Bayesian Causal Stress Testing: Posterior Fragility of Treatment-Effect Conclusions

Makoto Nakakita
Center for Advanced Intelligence Project
RIKEN
makoto.nakakita@riken.jp

Teruo Nakatsuma
Faculty of Economics
Keio University
nakatuma@econ.keio.ac.jp

Abstract

Bayesian causal inference quantifies uncertainty in treatment effects under a specified causal model. This uncertainty is often interpreted as evidence that the resulting causal conclusion is robust. We argue these are distinct targets: a posterior over an effect τ may be tightly concentrated above zero while the corresponding causal claim is fragile to small violations of ignorability, overlap, or target-population assumptions. We introduce *Bayesian causal stress testing* (BCS), a framework that assigns a posterior fragility score F_α to a causal conclusion by measuring the smallest calibrated stress under which the posterior probability of the claim falls to or below a target level. We give a stress-map formulation, decision and composition results for claim-level stress, a calibration theorem showing that $F_{0.5}$ recovers the partial-correlation form of the Cinelli–Hazlett robustness value under linear-Gaussian conditions on the posterior median and bias scale, and a proposition formalizing why posterior precision and posterior fragility can decouple. We evaluate BCS on IHDP, TWINS, and ACIC-cov+Hill, a semi-synthetic benchmark built from the 2016 Atlantic Causal Inference Conference (ACIC) covariate matrix, under Bayesian additive regression trees (BART), a BART variant modeled on Bayesian causal forests (BCF-style), Bayesian linear, and Bayesian-bootstrap augmented inverse-probability-weighted (AIPW) posteriors. On the Bayesian linear posterior, $F_{0.5}$ matches the square root of the Cinelli–Hazlett robustness value, $\sqrt{RV^{\mathrm{CH}}}$, with root-mean-square (RMS) discrepancy 0.001 on both IHDP and ACIC-cov+Hill, each over its ten main configurations; the BART and BCF-style posteriors show larger deviations when effects are small. In the original (unattenuated) IHDP configurations, mean posterior 95% interval widths differ about twofold across posterior generators while generator-level mean $F_{0.5}$ differs by less than 0.01, illustrating that posterior precision is not a robustness certificate. In simulations with a known Gaussian confounder, the mean absolute discrepancy between $F_{0.5}$ and the population flip-stress implied by the data-generating process is 0.033 over the 90 of 96 independently seeded zero-effect simulations whose sign claim is supported.

1 Introduction

A Bayesian causal analysis combines a model M with data D and returns a full posterior $p(\tau \mid D, M)$ for a treatment effect τ. A narrow 95% credible interval that excludes zero is often communicated as definitive evidence: the effect is positive with high posterior probability. However, the posterior is conditional on M, which encodes ignorability, overlap, and a target population. A narrow interval quantifies uncertainty *inside* M; whether the causal conclusion survives small violations of the assumptions that identify and define the effect is a separate question.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This paper proposes that a Bayesian causal analysis report a different object alongside the usual effect posterior: the *posterior fragility of the causal conclusion*. We call the resulting procedure *Bayesian causal stress testing* (BCS). For example, a posterior may put 0.99 mass on $\tau > 0$, yet the conclusion may hinge on an unmeasured confounder. We describe the strength of such a confounder by a single number γ, its partial correlation with the treatment given the observed covariates and with the outcome given the treatment and the covariates, the two taken to be equal; $\gamma = 0$ means no hidden confounding, and larger values mean a stronger confounder. Correcting for a confounder of strength only $\gamma = 0.20$ may cut the posterior probability of a positive effect to $1/2$. Likewise, an average treatment effect (ATE) may be positive for the empirical population, but a small population shift toward a subgroup with little treated–control overlap may reverse the claim. In both cases, the original analysis is statistically confident but causally fragile.

BCS builds on sensitivity analysis. It applies sensitivity-style perturbations, which we call stresses, to the posterior draws and reports a distance to the decision boundary: how far must the model or estimand move before the posterior no longer supports the claim? This distance can be computed for any conclusion, including "ATE is positive," "a subgroup effect is positive," or "the policy improves outcomes for the target population."

Contributions. (i) A posterior fragility score F_α, the size of the smallest stress that pushes the posterior probability of the claim down to a chosen level α or below. It is defined for a conclusion rather than a parameter, is computed from posterior draws, and applies to any family of stresses. (ii) Proposition 1: F_α equals the smallest stress that changes the Bayes-optimal action under a threshold utility; the same holds along any one-dimensional path of stresses, and a local bound covers effect thresholds other than zero. (iii) Proposition 2: when two stresses act together and their joint size is the larger of the two sizes, the joint score is at most the smaller of the two separate scores, so robustness to several stresses can be no larger than the smallest single-stress score. (iv) Theorem 1: in the linear-Gaussian case, under stated conditions on the posterior median and the bias scale, $F_{0.5}$ equals the square root of the Cinelli–Hazlett robustness value RV^{CH}, which anchors the stress scale to a classical sensitivity measure. At $\alpha = 0.5$ this score depends on the posterior only through its median; other levels α use other quantiles. (v) Proposition 3: posterior precision and posterior fragility can move independently, so a narrow credible interval is not a robustness certificate. (vi) Empirical evaluation on IHDP, TWINS, and ACIC-cov+Hill with posteriors from Bayesian additive regression trees (BART), a two-surface BART modeled on Bayesian causal forests (BCF-style), Bayesian linear regression, and the Bayesian bootstrap of augmented inverse-probability-weighted (AIPW) scores. The BART and BCF-style scores depart from $\sqrt{RV^{\mathrm{CH}}}$ when effects are small.

2 Related work

Classical sensitivity analysis. Rosenbaum bounds [1], E-values [2], and the Cinelli–Hazlett partial-R^2 robustness value [3] calibrate deviations from ideal causal assumptions on a point estimate or test statistic.

Bayesian sensitivity analysis for unmeasured confounding. A line of work extends sensitivity analysis to a Bayesian framework by placing priors on bias parameters and propagating uncertainty through the posterior [4–6]. The closest method is Dorie et al. [5], which combines BART with a two-parameter sensitivity surface; the `treatSens` package returns posteriors of τ at fixed sensitivity values. Veitch and Zaveri [7] provide a nonparametric formulation with Austen plots. These methods produce stress-conditional posteriors $p(\tau \mid D, \delta)$ indexed by the sensitivity parameters δ. Relative to them, we add the conclusion-level scalar F_α, a quantile-based summary of such a sensitivity surface, together with its calibration to the Cinelli–Hazlett robustness value (Theorem 1), the target-shift score G_α (Section 4.2), and the Bayes-action interpretation of Proposition 1. The omitted-variable bias formula and the BART posteriors are prior art.

Bayesian causal models. BART for causal inference [8] and Bayesian causal forests [9] produce posteriors over heterogeneous effects. The Bayesian bootstrap of Rubin [10] also supports weighted summaries of AIPW scores. BCS is estimator-agnostic and operates as a reporting diagnostic on top of any of these posteriors; model-based Bayesian sensitivity analysis remains the tool for inference under a specified latent confounder.

3 Setup

We observe data $D = \{(X_i, A_i, Y_i)\}_{i=1}^n$ on n units, with covariates X_i, treatment A_i, and outcome Y_i, and target a causal estimand τ such as the ATE. The Bayesian causal model M produces draws $\tau^{(\ell)} \sim p(\tau \mid D, M)$, $\ell = 1, \ldots, S$. We assume throughout: (i) ignorability conditional on X (no unmeasured confounding under the model M); (ii) positivity, or overlap ($0 < P(A = 1 \mid X) < 1$ on the support of X); (iii) the stable unit treatment value assumption (SUTVA). The treatment A is binary in $\{0, 1\}$ except in the calibration simulations of Appendix G, which use a continuous treatment. For the sign claim $C = \{\tau > 0\}$, the posterior probability is $q_0(D) = P(C \mid D) = P(\tau > 0 \mid D, M)$; below we suppress M from the notation. A large $q_0(D)$ indicates posterior confidence under M, but does not measure robustness to violations of M or to changes in the target population.

Stress maps. A stress map $T_\delta(\cdot, D)$ moves each posterior draw to its value under a stress δ, which ranges over a set Δ of stresses, giving $\tau_\delta^{(\ell)} = T_\delta(\tau^{(\ell)}, D)$; at $\delta = 0$ the draws are unchanged, so $\tau_0 = \tau$. We call a stress calibrated when its size is measured on an interpretable scale, such as a partial correlation. The fragility curve is the stressed posterior probability of the claim,

$$q(\delta; D) = P(\tau_\delta > 0 \mid D), \qquad \widehat{q}(\delta; D) = \frac{1}{S} \sum_{\ell=1}^{S} \mathbf{1}\{\tau_\delta^{(\ell)} > 0\},$$

where $\widehat{q}$ is its finite-draw estimate and $q(0; D) = q_0(D)$. BCS stress-tests the fitted posterior as it stands: the stress acts on the draws of M, with no refit and no prior on the stress.

4 Posterior fragility

Definition 1 (Posterior fragility score)**.** For the claim $C = \{\tau > 0\}$ and a level $\alpha \in (0, 1)$,

$$F_\alpha(D; \Delta) = \inf_{\delta \in \Delta} \big\{ \, \|\delta\| \; : \; P(\tau_\delta > 0 \mid D) \le \alpha \, \big\},$$

where $\|\cdot\|$ is a norm on Δ and $\inf \varnothing = +\infty$. Thus F_α is the size of the smallest stress at which the fragility curve $q(\delta; D)$ falls to α or below, and we also call it the flip distance of the claim; a smaller F_α means the claim fails under a weaker calibrated stress.

Unsupported-claim convention. If $P(C \mid D) \le \alpha$ before stress, the posterior claim is unsupported; we report *no-claim* and do not interpret the zero-distance infimum as a fragility score.

4.1 Hidden-confounding stress

This stress posits an omitted confounder U that has the same partial correlation $\gamma \in [0, 1)$ with the treatment (given X) and with the outcome (given A and X); the stress is γ itself, with size $\|\gamma\| = \gamma$. The bias that U induces is scaled by $s(D) = \sigma_{Y|A,X}/\sigma_{A|X}$, the ratio of the residual standard deviations from the ordinary least squares (OLS) regressions of Y on (A, X) and of A on X, computed from the data alone with the same variance divisor. By the omitted-variable bias formula of Cinelli and Hazlett [3], such a confounder biases the effect by $b(\gamma; D)$, and the stress map subtracts this amount from every draw:

$$b(\gamma; D) = \frac{\gamma^2}{\sqrt{1 - \gamma^2}} \, s(D), \qquad \tau_\gamma^{(\ell)} = \tau^{(\ell)} - b(\gamma; D).$$

The bias is subtracted because it works against the claim $\tau > 0$; for the claim $\tau < 0$ it is added. Because $s(D)$ does not depend on the chosen Bayesian model, all posteriors are stressed on the same scale. Throughout, F_α without a family argument denotes the score for this family; the score for the target-population stress of Section 4.2 is written G_α.

