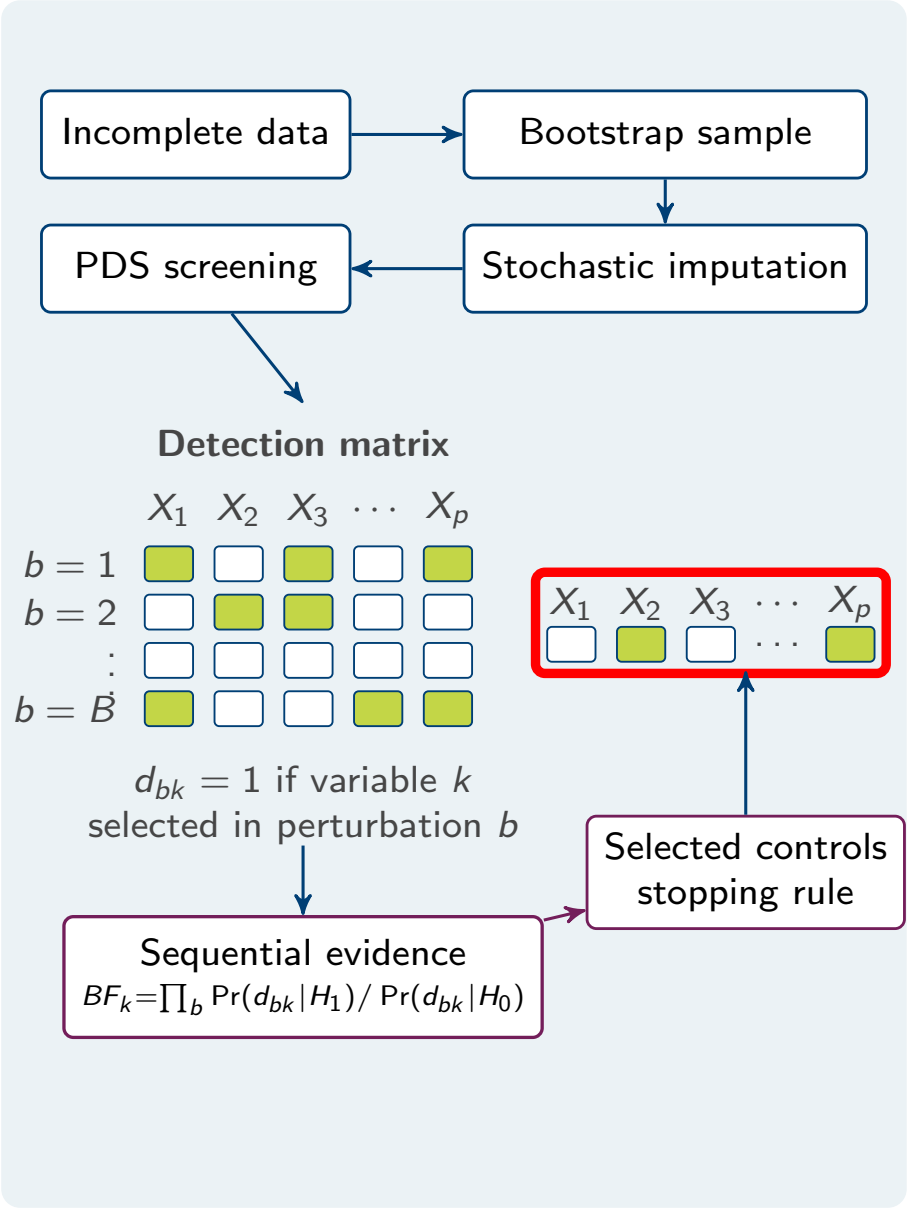
A Bayes-Factor-Guided Approach to Post-Double Selection

with Bootstrapped Multiple Imputation



Johannes Bleher
Claudia Tarantola

University of Hohenheim &
University of Milan

JSM 2026, Boston
August 1–6, 2026

Motivation
Starting Point

- Empirical researchers often estimate a target effect with many potential controls (Schwerter et al., 2025).
- Post-double selection (PDS) uses LASSO in both the outcome and treatment equations to reduce omitted-control risk (Belloni et al., 2014a,b).
- Missing covariates and sampling uncertainty add another layer: imputation and bootstrap perturb the data before selection (Rubin, 1987; Efron, 1979).
- Across BOOT-MI replications, the selected model is no longer stable.

