fails. The contour $r = 1$ (dashed) follows the empirical 50% failure line (solid). Experiment 1 is the bottom edge; restoring the assumption is the vertical direction.

### 4.6 The failure point moves as predicted

That $r \leq 1$ guarantees coverage is a matter of definition. The empirical content is that the failure occurs *at* $r = 1$ rather than past it, and this depends on a quantity held fixed so far: the alignment $\rho_{\text{align}}$ between the confounder's effect on the treatment and its effect on the outcome (Section 2). The bound assumes the worst case $\rho_{\text{align}} = 1$, which our single-channel

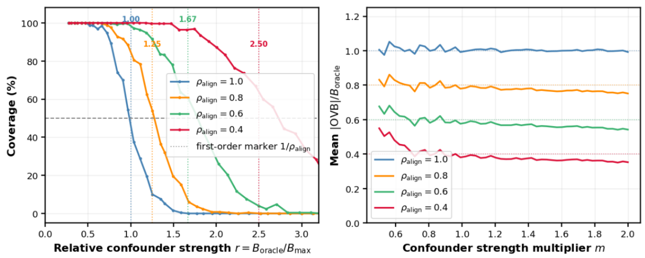


Figure 6: Alignment sweep. *Left:* coverage against relative strength $r$ for four alignment levels; the failure point tracks $1/\rho_{\text{align}}$ (dashed) at $1.00, 1.25, 1.67, 2.50$. *Right:* the realised bias as a fraction of the oracle bound, $|\text{OVB}|/B_{\text{oracle}} \approx \rho_{\text{align}}$.

confounder satisfies. If real confounding is less aligned than this worst case, the bound has slack and should keep covering past $r = 1$.

We test this by splitting the confounder into two channels correlated at $\rho_{\text{align}}$ (Appendix C), swept over $\{1.0, 0.8, 0.6, 0.4\}$. The relation $|\text{OVB}| = \rho_{\text{align}} B_{\text{oracle}}$ predicts the failure point should move to $r = 1/\rho_{\text{align}}$, and it does: the four coverage curves separate, with measured failure points $r \approx 0.99, 1.28, 1.78, 2.73$ against predicted $1.00, 1.25, 1.67, 2.50$ (Figure 6). The match is close at high alignment and looser at low alignment, because at the larger strengths needed to break coverage under weak alignment the realised bias does not reach the worst case; the predicted value is therefore a conservative marker.

What matters is that the failure point moves. A quantity fixed by definition can take only one value; a failure point that moves predictably as we change a known parameter cannot be fixed by definition alone. The guarantee $r \le 1$ is exact; how far past $r = 1$ the bound holds is empirical and is set by alignment.

**Summary of the chain.** Experiments 1 to 5 identify $r > 1$, the point at which the confounder overtakes the benchmark in the bound's own units, as the single condition under which the LOO bound stops containing the bias. A benchmarking violation is one of two levers that push $r$ past 1, and the alignment sets how far past 1 the bound survives. Section 5 draws out what this means for practice.

## 5 Discussion

**Conditional, not unconditional, robustness.** The results show that the `sensemakr` reporting template, used on its own, is not sufficient to establish robustness to unobserved confounding. The five reported statistics together fail to indicate a violation of the benchmarking assumption on which the bound depends. Reporting them and concluding "robust to confounders no stronger than $X_j$" does not establish unqualified robustness; it establishes robustness only if the benchmarking assumption holds, and no observed quantity can confirm that it does. This extends, rather than contradicts, the original warning of Cinelli and Hazlett [6]: they showed informal benchmarking can mislead; we show the corrected procedure carries a conditional validity its statistics do not convey.

**The two ways of mis-setting the multiplier are not symmetric.** The analyst's only control over the benchmark is the multiplier, and coverage depends only on the product $k_D k_Y$, since the bound scales as $\sqrt{k_D k_Y}$ (Appendix C). The two ways of getting it wrong have very different costs. If the product is too small, the bound is too narrow to contain the bias, and the analyst reports robustness that is not there; this error is silent, because every reported statistic moves toward greater apparent robustness as the bound fails. If the product is too large, the bound is too wide and the robustness check is seen to fail, so the analyst withholds a sound conclusion; this error is visible. The default $k = 1$ sits exactly at $r = 1$, the edge of the silent error, so a confounder even slightly stronger than assumed removes the reported robustness with no sign that it has.