4.2 Target-population stress

This stress moves the population toward a subgroup defined by a binary covariate H, whose share in the data is $p_H = P(H = 1) \in (0, 1)$. We fit a Bayesian linear regression of Y on A, H, their

product AH, and covariates (Appendix A). The draws $\beta_A^{(\ell)}$ and $\beta_{AH}^{(\ell)}$ of the coefficients of A and AH give the treatment effect $\beta_A^{(\ell)}$ outside the subgroup and $\beta_A^{(\ell)} + \beta_{AH}^{(\ell)}$ inside it. A stress $\lambda \geq 0$ raises the log-odds of subgroup membership by λ, which changes the subgroup share to $p_H(\lambda)$ and the ATE of the population to τ_λ:

$$p_H(\lambda) = \text{logit}^{-1}\big(\text{logit}(p_H) + \lambda\big), \qquad \tau_\lambda^{(\ell)} = \beta_A^{(\ell)} + p_H(\lambda)\beta_{AH}^{(\ell)}.$$

The target-shift score is $G_\alpha(D) = \inf\{\lambda \geq 0 : P(\tau_\lambda > 0 \mid D) \leq \alpha\}$, the fragility score of this family.

4.3 Computation

The hidden-confounding stress subtracts the same amount $b(\gamma; D)$ from every draw, and b increases in γ. Hence $P(\tau_\gamma > 0 \mid D) \leq \alpha$ exactly when $b(\gamma; D)$ reaches the posterior $(1-\alpha)$-quantile $Q_{1-\alpha}(\tau \mid D)$, and $F_\alpha = \inf\{\gamma : b(\gamma; D) \geq Q_{1-\alpha}(\tau \mid D)\}$. We invert b at the interpolated sample quantile. This approximates the score of the exact posterior and need not equal the first crossing of the finite-draw curve $\widehat{q}$ exactly (Appendix A). By this quantile formula, $F_{0.5}$ depends on the posterior only through its median, so differences in $F_{0.5}$ across estimators reflect their posterior medians, not their higher moments; the full curve depends on the whole distribution. For the target-population stress and for stresses that combine the two families (Proposition 2), we evaluate $\widehat{q}$ on a grid of stress values.

5 Theory

We give two structural results (Propositions 1–2), a calibration theorem linking the stress scale to the Cinelli–Hazlett robustness value (Theorem 1), and a final proposition formalizing why posterior precision and posterior fragility can decouple. Main-text proofs are sketches; full algebraic details are in Appendix O.

Proposition 1 (Decision-flip equivalence and local utility sensitivity)**.** *Fix a path of stresses $\xi : [0, L) \to \Delta$, $0 < L \leq \infty$, indexed by size, $\|\xi(t)\| = t$. Write $q(t) = P(\tau_{\xi(t)} > 0 \mid D)$ for the fragility curve along the path and $F_{\alpha,\xi} = \inf\{t : q(t) \leq \alpha\}$ for the score restricted to the path.*

(a) Consider acting on the claim ($a = 1$) or not ($a = 0$) under the threshold utility $\mathcal{U}(a, \tau) = a(\mathbf{1}\{\tau > 0\} - \alpha)$, and let a_δ^* denote the Bayes action under stress δ, the action with the larger posterior expected utility under τ_δ (0 at a tie). If $P(C \mid D) > \alpha$, then $F_\alpha(D; \Delta) = \inf\{\|\delta\| : a_\delta^* \neq a_0^*\}$ and $F_{\alpha,\xi} = \inf\{t : a_{\xi(t)}^* \neq a_0^*\} \geq F_\alpha(D; \Delta)$.

(b) Replace the threshold 0 by an effect threshold θ: take $\mathcal{U}_\theta(a, \tau) = a(\mathbf{1}\{\tau > \theta\} - \alpha)$, write $q(t, \theta) = P(\tau_{\xi(t)} > \theta \mid D)$ (so $q(t) = q(t, 0)$), and denote the resulting path score by $F(\theta)$, with $t_0 = F(0) = F_{\alpha,\xi} \in (0, L)$. Suppose q is continuously differentiable near $(t_0, 0)$ with partial derivatives $\partial_t q$ and $\partial_\theta q$, $\partial_t q(t_0, 0) < 0$, and nearby first crossings form the local root $t = F(\theta)$. Then, with f_t the posterior density of $\tau_{\xi(t)}$,

$$F'(0) = \frac{f_{t_0}(0)}{\partial_t q(t_0, 0)}.$$

If $|\partial_\theta q| \leq \overline{B}$ and $|\partial_t q| \geq \underline{B} > 0$ along these nearby roots, then $|F(\theta) - F(0)| \leq (\overline{B}/\underline{B})|\theta|$.

Proof sketch. In (a), the posterior expected utility of $a = 1$ minus that of $a = 0$ under stress δ is $P(\tau_\delta > 0 \mid D) - \alpha$, so $a_\delta^* = 1$ exactly when this probability exceeds α. In (b), implicit differentiation of $q(F(\theta), \theta) = \alpha$ gives $F' = -\partial_\theta q / \partial_t q$, with $\partial_\theta q(t, \theta) = -f_t(\theta)$. Integration gives the local bound. Appendix O gives details. □

The local coefficient $F'(0)$ involves both the posterior density at the effect threshold and the slope of the fragility curve. Decisions that maximize the expected net benefit $\tau - c$ at a treatment cost c, rather than compare τ with a threshold, are treated in Appendix K.

Proposition 2 (Composition upper bound)**.** *Suppose two stress families Δ_1 and Δ_2, with maps $T_{\delta_1}^{(1)}$ and $T_{\delta_2}^{(2)}$, are applied together as the joint family $\Delta = \Delta_1 \times \Delta_2$, whose stressed effect is τ_{δ_1,δ_2}.*

Assume the joint family contains both single-stress axes, $T_{\delta_1,0} = T^{(1)}_{\delta_1}$ and $T_{0,\delta_2} = T^{(2)}_{\delta_2}$, and measure joint size by the sup-norm $\|(\delta_1, \delta_2)\|_\infty = \max(\|\delta_1\|, \|\delta_2\|)$. Then

$$F_\alpha(D; \Delta) \;\le\; \min\big(F_\alpha(D; \Delta_1), F_\alpha(D; \Delta_2)\big).$$

If some joint stress with sup-norm below both single-stress scores already gives $P(\tau_{\delta_1,\delta_2} > 0 \mid D) \le \alpha$, the inequality is strict.

Proof sketch. The joint stress family contains each single-stress family as a coordinate axis. Therefore any single stress that flips the claim is also feasible for the joint problem, giving the stated upper bound. Strictness occurs when a joint stress off the axes reaches the decision boundary before either axis alone. Full details are in Appendix O. □

Single-stress scores thus upper-bound the joint score: reporting them alone can only overstate joint robustness, and the bound is informative when the two scores are comparable. On IHDP, after dividing γ and λ by the upper ends of their search ranges (0.95 and 6) so that both lie in $[0, 1]$, the joint score is up to about 30% smaller than the smaller single-stress score (Appendix D).

Theorem 1 (Calibration to the Cinelli–Hazlett robustness value in the linear-Gaussian case)**.** *In the OLS regression of Y on the design matrix $Z = [\mathbf{1}, A, X]$, assumed to have full column rank, let $\nu = n - \mathrm{rank}(Z) > 0$ be the residual degrees of freedom, $\hat\tau > 0$ the coefficient of A, $\widehat{\mathrm{se}}$ its homoskedastic standard error, and $s^2 = \mathrm{RSS}_{Y|A,X}/\mathrm{RSS}_{A|X} > 0$ the ratio of the residual sums of squares from the regressions of Y on (A, X) and of A on X, so that s is the scale $s(D)$ of Section 4.1. Suppose the posterior 0.5-quantile of τ equals $\hat\tau$, as in the flat-prior Gaussian linear model, and let RV^{CH} denote the robustness value of Cinelli and Hazlett [3] for reducing the estimate to zero, that is, the partial R^2 that a confounder must have with both the treatment and the outcome to bring the adjusted estimate to zero. Then*

$$F_{0.5} = \sqrt{RV^{\mathrm{CH}}}.$$

For $\hat\tau < 0$, the same result applies to the negative claim with the stress in the opposite direction; $\hat\tau = 0$ is a no-claim boundary.

Proof sketch. At $\alpha = 0.5$ the claim flips when the bias reaches the posterior median $\hat\tau$, that is, when $s\gamma^2/\sqrt{1-\gamma^2} = \hat\tau$. With $u = \gamma^2$ and the t-statistic $\hat t = \hat\tau/\widehat{\mathrm{se}}$, the OLS variance identity $\widehat{\mathrm{se}}^2 = s^2/\nu$ gives $u^2/(1-u) = \hat t^2/\nu$, the defining equation of RV^{CH}. The inverse is unique (Appendix O). □

Remark 1. Empirically (Section 6.1, Figure 1, Table 1, and Appendix B), the Bayesian linear posterior reproduces this identity to a root-mean-square (RMS) discrepancy of 0.0010 on IHDP and 0.0010 on ACIC-cov+Hill. Beyond this anchor, the score is also defined for non-Gaussian posteriors such as BART and for the decision rules and composed stresses of Propositions 1–2, which have no classical counterpart.

Proposition 3 (Precision and fragility decouple)**.** *Consider data D_n of size n whose posterior is $\tau \mid D_n \sim N(\mu, \sigma_n^2)$, with $\mu > 0$ fixed and $\sigma_n \downarrow 0$, and hold the bias scale $s = s(D_n) > 0$ of Section 4.1 fixed. Then, with Φ the standard normal distribution function, the width of the posterior 95% credible interval (CI) is $2\Phi^{-1}(0.975)\sigma_n \approx 3.92\sigma_n \downarrow 0$ and $P(\tau > 0 \mid D_n) \to 1$, while, for the hidden-confounding stress, $F_{0.5}(D_n) = F^* > 0$ for every n, where $F^* = h(\mu/s)$ is the unique solution of $\gamma^2/\sqrt{1-\gamma^2} = \mu/s$ and $h(z) = \sqrt{2z/(\sqrt{z^2+4}+z)}$. In contrast, for $\mu_n \downarrow 0$ with $\sigma_n = o(\mu_n)$, $P(\tau > 0 \mid D_n) \to 1$ and the CI width tends to zero, while $F_{0.5}(D_n) = (\mu_n/s)^{1/2}\{1 + o(1)\} \to 0$.*

In the first case the posterior concentrates while the flip distance stays bounded away from zero. In the second, the posterior probability of the claim tends to one while the flip distance vanishes, so improving precision coexists with an arbitrarily fragile conclusion.

6 Experiments

We evaluate BCS on three benchmarks with real covariates and known or simulated potential outcomes: IHDP, TWINS, and ACIC-cov+Hill. We analyze each with up to four posterior generators,

Bayesian models (estimators, in the tables) that produce the draws $\tau^{(\ell)}$; TWINS omits the BCF-style generator. We report $F_{0.5}$ alongside $\sqrt{RV^{\mathrm{CH}}}$; for reference, approximate E-values [2] computed from the standardized effects range from 1.49 to 10.43 over the ten main IHDP configurations defined below.