Motivation

Aggregation Problem

- Union rules are common when selections vary across imputations, but they can become dense (Chen and Wang, 2013).
- Frequency and stability-based rules retain variables selected often enough (Meinshausen and Bühlmann, 2010).
- Both are simple, but neither uses the sequential evidence in repeated detections.

Trade-off

Union rules are sensitive but dense.

High frequency thresholds are sparse but may lose persistent weaker signals.

Motivation
Contribution

- Within an analyst-defined admissible pool, reduce high-dimensional candidates to a stability-ranked adjustment recommendation.
- Treat variable-selection outcomes across perturbations as binary detection histories.
- Use an empirically calibrated likelihood-ratio process for inclusion and stopping.
- Derive working-model properties and add a BF-guided pilot-budget extension.
- Evaluate operating characteristics across 126 Monte Carlo scenarios and an ESS application.

Motivation

Causal Screening Must Precede Statistical Screening

Mandatory $\mathcal{M}$

Known pre-treatment controls enter every specification.

Prohibited $\mathcal{F}$

Known mediators, treatment descendants, and colliders stay out.

Eligible $\mathcal{E}$

Sequential aggregation screens the remaining high-dimensional pool.

$$\widehat{\mathcal{S}}_{\text{adj}} = \mathcal{M} \cup \widehat{\mathcal{S}}_{\text{BF}}(\mathcal{E})$$

Interpretation

Stable selection is statistical evidence within $\mathcal{E}$; it is not evidence that a variable is a confounder (Pearl, 2009; Hernán and Robins, 2020).

Method
BOOT-MI PDS Iteration

1. Draw a bootstrap sample from the incomplete data (Efron, 1979; Efron and Tibshirani, 1994).
2. Stochastically impute missing covariates (Rubin, 1987; Little and Rubin, 2019).
3. Run LASSO in the outcome equation and treatment equation.
4. Construct the within-iteration PDS candidate set.
5. Record binary detections d_{bk} for each candidate variable k.

Method
Detection Rule

- PDS screens both nuisance equations (Belloni et al., 2014a).
- Variables selected in the treatment equation are retained.
- Variables selected only in outcome equation must pass post-selection significance check.
- The asymmetric rule reflects their roles in treatment-effect estimation.

Detection History

$$d_{bk} = \begin{cases} 1, & \text{detected} \\ 0, & \text{not detected} \end{cases}$$

b: perturbation iteration

k: candidate variable

This is a working detection model, not a full Bayesian variable-selection model.

Working Detection Model

$$d_{bk} \mid H_0 \sim \text{Bernoulli}(\pi_0), \qquad d_{bk} \mid H_1 \sim \text{Bernoulli}(\pi_1).$$

Sequential Evidence

$$E_{kt} = \prod_{b=1}^{t} \frac{\Pr(d_{bk} \mid H_1)}{\Pr(d_{bk} \mid H_0)}, \qquad \log E_{kt} = \sum_{b=1}^{t} \ell(d_{bk}).$$

Conditional on fixed (π_0, π_1), E_{kt} is a likelihood ratio between two working detection regimes. The probabilities are estimated from pilot perturbations.

Method
Decision and Stopping

Variable-Level Decision

$$\log E_{kt} \geq c \quad \Rightarrow \quad \text{include } X_k.$$

$$\log E_{kt} \leq -c \quad \Rightarrow \quad \text{exclude } X_k.$$

$$|\log E_{kT_{\max}}| < c \Rightarrow \text{undecided.}$$

Stopping Rule

Continue perturbation iterations until all variables have crossed an evidence boundary, subject to a minimum delay and maximum iteration budget.