**A reporting recommendation.** We recommend two additions to the existing template. First, a defence of the benchmarking assumption. This is not satisfied by stating it, because it cannot be tested; it is an argument about the mechanism. The analyst should name the measured covariate the hidden confounder is expected to resemble, give the subject-matter reason it captures similar variation in the treatment and the outcome, and, where possible, bound the confounder's strength relative to it from prior studies or theory. Second, and because that argument cannot be verified, sensitivity to the multiplier: the analyst should report adjusted estimates across a small set of values, for example $k \in \{1, 2, 3, 5\}$, so the reader can see how far the conclusion depends on the assumed strength. Equivalently, the analyst can bypass benchmarking and set the strengths $C_Y$ and $C_D$ directly from subject-matter knowledge, then read the worst-case bias over that region; this uses the same bound without the benchmarking assumption, at the cost of an explicit strength claim that must be defended. Reporting only the default $k = 1$ commits to the most optimistic defensible case without saying so.

**No observed-data diagnostic can replace this.** It is natural to look for some other function of $(Y, D, X)$ that would detect the violation. Section 4.4 gives the reason this cannot work for any statistic of the regression itself: coverage is determined by the unobserved relationship between $A$ and $X$, and distinct confounding structures can produce the same observed data with different coverage. Information about $R^2(A \mid X)$ must come from outside the regression, from subject-matter argument or from auxiliary data such as a negative control [11].

**Limitations.** Our results concern the LOO bound under a partially linear data-generating process with five covariates and one confounder, estimated by ordinary least squares. The observed covariates are mutually independent, which is the most favourable case for benchmarking, because each provides a clean and separate reference. In applied work the covariates are usually correlated, and a confounder is often proxied by a combination of them rather than by any single covariate; whether the failure point holds there is the most important case our design leaves open. The results should also not be assumed to hold for nonparametric estimation, panel or time-series data, instrumental-variables designs, or heterogeneous effects. The most direct extension reverses the question: instead of asking whether the bound covers at $k = 1$, one could ask how accurately the LOO procedure estimates the confounder's true strength $r$ from observed data, which we take up next.

## 6 Responsible Research

**Reproducibility.** All experiments are implemented in Python using `numpy` and `statsmodels`. Each trial uses a fixed random seed determined by its trial and multiplier indices, so re-running an experiment produces identical samples. The data-generating process, the swept parameters, and the number of trials are stated in full in Section 3. The source code and the notebooks that produce every figure are available at *https://github.com/Endeville/Informal_Benchmarking*.

**Scope.** Our results apply to a partially linear data-generating process with five observed covariates and one unobserved confounder, estimated by ordinary least squares. They should not be assumed to hold for nonparametric estimation, panel or time-series data, instrumental-variables designs, or heterogeneous effects without further study.

**Use of synthetic data.** Synthetic data is necessary for this study. The benchmarking assumption concerns an unobserved quantity, and checking whether a bound covers the true effect requires knowing the confounder, which no real dataset provides. Our data-generating process is fully specified and the code is available at the repository above, so any reader can reproduce or modify it.

**Potential for misuse.** Our results could be read as a general argument against using `sensemakr` in applied work. That is not our claim. We recommend adding a defence of the benchmarking assumption to the existing reporting, not discarding the method or the results that use it. The reported statistics remain informative when the assumption holds. The broader concern is specific: a method that appears more robust as its validity fails can give false reassurance, including to decisions that depend on the result.

**Use of large language models.** A large language model was used as an assistive tool during this project, for editing and reviewing the text and for assistance while writing code. We designed the experiments, specified and ran the data-generating process, and made all methodological choices. Every numerical result, block of code, and claim produced with model assistance was checked independently, and no text was included without review. We take full responsibility for the content of this thesis, including any errors.

## 7 Conclusions and Future Work

An analyst who runs `sensemakr`, reads a reassuring robustness value, and concludes that an estimate is safe from hidden confounding is relying on an assumption they did not state and cannot check: that some measured covariate is an appropriate reference for the unknown confounder. This paper asked what happens to the reported bound when that assumption is false. The question was not whether the bound eventually fails, but where it fails, how abruptly, and whether the reported statistics reveal it.

Our simulations established four results. First, the bound contains the true bias until the confounder reaches roughly the strength of the covariate it is benchmarked against, and then fails over a narrow range rather than gradually. Second, the strength at which it fails depends on the covariate set in raw coefficient units, so we derived a coordinate, $r$, that measures the confounder in the bound's own units; in these units the failure points coincide at $r \approx 1$, the point where the confounder overtakes the benchmark. Third, none of the reported statistics indicates this failure: every one moves toward greater apparent robustness as coverage collapses. Fourth, the failure is governed by the single condition $r > 1$, reached through two levers, the confounder's strength and how well the covariates proxy it, so that restoring the benchmarking assumption restores coverage; how far past $r = 1$ the bound keeps covering is set by the alignment between the confounder's two effects.