Datasets. IHDP [8]: Infant Health and Development Program covariates (747 units, 25 covariates) from a randomized intervention for low-birth-weight infants, with simulated outcomes. The benchmark comes in replications, each an independently simulated set of outcomes on the same covariates, and each replication is used in an *original* and an *attenuated 10%* setting; the latter multiplies the individual treatment-effect component by 0.1 before restandardizing the outcome, so it does not preserve the noise-to-signal ratio. We call each replication in one setting a configuration. The main comparison of generators uses five replications (10 configurations); across them, mean standardized true effects in the original and attenuated settings are 1.676 and 0.219, respectively. The target-shift and posterior-predictive-scale checks use replications 1–10 (Appendix A). TWINS [11]: $4{,}821$ same-sex twin pairs (complete records, both twins under 2 kg). Because mortality is recorded for both twins, both potential outcomes are known. The treatment is being the heavier twin and the outcome is mortality, so the claim of interest is negative, $\tau < 0$. Each observational subsample keeps one twin from each of $4{,}000$ pairs, chosen with a covariate-dependent probability. ACIC-cov+Hill (ACIC): covariate matrix from the 2016 Atlantic Causal Inference Conference competition ($4{,}802$ units, 79 covariates after one-hot encoding; each analysis uses a random subsample of $2{,}000$ units); we overlay a nonlinear outcome surface in the style of Hill [8] and use two constant treatment effects (large and small), each with five random seeds (10 configurations, one per effect size and seed).

Posterior generators. We use (i) a Bayesian linear regression pooled over both arms, with a ridge-stabilized normal coefficient approximation and plug-in residual variance; (ii) Bayesian-bootstrap AIPW with ridge nuisance models; (iii) a BART T-learner with two BART models, one per arm; and (iv) a BCF-style model, a simplified two-surface BART with separate prognostic and treatment-effect surfaces and the propensity score as an auxiliary feature in the prognostic surface, inspired by Hahn et al. [9] but using a linear-probability (OLS) propensity rather than logistic regression and without the half-Cauchy variance prior of the original, hence "BCF-style" rather than BCF. The main tree fits use one chain, 50 trees per BART model (50 prognostic and 30 effect trees for BCF-style), and 100–200 tuning/draw iterations depending on the experiment. Appendix F reports a limited two-chain BART check.

6.1 Agreement with $\sqrt{RV^{\mathrm{CH}}}$ for the linear posterior

Figure 1 plots $F_{0.5}$ against $\sqrt{RV^{\mathrm{CH}}}$ for the four posterior generators on IHDP and ACIC-cov+Hill. The dashed line is the Theorem 1 reference. On the linear posterior, $F_{0.5}$ matches $\sqrt{RV^{\mathrm{CH}}}$ to RMS 0.0010 on IHDP and 0.0010 on ACIC, with correlation above 0.99997 on both. AIPW has RMS discrepancies 0.0128 (IHDP) and 0.0257 (ACIC). BART and BCF-style show larger differences, particularly for small effects. BART has RMS 0.040 (IHDP) and 0.044 (ACIC); BCF-style has 0.025 and 0.036. These discrepancies reflect differences in posterior medians under the common scale $s(D)$. Full numbers are in Appendix B, and the ACIC-cov+Hill rows in Appendix C. Appendix H repeats the comparison with a bias scale computed from posterior-predictive residuals.

6.2 Credible but fragile on IHDP

Figure 2 shows a separate illustrative BART fit for IHDP replication 1 in the attenuated setting. The BART ATE posterior is concentrated entirely above zero with $P(\tau > 0 \mid D) = 1.000$, but a hidden-confounding stress of $\gamma = 0.318$ drives the posterior probability of a positive effect to 0.5, that is, $F_{0.5} = 0.318$. The main comparison fits the same configuration with a different seed and sampler length (seed 1; 200 tuning steps and 200 draws) and gives $F_{0.5} = 0.321$. A conventional report shows the posterior alone; BCS adds its fragility curve.

6.3 Posterior precision is not posterior fragility

Figure 3 plots $F_{0.5}$ against the 95% CI width across all 10 configurations and the four posterior generators. The two horizontal clusters (top: original, bottom: attenuated) show that $F_{0.5}$ separates

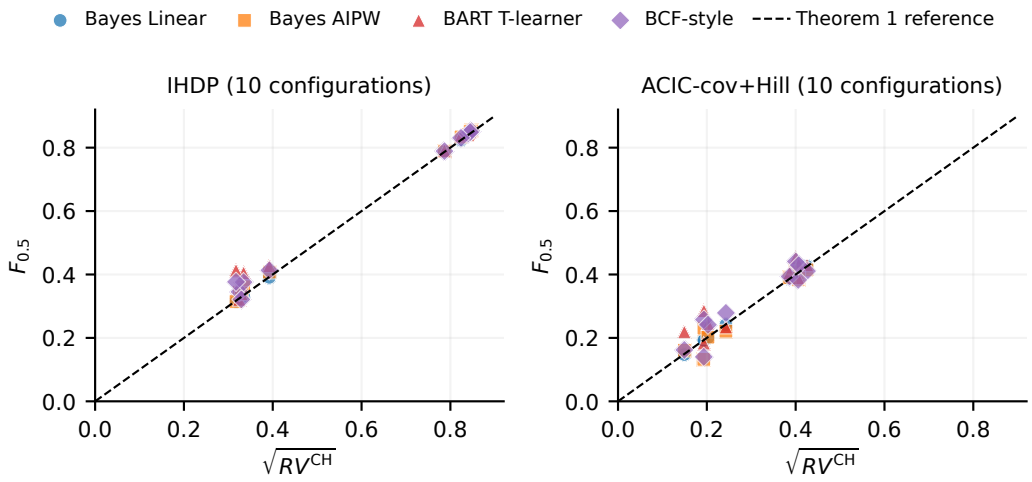

Figure 1: $F_{0.5}$ versus $\sqrt{RV^{\mathrm{CH}}}$ on IHDP (left) and ACIC-cov+Hill (right). The linear posterior follows the Theorem 1 reference; AIPW, BART, and BCF-style have model-dependent discrepancies.

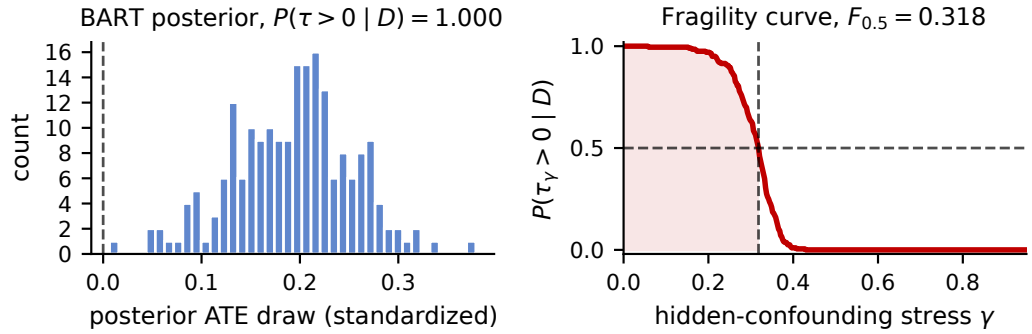

Figure 2: Credible but fragile: separate illustrative BART fit (IHDP replication 1, attenuated setting; 50 trees, 150 tuning steps, 200 draws, seed 11). Left: posterior of the ATE; right: fragility curve $q(\gamma; D)$ under the hidden-confounding stress.

the two settings, while posterior CI widths vary visibly across generators. Averaged over replications within each generator (Table 1), $F_{0.5}$ varies by only 0.008 in the original setting and 0.042 in the attenuated setting.

6.4 Target-shift fragility on IHDP

We take the binary IHDP covariate x_{15} as the subgroup indicator H of Section 4.2 (subgroup share 14.1%). The true difference between the subgroup effects varies across replications and settings. Figure 4 shows the curves $\lambda \mapsto P(\tau_\lambda > 0 \mid D)$ for four representative configurations. Replication 9 attenuated has $G_{0.5} = 0.70$: a logit shift of 0.70 toward $H = 1$ overturns a positive-ATE claim whose unstressed support is only $P(\tau > 0 \mid D) = 0.578$. Replication 9 original has $G_{0.5} = 1.54$. Table 2 lists the configurations with finite $G_{0.5}$.

6.5 TWINS

Table 3 reports TWINS aggregates for the negative claim $\tau < 0$, for which the hidden-confounding stress adds the bias. The mean paired ATE across the five observational subsamples is -0.023 in raw mortality-probability units; posterior effect summaries in the table use standardized outcomes. Across the five subsamples and three generators, the posterior probability of the negative claim

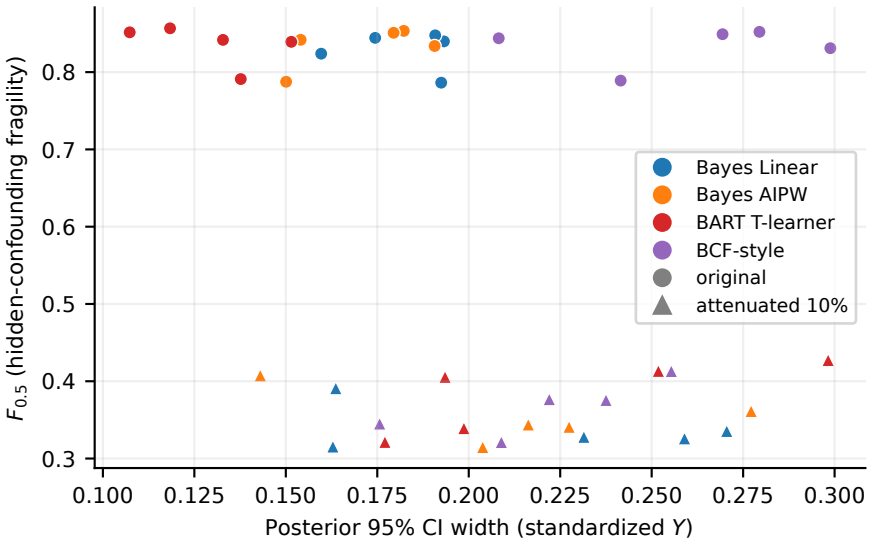

Figure 3: IHDP across all 10 configurations. Different posterior generators give visibly different CI widths while $F_{0.5}$ remains clustered by setting (original versus attenuated). Posterior precision and posterior fragility carry different information.

Table 1: IHDP main results by posterior generator and setting, averaged over the five replications: posterior mean of the ATE ($\bar{\tau}_{\text{post}}$, in standardized Y units), 95% CI width, posterior probability of a positive ATE, $F_{0.5}$ (mean and standard deviation across replications), and $\sqrt{RV^{\text{CH}}}$. The linear posterior follows the Theorem 1 reference; other generators have model-dependent deviations.

estimator	setting	$\bar{\tau}_{\text{post}}$	95% CI width	$P(\tau > 0 \mid D)$	$F_{0.5}$ (mean±sd)	$\sqrt{RV^{\text{CH}}}$
BART T-learner	original	1.677	0.130	1.000	0.836±0.026	0.828
BCF-style	original	1.658	0.259	1.000	0.833±0.026	0.828
Bayes AIPW	original	1.657	0.171	1.000	0.833±0.027	0.828
Bayes Linear	original	1.613	0.182	1.000	0.828±0.025	0.828
BART T-learner	attenuated 10%	0.240	0.224	1.000	0.381±0.048	0.340
BCF-style	attenuated 10%	0.226	0.220	0.998	0.366±0.035	0.340
Bayes AIPW	attenuated 10%	0.209	0.214	0.998	0.354±0.034	0.340
Bayes Linear	attenuated 10%	0.189	0.217	0.999	0.339±0.030	0.340

ranges from 0.863 to 1.000. The generator-level mean fragility scores are 0.176–0.186, close to $\sqrt{RV^{\text{CH}}} = 0.177$; Table 3 reports the corresponding seed-level $F_{0.5}$ ranges in brackets.