Baseline

Pilot calibration, $T_{\text{pilot}} = 20$, $T_{\max} = 200$, and $c = \log(10)$ in the Monte Carlo study.

Method

What the Theory Establishes

Working-null boundary

If the detection process is dominated by the low-detection regime,

$$\Pr\left(\sup_t \log E_{kt} \geq c\right) \leq e^{-c}.$$

A union bound extends this to multiple variables.

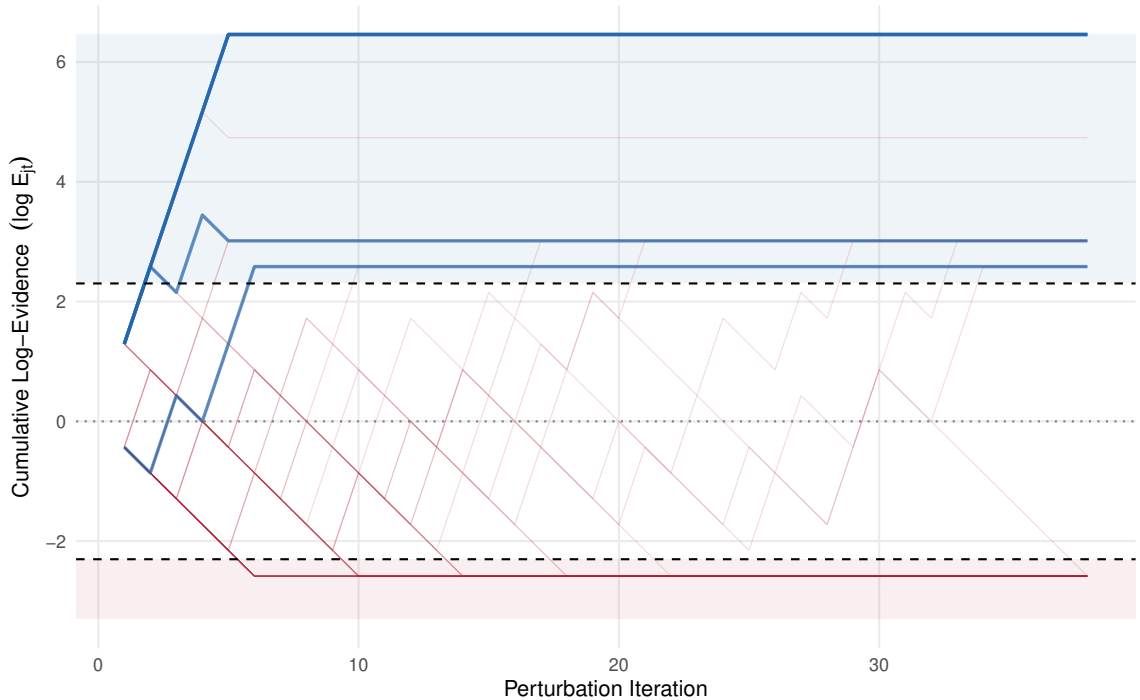
Long-run classification

With fixed working probabilities, evidence drifts according to whether the effective detection probability lies above or below the break-even rate q^*.

Interpretation Boundary

These working-process results are not unconditional structural-support FWER guarantees and do not provide post-selection treatment-effect inference.

Method
Evidence Paths



- Colors show simulation truth: blue paths are $X_1, \ldots, X_5$; red paths are truly irrelevant controls.
- Boundary crossings show decisions; red upper-bound crossings are false positives in this realization.

Simulation
Monte Carlo Design

- Partially linear model following the PDS simulation tradition:

$$Y_i = D_i\alpha_0 + X_i'\beta + \varepsilon_i,$$
$$D_i = X_i'\gamma + \nu_i.$$

- $p = 50$ candidate controls; $k_0 = 5$ are relevant.
- 500 replications per scenario.
- All candidates are pre-treatment; this design contains no collider.
- $n \in \{100, 500, 1000\}$.
- $R^2 \in \{0.2, 0.6\}$.
- Homoscedastic and heteroscedastic errors.
- MCAR, MAR, MNAR missingness at 20%, 40%, 60%.
- 126 scenarios in total.