These results lead to a concrete recommendation. A robustness section using `sensemakr` should defend the benchmarking assumption, naming which measured covariate the hidden confounder is expected to resemble and why, and should report adjusted estimates across a range of multipliers $k$ rather than the default $k = 1$ alone, which sits exactly at the point where the bound begins to fail.

Two directions extend this work. The first is to test whether the same failure appears under nonparametric estimation [5; 4]. The second follows from the role of $r$. We used $r$ only as a diagnostic, computed from the known confounder to locate and explain the failure; a natural next question is how accurately benchmarking can estimate $r$ from observed data alone, which would shift the object of study from whether the bound covers at a fixed $k$ to how well benchmarking recovers the confounder's true strength.

## References

[1] Joseph G Altonji, Todd E Elder, and Christopher R Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184, 2005.

[2] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[3] 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. *HAL preprint hal-05330588*, 2025.

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[5] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. *arXiv preprint arXiv:2112.13398v5*, 2024.

[6] 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, 2020.

[7] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203, 1959.

[8] Kenneth A Frank. Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2):147–194, 2000.

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

[10] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[11] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

[12] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

[13] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[14] Paul R Rosenbaum. *Observational studies*. Springer, 2002.

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

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

[17] Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

# A Implementation details

This appendix records the constants behind the sweeps so a reader can reproduce them exactly. Each trial draws $n = 2000$ observations, with $\tau = 1$ and five covariates $X \sim \mathcal{N}_5(0, I_5)$. The uniform panel uses $\beta_{DX} = \beta_{YX} = 0.5$; the dominant-covariate panel uses $\beta_{DX} = \beta_{YX} = (0.9, 0.2, 0.2, 0.1, 0.1)$. The confounder enters both equations with coefficient $0.5\,m$. The one-dimensional strength sweep uses 50 values of $m$ in $[0.5, 2.0]$ with 250 trials each; the joint $(m, \rho_{\text{proxy}})$ sweep uses 150 trials per cell.

The treatment is binary, $D = \mathbf{1}[D^* > 0]$ with $D^* = X\beta_{DX} + A\beta_{DA} + \varepsilon_D$, which gives a balanced treatment in every configuration; the continuous case is in Appendix D. All randomness comes from a single seeded generator per trial, so that the experiments are deterministic. The draws occur in a fixed order, with the auxiliary draw used only for the alignment sweep placed last; this makes the two-channel construction at $\rho_{\text{align}} = 1$ identical to the single-confounder case, so the alignment sweep contains Experiment 1 exactly. The confounder is built as $A = \rho_{\text{proxy}} X_1 + \sqrt{1 - \rho_{\text{proxy}}^2}\,\eta$, which keeps $\text{Var}(A) = 1$ for every $\rho_{\text{proxy}}$, so that changing $\rho_{\text{proxy}}$ changes only how much of the confounder $X_1$ can explain, not how strong the confounder is.

# B Estimation and the reported statistics

The two ingredients of the bound are partial $R^2$ values and a scale factor, and both come from ordinary least squares residuals. For a target $y$, controls $X$, and added regressor $A$, the partial $R^2$ is the share of the residual variance of $y$ that $A$ removes,

$$\eta^2_{y\sim A|X} \;=\; \frac{\text{RSS}(y \sim X) - \text{RSS}(y \sim [X, A])}{\text{RSS}(y \sim X)},$$

which we compute through a reduced QR decomposition for speed and stability. The scale factor $S^2 = E[Y^2_{\text{res}}]/E[D^2_{\text{res}}]$ converts these shares into the units of the estimate, where $D_{\text{res}}$ is the residual of $D \sim X$ and $Y_{\text{res}}$ the residual of $Y \sim [D, X]$. The two terms of the bound are $C_Y^2 = \eta^2_{Y\sim A|D,X}$ and $C_D^2 = \eta^2_{D\sim A|X}/(1 - \eta^2_{D\sim A|X})$, and the bound is $B = \sqrt{C_Y^2 C_D^2 S^2}$. The factor $1/(1 - \eta^2)$ on the treatment side is what makes a benchmark covariate that is nearly collinear with the others diverge; the code flags this, and it does not occur for the independent covariates used in the main text.

The reported statistics follow from the same quantities. The significance robustness value has the closed form of Cinelli and Hazlett,

$$RV_\alpha = \tfrac{1}{2}\left(\sqrt{f^4 + 4f^2} - f^2\right), \qquad f = \frac{\max(0,\ |\hat\tau| - z_{\alpha/2}\,\text{SE})}{\sqrt{S^2}},$$

and reads as the confounder strength that would erase statistical significance. The adjusted estimate is $\hat\tau - \text{sign}(\hat\tau)\,B$ and the adjusted $t$-statistic is $(\hat\tau - \text{sign}(\hat\tau)\,B)/\text{SE}$; both can change sign once $B > |\hat\tau|$, that is, once the assumed confounder is strong enough to explain the effect away. We checked the leave-one-out partial $R^2$ against the closed-form Frisch–Waugh–Lovell identity $R^2(Y \sim X_j \mid D, X_{-j}) = t_j^2/(t_j^2 + \text{df})$, which expresses a covariate's partial $R^2$ through its regression $t$-statistic and the residual degrees of freedom; the two agree to machine precision ($|\Delta| < 10^{-15}$).