Table 3: TWINS aggregates (4,821 eligible pairs; 4,000 per observational subsample, five seeds). $\bar{\tau}_{\text{pair}}$ is the benchmark ATE from the paired potential outcomes, in raw mortality-probability units; the posterior mean $\bar{\tau}_{\text{post}}$ and CI widths are standardized. $P(\tau < 0 \mid D)$ is the posterior probability of the negative claim; means are over seeds and brackets give seed ranges.

estimator	$\bar{\tau}_{\text{pair}}$	$\bar{\tau}_{\text{post}}$ (std.)	95% CI width	$P(\tau < 0 \mid D)$ mean [range]	$F_{0.5}$ mean±sd [range]	$\sqrt{RV^{\text{CH}}}$
Bayes Linear	-0.0231	-0.0650	0.120	0.958 [0.872,1.000]	0.178±0.037 [0.129,0.220]	0.177
Bayes AIPW	-0.0231	-0.0627	0.120	0.953 [0.863,0.998]	0.176±0.032 [0.134,0.211]	0.177
BART T-learner	-0.0231	-0.0685	0.107	0.978 [0.960,1.000]	0.186±0.024 [0.174,0.230]	0.177

6.6 Decision, composition, and calibration checks

Figure 8 summarizes three checks; Appendices D, K, and G give the details. (1) Decision: when the action maximizes the posterior expected net benefit $\tau - c$ at treatment cost c, the stress at which the action switches (the mean-utility boundary) decreases with the cost and approximately agrees

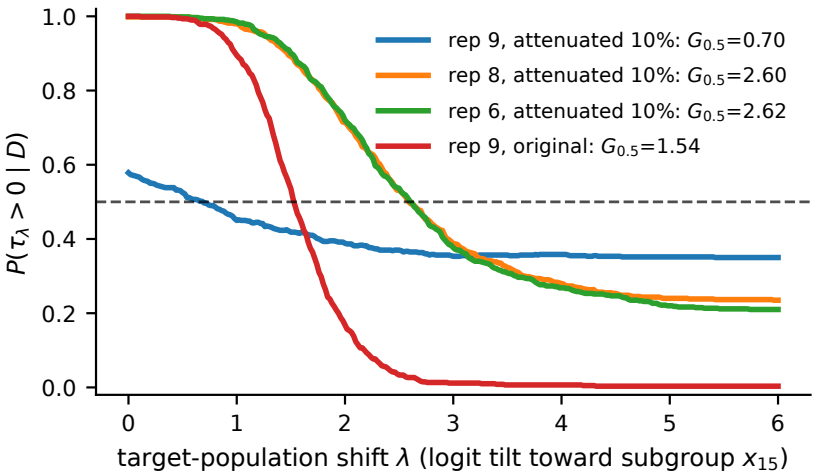

Figure 4: Target-shift fragility curves on IHDP using subgroup x_{15}. The curves show how the posterior probability of a positive ATE moves as the log-odds of membership in $H = 1$ increase by λ.

Table 2: Target-shift fragility on IHDP (subgroup x_{15}): replication (rep), setting, posterior mean and probability of a positive ATE, $F_{0.5}$, and $G_{0.5}$ for shifts toward $H = 1$; configurations with finite $G_{0.5}$, sorted by $G_{0.5}$.

rep	setting	$\bar{\tau}_{\text{post}}$	$P(\tau > 0 \mid D)$	$F_{0.5}$ (hidden)	$G_{0.5}$ (target $\to H{=}1$)
9	attenuated 10%	0.007	0.578	0.085	0.70
9	original	0.193	1.000	0.352	1.54
8	attenuated 10%	0.171	0.998	0.332	2.60
6	attenuated 10%	0.233	1.000	0.353	2.62
7	attenuated 10%	0.196	0.997	0.289	2.70

with the median boundary $F_{0.5}$ at zero cost. (2) Composition: with all radii read on one grid, the normalized joint stress agrees with Proposition 2 in all 18 configurations, with reductions of up to about 30%. (3) Calibration: in data simulated from a specified data-generating process (DGP) with a Gaussian confounder, we compare $F_{0.5}$ with the population flip-stress F^*, the stress that flips the claim in the population implied by the DGP parameters. With independent seeds at every grid point, over the simulations whose sign claim is supported (90 of 96 zero-effect simulations and 119 of 120 on the nonzero-effect grid, which also includes $\tau_{\text{true}} = 0$), the mean absolute discrepancies from the population flip-stress are 0.0328 and 0.0182 and the mean signed discrepancies -0.0009 and -0.0023; the remaining simulations are no-claim (Appendix G, Figure 6). Inverting the bias at the posterior mean instead of the median changes the score by only 0.0006.

6.7 Real-data audits beyond the benchmarks

Large-scale ACSIncome audit. We analyze 246,139 ACSIncome person records (American Community Survey Public Use Microdata Sample, PUMS; Appendix N, Table 13), with a bachelor's or higher degree as exposure and the standardized asinh of nonnegative total personal income as outcome. With all covariates, the adjusted posterior sign claim is unsupported ($P(C \mid D) = 0.4996$), so the diagnostic reports *no-claim*. Excluding income-related variables yields a supported claim ($F_{0.5} = 0.279$); the income+labor exclusion, which also removes labor-market variables, yields one as well ($F_{0.5} = 0.303$). Each set is capped at 60 features, so these are non-nested covariate-set comparisons, not evidence of a causal education effect.

Benchmark dashboard. To put a flip distance in context, Appendix M (Table 12) compares $F_{0.5}$ with b_{ref}, the largest residual association of a single observed covariate with both treatment and outcome, and reports the ratio $F_{0.5}/b_{\text{ref}}$ for supported claims. The reference ratios are descriptive, not

calibrated bounds on omitted-confounder strength: 1.37 (TWINS), 2.22 (IHDP attenuated), and 3.01 (NHEFS, the NHANES I Epidemiologic Follow-up Study; smoking cessation and weight change, Appendix L, Figure 9). LaLonde, a job-training evaluation with Current Population Survey (CPS) or National Supported Work (NSW) experimental controls, has flip distances 0.100 (CPS controls) and 0.347 (NSW controls) (Appendix J, Figure 7); we do not report reference ratios for these rows. The ACSIncome ratios are 2.43 and 2.62 (Table 13).

7 Discussion, limitations, and conclusion

What the score adds. A credible interval answers how uncertain τ is under the model M; the fragility score answers how far M or the target population must move before the conclusion changes. On IHDP the posterior generators differ about twofold in interval width yet agree on $F_{0.5}$ to within 0.01 (Section 6.3), because the score is set by the posterior median and the scale $s(D)$, not by the model's own uncertainty, and Proposition 3 gives posteriors whose probability of the claim tends to one while the flip distance vanishes; a narrow interval is therefore not a robustness statement. The score is computed from posterior draws without refitting, applies to any generator and any claim, and is read on a known scale: Theorem 1 ties γ to the Cinelli–Hazlett robustness value, and Proposition 1 reads F_α as the smallest stress that changes the Bayes action.

What to report. We recommend reporting the posterior mean, the 95% CI, $P(C \mid D)$, $F_{0.5}$ with its fragility curve, covariate benchmarks such as b_{ref}, and target-shift diagnostics by subgroup (Appendix I).

Limitations. The hidden-confounding stress is symmetric in the two partial correlations and represents a single omitted confounder; extensions to several confounders, to minimax summaries over sets of stresses, and to a fully Bayesian variant with a prior on the stress are future work. TWINS posterior summaries are on the standardized outcome scale rather than in mortality units. Because $F_{0.5}$ depends on the posterior only through its median, it cannot isolate higher moments (Appendix E).

Conclusion. A Bayesian causal analysis reports uncertainty inside its model; BCS reports how much of the model the conclusion rests on. The two belong together: the posterior with its interval, and the flip distance with its fragility curve.

Data availability

All data are existing public data; none were collected for this study. IHDP is the semi-synthetic benchmark of Hill [8] (Infant Health and Development Program covariates, simulated outcomes). TWINS is the same-sex twin-birth benchmark of Louizos et al. [11], built from the 1989–1991 U.S. linked birth and infant death records analyzed by Almond et al. [12]. The IHDP and TWINS files are from the GitHub repository `AMLab-Amsterdam/CEVAE` accompanying Louizos et al. [11]. ACIC-cov+Hill uses the covariate matrix of the 2016 ACIC competition [13] from the repository `vdorie/aciccomp`; its outcomes are simulated here. LaLonde uses the National Supported Work experimental sample of LaLonde [14] and the CPS comparison group, both in the form of Dehejia and Wahba [15]. NHEFS is the NHANES I Epidemiologic Follow-up Study extract of Hernán and Robins [16]; both come through the `causaldata` Python package. ACSIncome consists of 2018 one-year American Community Survey PUMS records prepared for the ACSIncome task of Ding et al. [17]; the analysis uses the $246{,}139$ complete records among the first $300{,}000$ rows of the public Hugging Face copy `davidboetius/ACSIncome-2018-1-Year`.

Acknowledgments and Disclosure of Funding

This research was funded by JSPS KAKENHI under grant number JP25K00626.

References

[1] Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed. New York: Springer.

[2] VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine* 167(4):268–274.

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

[4] McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine* 26(11):2331–2347.

[5] Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine* 35(20):3453–3470.

[6] Hebdon, R., Stamey, J., Kahle, D., and Zhang, X. (2024). `unmconf`: Modeling with Unmeasured Confounding. R package version 1.0.0. `https://doi.org/10.32614/CRAN.package.unmconf`.

[7] Veitch, V. and Zaveri, A. (2020). Sense and sensitivity analysis: simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems* 33.

[8] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1):217–240.

[9] Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis* 15(3):965–1056.

[10] Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* 9(1):130–134.

[11] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems* 30.

[12] Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics* 120(3):1031–1083.

[13] Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Statistical Science* 34(1):43–68.

[14] LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76(4):604–620.

[15] Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448):1053–1062.

[16] Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

[17] Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring Adult: new datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34.

A Computational details

IHDP, TWINS, and ACIC posterior effects use standardized outcomes; LaLonde and NHEFS effect tables use dollars and kilograms. Hidden-confounding estimates invert interpolated sample quantiles to approximate the score of the exact posterior. Reported probabilities are finite-draw estimates. For the strict event $\tau > b$, the exact empirical boundary is the ascending order statistic $\tau_{(k)}$, $k = \lceil S(1-\alpha) \rceil$; interpolation approximates it. Target-shift searches use $\lambda \in [0, 6]$; > 6 means that no crossing was detected on the evaluated grid. The main IHDP comparison has five replications in two settings; the target-shift and posterior-predictive-scale checks use replications 1–10, and the composition and multi-subgroup checks use replications 1–9. The target-shift fits of Section 6.4 use 600 draws. Tree fits use one chain: BART uses 200 tuning and 200 retained draws on IHDP and 100 and 100 on TWINS and ACIC, BCF-style uses 100 and 100, and the separate fits of Figure 2 and Appendix E use 150 and 200. The interaction regressions of Sections 4.2 and 6.4 and Appendices D and I regress the standardized outcome on an intercept, A, H, AH, and the remaining covariates as standardized linear terms, with a small ridge term; the bias scale $s(D)$ used with them conditions on all covariates.

B Theorem 1 numerical calibration on IHDP and ACIC-cov+Hill

Table 4: Discrepancy between $F_{0.5}$ and $\sqrt{RV^{\mathrm{CH}}}$ by posterior generator over the ten main configurations of each benchmark; $\sqrt{RV}$ is short for $\sqrt{RV^{\mathrm{CH}}}$.

benchmark	estimator	mean$(F - \sqrt{RV})$	RMS$(F - \sqrt{RV})$	max$\lvert F - \sqrt{RV}\rvert$	$\rho(F, \sqrt{RV})$
IHDP	Bayes Linear	-0.0003	0.0010	0.0018	1.0000
IHDP	Bayes AIPW	+0.0096	0.0128	0.0270	0.9995
IHDP	BART T-learner	+0.0247	0.0400	0.0961	0.9934
IHDP	BCF-style	+0.0157	0.0250	0.0595	0.9977
ACIC	Bayes Linear	-0.0006	0.0010	0.0018	1.0000
ACIC	Bayes AIPW	-0.0047	0.0257	0.0608	0.9733
ACIC	BART T-learner	+0.0272	0.0441	0.0923	0.9459
ACIC	BCF-style	+0.0132	0.0359	0.0645	0.9509

Table 4 shows that, across both benchmarks, the Bayesian linear posterior has an RMS discrepancy of approximately 10^{-3} from the RV reference. AIPW deviates mildly. BART deviates the most, with mean $F - \sqrt{RV} = +0.025$ on IHDP and $+0.027$ on ACIC and RMS 0.040 and 0.044; BCF-style deviates less (RMS 0.025 on IHDP, 0.036 on ACIC).

C ACIC-cov+Hill main results

Table 5: ACIC-cov+Hill results by posterior generator and effect size, averaged over five random seeds ($n = 2{,}000$ subsample of the $4{,}802$-row covariate matrix).

estimator	effect	$\bar{\tau}_{\text{true}}$	$\bar{\tau}_{\text{post}}$	95% CI width	$P(\tau > 0 \mid D)$	$F_{0.5}$ (mean±sd)	$\sqrt{RV^{\mathrm{CH}}}$
BART T-learner	large	0.222	0.233	0.120	1.000	0.422±0.025	0.405
BCF-style	large	0.222	0.223	0.126	1.000	0.412±0.024	0.405
Bayes AIPW	large	0.222	0.212	0.112	1.000	0.402±0.015	0.405
Bayes Linear	large	0.222	0.215	0.107	1.000	0.405±0.014	0.405
BART T-learner	small	0.056	0.070	0.128	0.956	0.234±0.037	0.196
BCF-style	small	0.056	0.062	0.131	0.920	0.216±0.061	0.196
Bayes AIPW	small	0.056	0.046	0.113	0.921	0.190±0.042	0.196
Bayes Linear	small	0.056	0.049	0.106	0.943	0.195±0.034	0.196

Table 5 reports the results. The linear posterior closely matches the Theorem 1 reference; AIPW is a comparison method, not covered by that identity; BART and BCF-style deviate upward for the small effect.

D Proposition 2 numerical check on IHDP

Table 6: Single-stress and joint stress radii on IHDP, replications 1–9. The radii r^F, r^G, and r^{joint} are normalized sup-norm radii read on one grid.

rep	setting	$F_{0.5}$	$G_{0.5}$	r^F	r^G	r^{joint}
1	orig	0.840	> 6	0.884	> 1	0.836
1	atten 10%	0.322	> 6	0.342	> 1	0.320
2	orig	0.848	> 6	0.895	> 1	0.888
2	atten 10%	0.337	> 6	0.358	> 1	0.356
3	orig	0.844	> 6	0.889	> 1	0.888
3	atten 10%	0.324	> 6	0.342	> 1	0.340
4	orig	0.824	> 6	0.868	> 1	0.808
4	atten 10%	0.317	> 6	0.337	> 1	0.336
5	orig	0.786	> 6	0.832	> 1	0.830
5	atten 10%	0.390	> 6	0.411	> 1	0.382
6	orig	0.853	> 6	0.900	> 1	0.846
6	atten 10%	0.353	2.62	0.374	0.438	0.262
7	orig	0.857	> 6	0.905	> 1	0.882
7	atten 10%	0.289	2.70	0.305	0.450	0.228
8	orig	0.805	> 6	0.847	> 1	0.782
8	atten 10%	0.332	2.60	0.353	0.433	0.248
9	orig	0.352	1.54	0.374	0.258	0.198
9	atten 10%	0.085	0.70	0.095	0.117	0.066

Table 6 covers IHDP replications 1–9 in both settings (all 18 configurations support the positive claim). The joint stress applies both stress maps, $\tau^{(\ell)}_{\gamma,\lambda} = \tau^{(\ell)}_{\lambda} - b(\gamma; D)$. We compute the joint posterior probability $P(\tau_{\gamma,\lambda} > 0 \mid D)$ on a grid with γ in steps of 0.005 on $[0, 0.95]$ and λ in steps of 0.025 on $[0, 6]$, and the joint fragility in the normalized sup-norm, $r^{\text{joint}} = \inf\{\max(\gamma/0.95, \lambda/6) : P(\tau_{\gamma,\lambda} > 0 \mid D) \leq 0.5\}$, in radius steps of 0.002. The radii r^F and r^G are the first crossings on the two axes of the same grid, so r^F can differ from $F_{0.5}/0.95$ by the grid resolution; $F_{0.5}$ and $G_{0.5}$ are the scores of Section 4. Proposition 2 predicts $r^{\text{joint}} \leq \min(r^F, r^G)$. On the common grid this holds in $18/18$ configurations. The reduction is small when one single-stress score dominates and reaches about 30% when both stresses are active (rep 9 attenuated: 0.066 versus 0.095).

E Posterior-shape diagnostic

Figure 5 compares the four posterior generators on IHDP replication 1 in the attenuated setting.

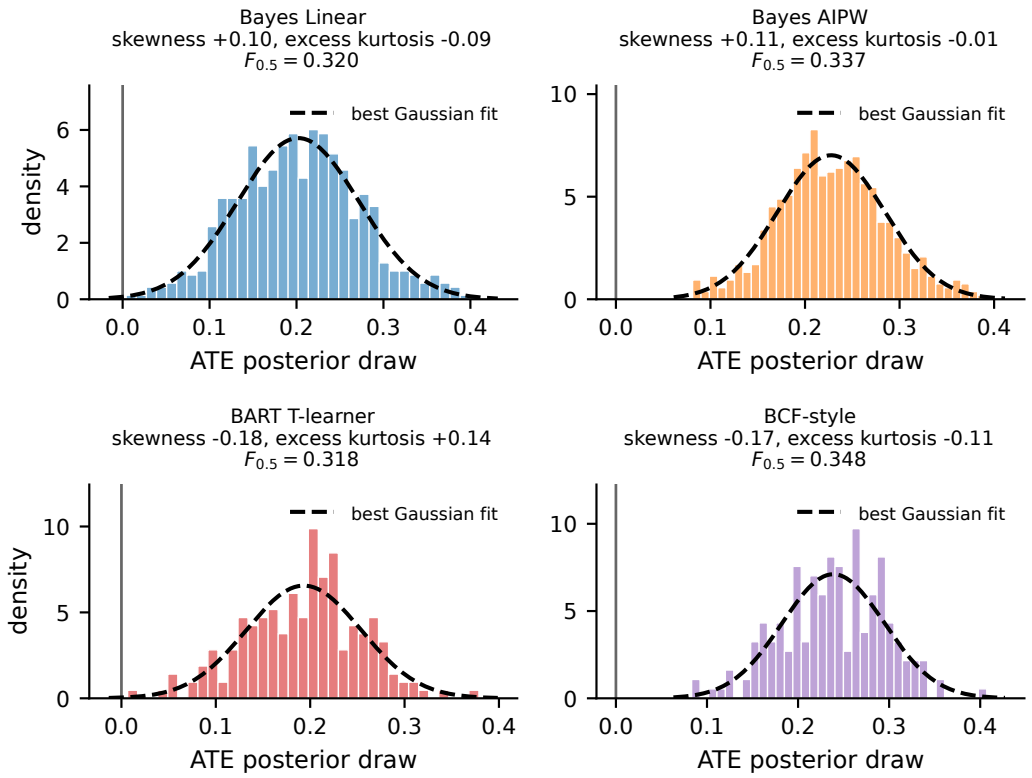

Figure 5: Posterior densities and Gaussian fits on IHDP replication 1, attenuated setting. Linear and AIPW are nearly Gaussian (absolute skewness below 0.12, absolute excess kurtosis below 0.10); BART and BCF-style are left-skewed (skewness -0.18 and -0.17). The Theorem 1 prediction $F_{0.5} \approx \sqrt{RV^{\mathrm{CH}}}$ is recovered in the linear-Gaussian flat-prior limit; these density summaries illustrate distributional differences, but higher moments alone do not determine the median-flip score.

F Convergence diagnostics and draw-count sensitivity

The Bayesian linear posterior draws normal coefficients using a plug-in residual variance. Its draws are independent, so we report draw-count sensitivity rather than Markov-chain diagnostics.

Table 7: Bayesian linear posterior: $F_{0.5}$ at 200 and 800 analytic draws on six IHDP configurations.

rep	setting	$F_{0.5}^{(200)}$	$F_{0.5}^{(800)}$	$\lvert F_{0.5}^{(200)} - F_{0.5}^{(800)} \rvert$
1	orig	0.840	0.840	0.0005
1	atten 10%	0.332	0.327	0.0049
2	orig	0.848	0.847	0.0001
2	atten 10%	0.336	0.334	0.0013
3	orig	0.845	0.844	0.0002
3	atten 10%	0.330	0.328	0.0021

Draw-count sensitivity of the Bayesian linear posterior. For the Bayesian linear posterior at 200 versus 800 analytic draws across six representative IHDP configurations (Table 7), $|F_{0.5}^{(200)} - F_{0.5}^{(800)}| \leq 0.005$.

BART between-chain check. On IHDP replications 1 and 2 in both settings, we fit two independent BART chains per configuration with different random seeds (each: 200 draws after 80 tuning steps, 40 trees). Table 8 reports the split potential scale reduction factor $\hat{R}$ from the two chains, the effective

sample size out of the 400 pooled draws, and the mean and standard deviation of the two chain-level $F_{0.5}$ values. Across the four configurations, $\hat{R} \leq 1.011$, the effective sample size is at least 330, and the chain-level $F_{0.5}$ values have standard deviation at most 0.008. Two short chains give a limited between-chain comparison.