# C The coverage condition and the alignment law

This appendix explains, rather than just states, why coverage hinges on $r$. Coverage on a trial is the event $|\text{OVB}| \le B_{\max}$, and the method guarantees $|\text{OVB}| \le B_{\text{oracle}}$ [6; 5]. With $r = B_{\text{oracle}}/B_{\max}$,

$$r \le 1 \iff B_{\text{oracle}} \le B_{\max} \implies |\text{OVB}| \le B_{\text{oracle}} \le B_{\max}.$$

So $r \le 1$ implies coverage on every trial, for any covariate set, which is why the failure cannot occur before $r = 1$ and why sets of different structure fail at the same $r$. The Experiment 1 numbers make the transition concrete. Just past the failure ($m \approx 0.96$) the mean bias and oracle bound nearly coincide, $|\mathrm{OVB}| \approx B_{\mathrm{oracle}} \approx 0.35$, while the strongest benchmark gives $B_{\max} \approx 0.33$, so $r \approx 1.04$ and the bound no longer contains the bias; one step earlier ($m \approx 0.93$) the same quantities give $r \approx 0.97$ and the bound contains it. The coverage curve is the population count of this sign change, and because the two bounds are so close near the crossing, sampling noise alone decides each trial there, which places mean coverage at 50%.

The alignment law explains how far past $r = 1$ the bound survives. Chernozhukov et al. [5] write the bias as $|\hat{\tau}_s - \tau| = \rho_{\mathrm{align}} B$, where $\rho_{\mathrm{align}} \in [0, 1]$ is the alignment of the confounder's two effects and $B$ is the worst-case bound. Identifying $B$ with $B_{\mathrm{oracle}}$ gives $|\mathrm{OVB}| = \rho_{\mathrm{align}} B_{\mathrm{oracle}}$, so coverage fails when $\rho_{\mathrm{align}} B_{\mathrm{oracle}} > B_{\max}$, that is, when $r > 1/\rho_{\mathrm{align}}$. This relation is exact in the small partial-$R^2$ limit; at finite strength the realised bias falls short of the worst case, so the failure point sits at or beyond $1/\rho_{\mathrm{align}}$, with the gap growing in strength. For the alignment sweep the single confounder is replaced by two unit-variance channels $A_D$ and $A_Y$ correlated at $\rho_{\mathrm{align}}$, with $A_D$ entering the treatment and $A_Y$ the outcome; each has unit variance, so $B_{\mathrm{oracle}}$ is nearly fixed across $\rho_{\mathrm{align}}$ and only the realised bias changes. At $\rho_{\mathrm{align}} = 1$ the two channels coincide and the single-confounder process is recovered. Finally, because $C_Y^2 \approx k_Y C_{Y,j}^2$ and $C_D^2 \approx k_D C_{D,j}^2$, a confounder benchmarked at $(k_D, k_Y)$ has bound $\sqrt{k_D k_Y}\, B_j$, so coverage depends only on the product $k_D k_Y$: moving strength between $k_D$ and $k_Y$ at fixed product leaves the verdict unchanged.

## D Robustness checks

Four checks confirm that the main findings are not artefacts of specific choices, and we state what each rules out. The implementation is anchored by three internal consistencies: the leave-one-out partial $R^2$ matches the Frisch–Waugh–Lovell identity to machine precision, the two-channel construction at $\rho_{\mathrm{align}} = 1$ reproduces the single-confounder data trial for trial, and the $\rho_{\mathrm{proxy}} = 0$ column of the joint sweep reproduces the Experiment 1 data. Replacing the binary treatment with a continuous one leaves the results qualitatively unchanged, with the sharp failure and its location at $r \approx 1$ intact and a slightly cleaner map from coefficients to partial $R^2$, which shows the binary threshold is not driving the result. Replacing the single confounder with three, variance-locked so the total confounding variance matches the baseline, leaves the failure point indistinguishable from Experiment 1, which shows the result is not specific to a single confounder. Negating the outcome-side coefficient leaves the failure point in the same place; the adjusted estimate changes sign at large $m$, but the reported statistics still fail to flag the failure, which shows the diagnostic blindness is not specific to the sign of the confounding.