Table 8: BART between-chain check on IHDP replications 1 and 2: split-$\hat{R}$, effective sample size out of 400 pooled draws, and the mean and standard deviation of the two chain-level $F_{0.5}$ values.

rep	setting	$\hat{R}$ (2 chains)	ESS	$F_{0.5}$ (chain mean)	$F_{0.5}$ (chain sd)
1	orig	1.000	330	0.843	0.0005
1	atten 10%	1.011	378	0.315	0.0077
2	orig	1.002	400	0.856	0.0005
2	atten 10%	1.005	355	0.410	0.0041

G Calibration with nonzero causal effects

The simulations use $A = m_A(X) + gU + \sqrt{1-g^2}\,\varepsilon_A$ and $Y = m_Y(X) + \tau_{\text{true}} A + gU + \sqrt{1-g^2}\,\varepsilon_Y$, with $U, \varepsilon_A, \varepsilon_Y$ independent standard normal and independent of X, and linear $m_A(X), m_Y(X)$ in six standardized Gaussian covariates. Thus g is a loading: the population partial correlations are $r_A = g$ and $r_Y = g/\sqrt{1+g^2}$, not a symmetric pair equal to g. The observed-only coefficient is $\tau_{\text{true}} + g^2$, and its bias-calibration scale is $\sqrt{1-g^4}$. With h as in Proposition 3, the population flip-stress is

$$F^*(g, \tau_{\text{true}}) = h\left(\frac{\tau_{\text{true}} + g^2}{\sqrt{1-g^4}}\right).$$

Outcome standardization cancels from this ratio. The zero-effect grid uses $g \in \{0.05, 0.10, \ldots, 0.60\}$, $n = 2{,}000$, and eight seeds per grid point; the nonzero grid uses $\tau_{\text{true}} \in \{0, 0.1, 0.3, 0.5\}$, $g \in \{0.1, \ldots, 0.5\}$, $n = 2{,}000$, and six seeds per grid point. Every grid point has its own seeds, so covariates, confounder, and noise are drawn independently across simulations. Following the no-claim convention of Section 4, simulations whose unstressed posterior does not support the sign claim ($P(C \mid D) \leq 0.5$) are not scored: 6 of the 96 zero-effect simulations (all with $g \leq 0.15$) and 1 of the 120 simulations on the nonzero grid. The 90 scored zero-effect simulations have mean absolute discrepancy 0.0328 from F^* and mean signed discrepancy -0.0009, with 44 of 90 errors positive; $|F_{0.5} - g|$ averages 0.0404, but g is not the exact target. The 119 scored simulations on the nonzero grid have mean absolute discrepancy 0.0182 and mean signed discrepancy -0.0023, with 60 of 119 errors positive (Figure 6). The errors are largest at small g, where F^* is below the seed-to-seed standard deviation of $F_{0.5}$ (0.081 at $g = 0.05$); the unsupported simulations occur there as well. Inverting the same fitted scale at the posterior mean rather than its median gives discrepancy 0.0006 on the scored simulations.

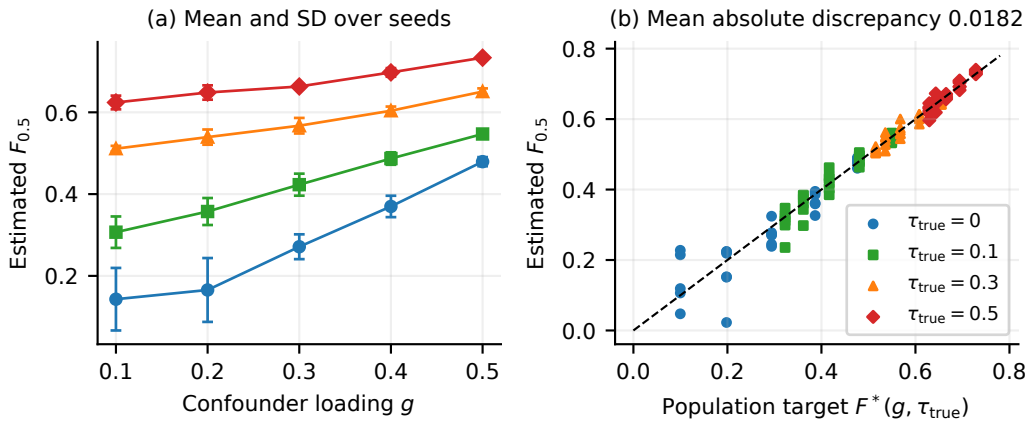

Figure 6: Fitted scores compared with the specified DGP (scored simulations). (a) $F_{0.5}$ across the loading g and true effects, with seed variability. (b) The fitted scores against the population flip-stress F^* (the population target), which uses DGP parameters rather than a fitted posterior mean. Mean absolute discrepancy is 0.0182.

H Theorem 1 calibration check under a posterior-predictive bias scale

The bias scale used in $F_{0.5}$ is $s(D) = \sigma_{Y|A,X}/\sigma_{A|X}$, which can be estimated from OLS residuals (σ^{OLS}) or from posterior-predictive residuals (σ^{PP}). The latter pools residuals under coefficient draws from the same fitted regressions; it is a same-data scale perturbation, not independent validation. Across 20 IHDP configurations, $\sigma^{\text{PP}}/\sigma^{\text{OLS}} = 1.0009 \pm 0.0003$ (mean$\pm$sd). For the linear posterior, the Theorem 1 calibration check holds under either bias-scale estimator: mean $|F_{0.5} - \sqrt{RV^{\text{CH}}}| = 0.0015$ with the OLS scale and 0.0015 with the posterior-predictive scale; with the OLS scale the RMS is 0.0024 and the maximum 0.0084, larger than in the ten main configurations of Appendix B.

I Multi-subgroup target-shift fragility

We fit an interaction regression separately for each binary subgroup (16 per IHDP replication; replications 1–9) and summarize the subgroup-specific target-shift scores $G_{0.5}$, written G below. These are model-and-target comparisons, not a worst-case population shift from one fixed posterior of the conditional average treatment effect. A subgroup fit whose unstressed posterior does not support the sign claim is counted as no-claim and excluded from the minimum. The x_{15} column uses 400 draws here, versus 600 in Table 2.

Table 9 gives the results. For replication 9 in the attenuated setting, three of the sixteen subgroup fits are no-claim, and the smallest supported distance is 0.26 (subgroup x_{12}). The minimum supported distances range from 0.26 to 3.46, and finite values of $G(x_{15})$ from 0.70 to 2.94; entries > 6 indicate no crossing detected on the search grid (Appendix A).

Table 9: Subgroup-specific target-shift distances on IHDP replications 1–9: number of subgroup fits, number of no-claim fits, the smallest supported distance G (with its covariate), and G for x_{15}. A value > 6 means no crossing was detected on the search grid.

rep	setting	#subgroups	#no-claim	min. G (supported)	covariate	$G(x_{15})$
1	orig	16	0	> 6	—	> 6
1	atten 10%	16	0	2.44	x_{23}	> 6
2	orig	16	0	> 6	—	> 6
2	atten 10%	16	0	1.98	x_{23}	> 6
3	orig	16	0	> 6	—	> 6
3	atten 10%	16	0	3.46	x_8	> 6
4	orig	16	0	> 6	—	> 6
4	atten 10%	16	0	2.38	x_{23}	> 6
5	orig	16	0	> 6	—	> 6
5	atten 10%	16	0	> 6	—	> 6
6	orig	16	0	> 6	—	> 6
6	atten 10%	16	0	1.76	x_{23}	2.62
7	orig	16	0	> 6	—	> 6
7	atten 10%	16	0	1.90	x_{24}	2.94
8	orig	16	0	> 6	—	> 6
8	atten 10%	16	0	2.60	x_{15}	2.60
9	orig	16	0	1.50	x_7	1.54
9	atten 10%	16	3	0.26	x_{12}	0.70

J LaLonde NSW versus CPS: real-world case study

The National Supported Work (NSW) job-training experiment [14], in the subsample of Dehejia and Wahba [15], gives an experimental benchmark for the effect on 1978 earnings. Two observational analyses use the treated NSW units: NSW+CPS takes its controls from the nonexperimental CPS comparison group of Dehejia and Wahba [15], and NSW+NSW takes the experimental NSW controls but analyzes them as if observational (a pseudo-observational analysis).

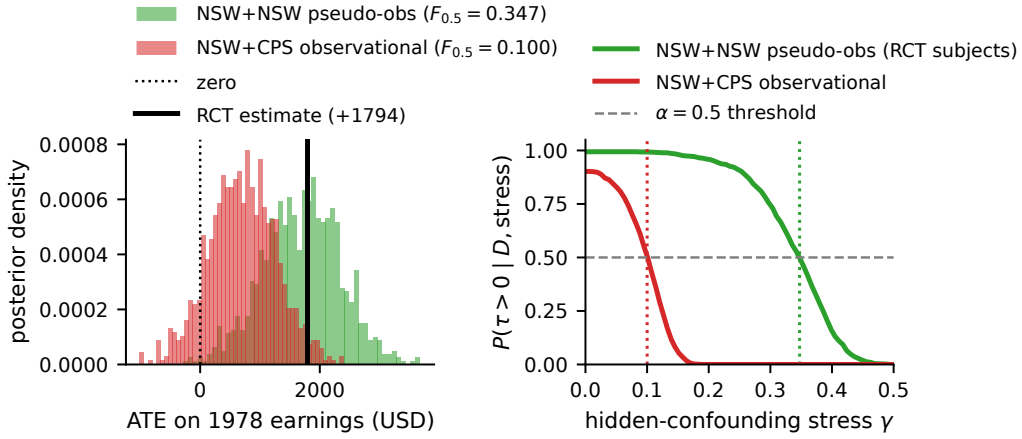

Figure 7: LaLonde NSW versus CPS. Left: posterior densities for two observational analyses with the same treated units. NSW+CPS observational (red) has a lower point estimate than the experimental reference ($1,794, vertical line); the NSW+NSW pseudo-observational posterior (green) covers it. Right: fragility curves. The NSW+CPS claim flips at $\gamma \approx 0.10$; the NSW+NSW claim at $\gamma \approx 0.35$. The flip distances are computed without using the experimental reference as a target.

Figure 7 and Table 10 compare the analyses. The fitted CPS comparison gives a point estimate about $1,100 below the NSW experimental estimate; their reported intervals overlap. The CPS analysis has flip distance $F_{0.5} = 0.100$, compared with 0.347 for the NSW-control analysis. The two analyses differ in stress sensitivity; neither flip distance identifies the bias of the CPS comparison.

Table 10: LaLonde analyses: the experimental estimate and two observational analyses that share the treated units. Effects are in dollars.

analysis	n_{treat}	n_{ctrl}	$\hat{\tau}$ ($)	95% CI ($)	$F_{0.5}$
NSW RCT estimate	185	260	+1794	[+479, +3109]	—
NSW + NSW obs.	185	260	+1692	[+355, +2933]	0.347
NSW + CPS obs.	185	15992	+707	[-429, +1824]	0.100

K Decision, composition, and calibration diagnostics

For the net-benefit utility $\mathcal{U}(1,\tau) = \tau - c, \mathcal{U}(0,\tau) = 0$ at treatment cost c, the mean-utility boundary is the smallest γ at which the Bayes action switches; because expected net benefit depends on the posterior mean, this boundary uses the mean where $F_{0.5}$ uses the median. Figure 8 collects the mean-utility boundary, the zero-effect calibration check, and the joint stress surface of Proposition 2.

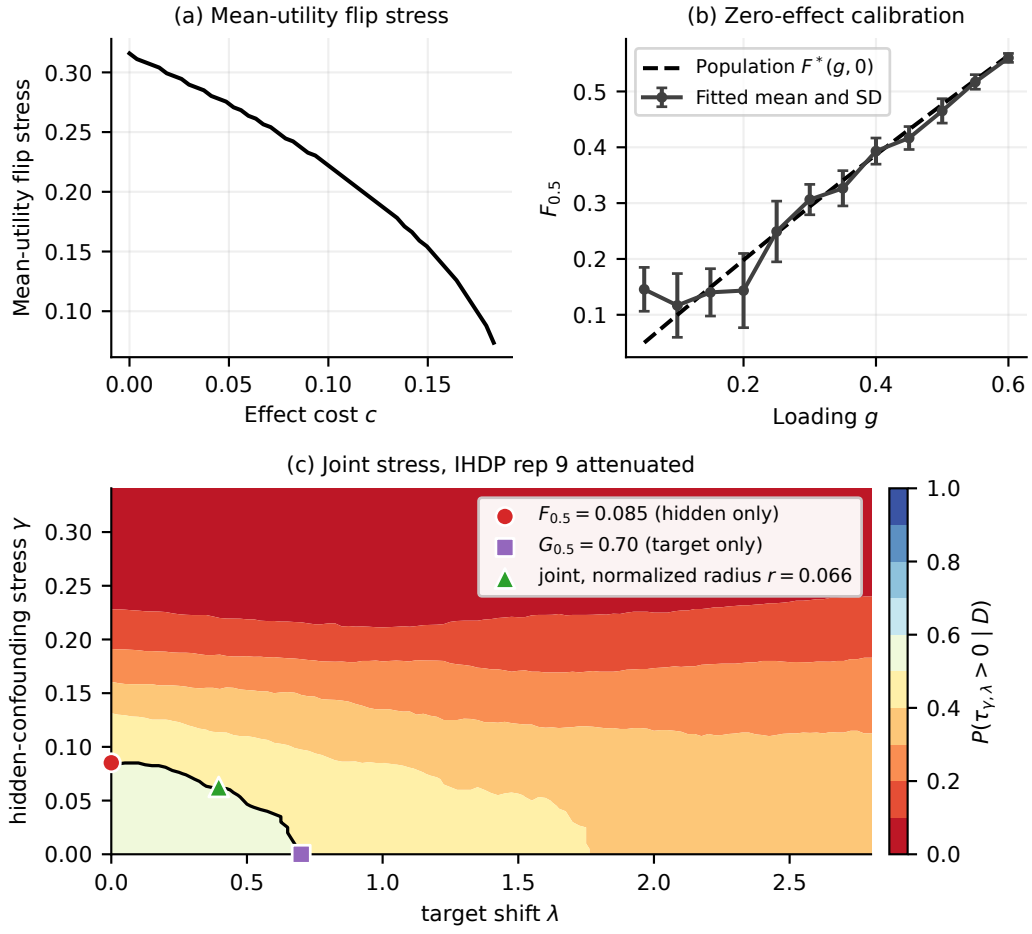

Figure 8: (a) Mean-utility boundary for $\mathcal{U}(1,\tau) = \tau - c$, $\mathcal{U}(0,\tau) = 0$ on IHDP replication 1, attenuated setting. At zero cost it is close to the median boundary; equality requires matching mean and median, unlike the threshold utility in Proposition 1. (b) The 90 scored zero-effect simulations compared with the population target $F^*(g,0)$ from Appendix G; mean absolute discrepancy 0.0328. (c) Proposition 2: joint posterior probability of a positive effect on replication 9, attenuated setting; the single-stress flip points lie on the axes (red, purple) and the joint flip point lies strictly inside (green); the black contour is the 0.5 level.

L NHEFS smoking cessation: real-world textbook case study

NHEFS, the National Health and Nutrition Examination Survey (NHANES) I Epidemiologic Follow-up Study, is the dataset used throughout Hernán and Robins [16]. The causal question is whether quitting smoking between 1971 and 1982 changes body weight over the same period (kg). We follow the textbook adjustment set (complete cases, $n = 1{,}566$): age, sex, race, education, smoking intensity, years of smoking, exercise, physical activity, and 1971 weight.

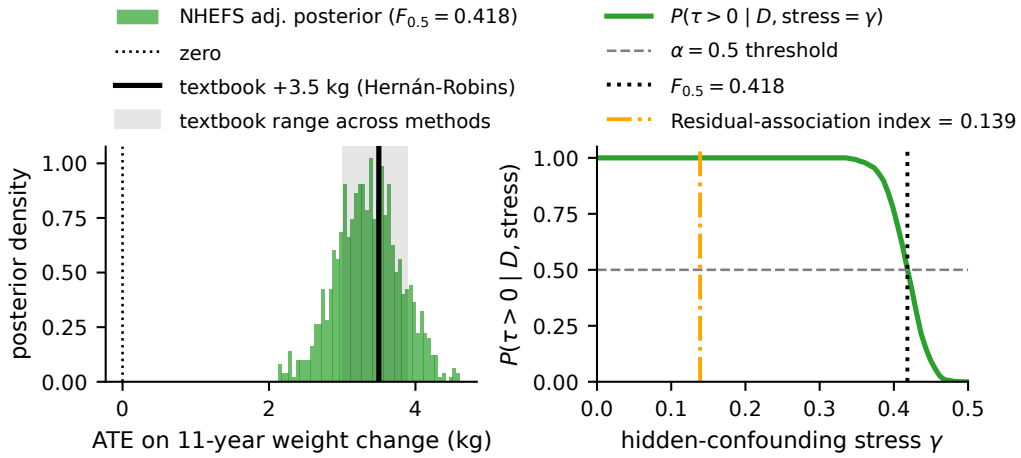

Figure 9: NHEFS smoking cessation → weight gain. Left: posterior of the adjusted Bayesian linear model. Posterior mean is +3.35 kg, in agreement with the textbook adjusted estimates of 3.4–3.6 kg across inverse-probability weighting, standardization, g-estimation, and outcome and propensity-score regression (vertical line at +3.5 kg; the gray band spans 3.0–3.9 kg around them). Right: fragility curve. $F_{0.5} = 0.418$ is about 3.0 times the residual-association reference index ($b_{\text{ref}} = 0.139$, age; Appendix M).

Table 11: NHEFS smoking cessation and weight change: unadjusted and adjusted estimates, fragility score, $\sqrt{RV}$ (short for $\sqrt{RV^{\text{CH}}}$), and the residual-association reference index.

estimator	n	$\hat{\tau}$ (kg)	95% CI (kg)	$F_{0.5}$	$\sqrt{RV}$	b_{ref}
naive (no adj.)	1566	+2.54	[+1.59, +3.50]	—	—	—
adjusted (linear)	1566	+3.35	[+2.48, +4.21]	0.418	0.418	0.139

Figure 9 and Table 11 report the analysis. The naive estimate is +2.54 kg and the adjusted estimate is +3.35 kg; only the latter is in line with the textbook adjusted estimates of 3.4–3.6 kg. The positive sign claim flips at the hidden-confounding stress $\gamma = 0.418$, that is, $F_{0.5} = 0.418$. The reference index is $b_{\text{ref}} = 0.139$ (Appendix M).

M Dashboard analyses

The dashboard compares each result with a residual-association reference index. Claims are positive except for TWINS (negative); support here means $P(C \mid D) > 0.5$, distinct from a 95% reporting standard. Where the covariates are available, we write R_W for the residual after regressing on an intercept and W (with a small ridge stabilizer), X_{-j} for all covariates except X_j, and $\kappa(u, v) = u^\top v / \max\{\|u\|\|v\|, 10^{-12}\}$ for the normalized inner product; the index for covariate j is

$$b_j = \sqrt{\left|\kappa(R_{X_{-j}} A, R_{X_{-j}} X_j)\, \kappa(R_{A,X_{-j}} Y, R_{X_{-j}} X_j)\right|}.$$

The reference index is $b_{\text{ref}} = \max_j b_j$. The outcome and covariate in the second factor have different residualization sets: b_j is a residual-association index rather than a conventional partial correlation

or bias-equivalent symmetric γ. Under exact OLS with nondegenerate residuals and an inactive norm floor, its second factor equals $r_{Y,j}\sqrt{1-r_{A,j}^2}$, where $r_{A,j} = \kappa(R_{X_{-j}}A, R_{X_{-j}}X_j)$ is the first factor and $r_{Y,j} = \kappa(R_{A,X_{-j}}Y, R_{A,X_{-j}}X_j)$ is the partial correlation of Y and X_j given A and X_{-j}. Thus $F_{0.5}/b_{\mathrm{ref}}$ is not an exact multiplier of omitted-confounder strength. For IHDP and TWINS, the numerator is the Bayesian-linear mean score of Table 1 or 3, averaged over replications or seeds, and the denominator is b_{ref} of a single reference covariate set (IHDP replication 1, attenuated setting; TWINS seed 0). LaLonde is reported without a covariate-reference denominator.

Table 12 lists the reference comparisons. LaLonde and the ACIC rows have no covariate-reference denominator, so no ratio is formed for them; for ACIC-cov+Hill the potential-outcome DGP is known, and the table reports BCF-style $F_{0.5}$ scores. For ACSIncome with all covariates the posterior claim probability is 0.4996, so the no-claim convention applies and no score is reported.

Table 12: Reference comparisons. b_{ref} is the largest residual-association index, not a calibrated hidden-confounding strength. Ratios appear only where a covariate reference exists and the claim is supported; they are computed before rounding.

benchmark	$F_{0.5}$	b_{ref}	$F_{0.5}/b_{\mathrm{ref}}$
TWINS	0.178	0.131	1.37
IHDP atten.	0.339	0.153	2.22
NHEFS	0.418	0.139	3.01
LaLonde CPS	0.100	—	—
LaLonde NSW	0.347	—	—
ACIC-cov+Hill small	0.216	—	—
ACIC-cov+Hill large	0.412	—	—
ACSIncome, all covariates	no claim	0.121	—

ACSIncome is used as a covariate-set audit rather than as a truth-labeled causal benchmark (Table 13). The all-covariate adjustment includes direct income and labor-market proxies and yields no supported sign claim. Both exclusion-based sets support the positive claim, with positive flip distances. The fixed feature cap admits replacement variables, so this is not a nested exclusion experiment.

Table 13: ACSIncome covariate-set audit. The all-covariate adjustment yields no supported sign claim, and both alternative sets support one; the ratios refer only to the residual-association benchmarks. The last column gives the covariate with the largest index (ACS codes: RETP, retirement income; OC, own child; RELP, relationship to householder).

covariate set	$P(C \mid D)$	$F_{0.5}$	b_{ref}	ratio	top covariate
all cov.	0.4996	—	0.121	—	RETP
exclude income	1.0000	0.279	0.115	2.43×	OC
exclude income+labor	1.0000	0.303	0.116	2.62×	RELP

Figure 10 shows the residual-association indices behind the ACSIncome ratio after the income+labor exclusion. The largest index belongs to the relationship-to-householder variable (RELP) and is well below the flip distance $F_{0.5} = 0.303$, giving a ratio of 2.62.

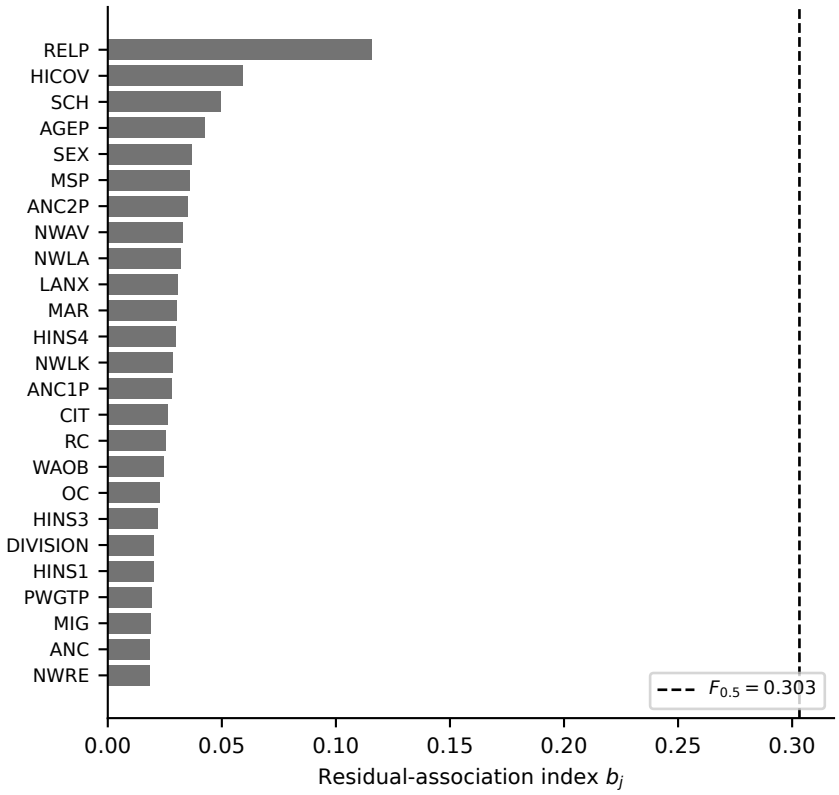

Figure 10: ACSIncome residual-association indices for the income+labor exclusion set (the 25 largest, labeled by ACS PUMS variable code). After excluding income and labor-market proxies, RELP (relationship to householder) has the largest residual-association index and the supported sign claim has $F_{0.5} = 0.303$.

N Large-scale ACSIncome audit

We include a large observational ACSIncome audit (American Community Survey PUMS records) to examine whether the diagnostic remains informative at scale (Table 13). With all covariates, the posterior sign claim is unsupported: posterior mean 2.0×10^{-5} and $P(C \mid D) = 0.4996$. The retrospective exclusion audit removes direct income components, employment status, occupation/industry, work or commuting variables, and clear post-treatment labor-market proxies as alternative adjustment choices. After income exclusion, $F_{0.5} = 0.279$, $b_{\mathrm{ref}} = 0.115$, and the descriptive ratio is 2.43. After income+labor exclusion, these values are 0.303, 0.116, and 2.62. Each set is selected separately up to 60 features; the outcome is the standardized asinh of nonnegative total personal income, and the fits are unweighted.

O Proof details

This appendix gives the algebraic details omitted from the main-text proof sketches.

Proposition 1. For (a), the posterior expected utility of $a = 1$ minus that of $a = 0$ under stress δ equals $P(\tau_\delta > 0 \mid D) - \alpha$. With ties assigned to action 0, $a^*_\delta = 1$ if and only if $P(\tau_\delta > 0 \mid D) > \alpha$. Since $P(C \mid D) > \alpha$, $a^*_0 = 1$, so $\{\delta : a^*_\delta \neq a^*_0\} = \{\delta : P(\tau_\delta > 0 \mid D) \leq \alpha\}$ and the two infima coincide; restricting to the path gives $F_{\alpha,\xi}$. The path image is a subset of Δ, so minimizing over Δ can only reduce the infimum. For example, with $T_{(\delta_1,\delta_2)}(\tau) = \tau - \delta_1 - 10\delta_2$ and the sup-norm, the joint stress (t, t) subtracts $11t$ while the path along the first axis subtracts only t.

For (b), recall $q(t,\theta) = P(\tau_{\xi(t)} > \theta \mid D)$ and $t_0 = F(0)$. Continuity of $q(\cdot,0)$ at t_0 gives $q(t_0,0) = \alpha$, and continuous differentiability with $\partial_t q(t_0,0) < 0$ gives a local root $F(\theta)$ with $q(F(\theta),\theta) = \alpha$. The stated first-crossing assumption identifies this root with the decision boundary. Differentiating gives

$$\partial_t q(F(\theta),\theta)\, F'(\theta) + \partial_\theta q(F(\theta),\theta) = 0, \qquad \partial_\theta q(t,\theta) = -f_t(\theta),$$

so $F'(0) = f_{t_0}(0)/\partial_t q(t_0,0)$. Therefore $|F(\theta) - F(0)| = f_{t_0}(0)|\theta|/|\partial_t q(t_0,0)| + o(|\theta|)$. If the derivative bounds hold along the roots between 0 and θ, integration gives $(\overline{B}/\underline{B})|\theta|$. A bound for a finite threshold cannot be obtained from the slope at a single point alone. For a stress map that shifts location, $\tau_t = \tau - b(t)$ with differentiable b, this derivative reduces to $-1/b'(t_0)$ when the density is positive. □

Proposition 2. Because the joint family contains both single-stress axes, every single stress that reaches the threshold embeds in the joint family with the same norm; no monotonicity is needed. If a finite single-stress infimum is not attained, choose a sequence of feasible single stresses whose norms decrease to that infimum. The corresponding joint stresses prove the inequality by taking limits. Empty feasible sets have infimum $+\infty$ and give the same extended-real inequality. Strictness follows whenever a feasible joint stress has norm below both single-stress infima. □

Theorem 1. The Frisch–Waugh–Lovell identity and the homoskedastic standard error give $\widehat{\mathrm{se}}^2 = \mathrm{RSS}_{Y|A,X}/(\nu\,\mathrm{RSS}_{A|X}) = s^2/\nu$, hence $(\hat\tau/s)^2 = \hat t^2/\nu$. Since $P(\tau - b > 0 \mid D) \le 0.5$ exactly when b is at least the posterior 0.5-quantile $\hat\tau > 0$, the boundary is reached when $b(\gamma; D) = \hat\tau$. With $u = \gamma^2$,

$$\psi(u) = \frac{u^2}{1-u} = \frac{\hat t^2}{\nu}, \qquad \psi^{-1}(v) = \frac{\sqrt{v^2+4v} - v}{2}, \quad v \ge 0.$$

The classical symmetric partial-R^2 value for reducing the coefficient to zero solves the identical equation. Since ψ is strictly increasing on $[0,1)$, $u = RV^{\mathrm{CH}}$. The exact identity requires the stated quantile and scale; ridge regularization and finite-draw quantiles produce approximation discrepancies, quantified in Appendix B. The negative-claim case follows by reversing the sign of the estimand and stress. □

Proposition 3. In the fixed-μ regime, posterior interval width is proportional to σ_n, hence vanishes, and $P(\tau > 0 \mid D_n) = \Phi(\mu/\sigma_n) \to 1$. The median is μ, so with fixed positive s the median-flip condition is $b(\gamma; D) = \mu$, that is, $u^2/(1-u) = (\mu/s)^2$ with $u = \gamma^2$, whose solution is $\gamma = h(\mu/s) = F^*$ for every n. In the local-effect regime $\mu_n \downarrow 0$ with $\sigma_n = o(\mu_n)$, $P(\tau > 0 \mid D_n) = \Phi(\mu_n/\sigma_n) \to 1$, while the median is μ_n and the flip condition is $b(\gamma; D) = \mu_n$; since $b(\gamma; D) = s\gamma^2\{1 + O(\gamma^2)\}$ as $\gamma \to 0$, $F_{0.5} = (\mu_n/s)^{1/2}\{1 + o(1)\} \to 0$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction list the posterior-fragility definition (Section 4), the stress-map formulation, three propositions and the Cinelli–Hazlett calibration theorem (Section 5), and the empirical evaluation (Section 6). These claims are supported by proofs, numerical verification, and dashboard analyses.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses (i) the symmetric, single-confounder partial-correlation stress, (ii) the TWINS posterior summaries being on the standardized outcome scale rather than in mortality units, (iii) the dependence of $F_{0.5}$ on the posterior median only, and (iv) extensions to several confounders, to minimax summaries over sets of stresses, and to a fully Bayesian variant, left as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Propositions 1–3 and Theorem 1 list their assumptions explicitly (differentiability for Proposition 1(b), axis embedding for Proposition 2, the posterior 0.5-quantile and residual scale for Theorem 1, a Gaussian posterior with a fixed bias scale for Proposition 3). Proof sketches follow each statement, with complete proofs in Appendix O. Numerical sanity checks are in the experiments and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6 and Appendix A specify the data, posterior generators, tree counts, draw counts, and seed counts.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All datasets are public, and their sources are given in the Data availability section. The code is not released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The "Posterior generators" paragraph of Section 6 lists tree counts and draw counts, and Appendix A gives the per-experiment settings. Datasets, configurations, and seed counts are itemized in the same section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1, 3, and 5 report $F_{0.5}$ as mean$\pm$sd across replications or seeds (TWINS also gives seed ranges). Figures 6(a) and 8(b) show seed-level standard deviations.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All analyses run on one desktop CPU (16 cores, 128 GB memory) without GPUs. Each analysis is single-threaded, and the complete set of experiments takes under one hour of wall-clock time; the tree-based posterior generators account for most of it.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public, de-identified benchmark datasets and contributes a methodological diagnostic; no human-subject data collection or model deployment is involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is a theoretical and methodological contribution to causal-inference reporting. The intended impact is to make the robustness of Bayesian causal conclusions more transparent; we do not foresee specific direct misuse paths beyond those already present in any sensitivity-analysis tool.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No data or models with misuse risk are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data creators are credited by citation in the Data availability section (IHDP, TWINS, ACIC 2016, LaLonde, NHEFS, and ACSIncome). The Data availability section also names the distributing repositories; the data are used as distributed there, under their stated terms, for research analysis only.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for editing assistance (grammar suggestions), which is not a core method per the NeurIPS guidelines.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.