Simulation
Benchmarks and Metrics

Benchmarks

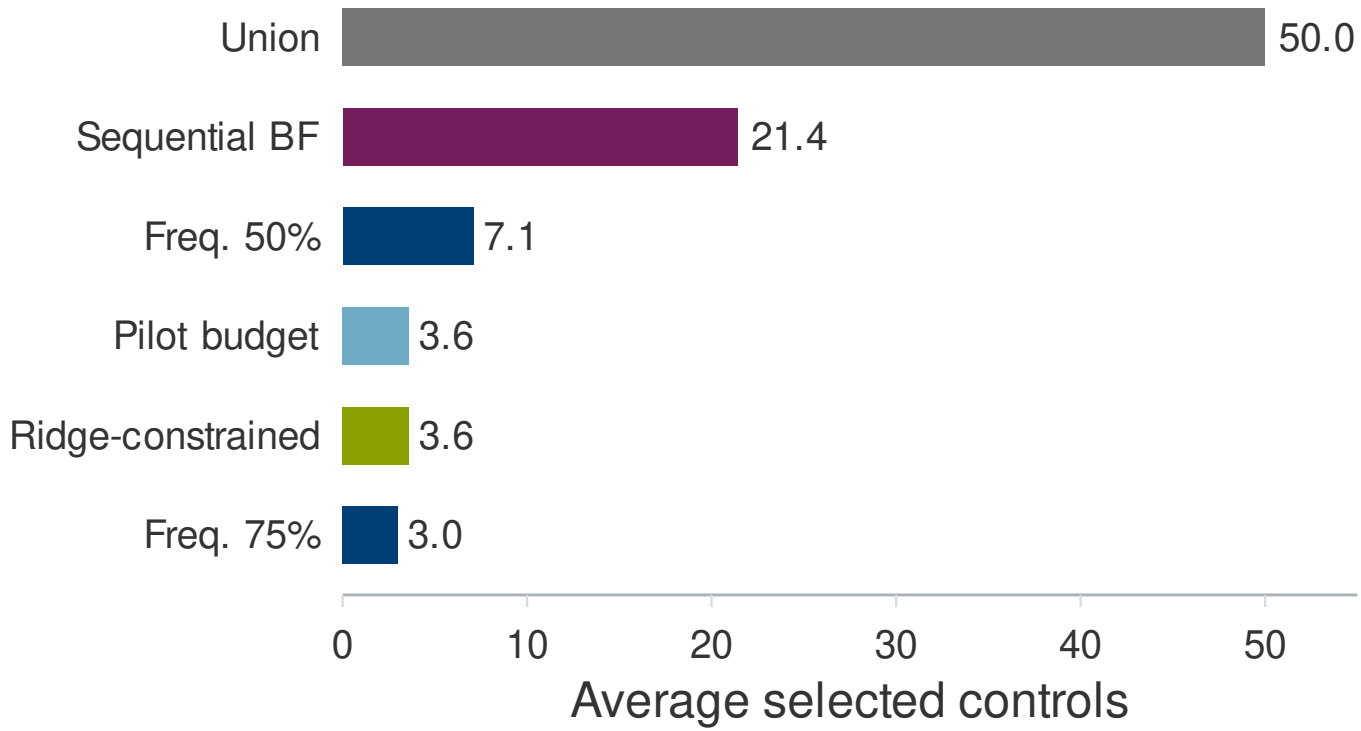
- Union and 50%/75% frequency rules.
- Pilot-budget and ridge-constrained extensions.
- Perturbation-average estimator benchmark.
- Fixed-budget and matched-budget comparisons.

Evaluation

- Variable selection: TPR, FPR, model size.
- Treatment effect: bias, RMSE, coverage.
- Efficiency: stopping iterations.

Simulation

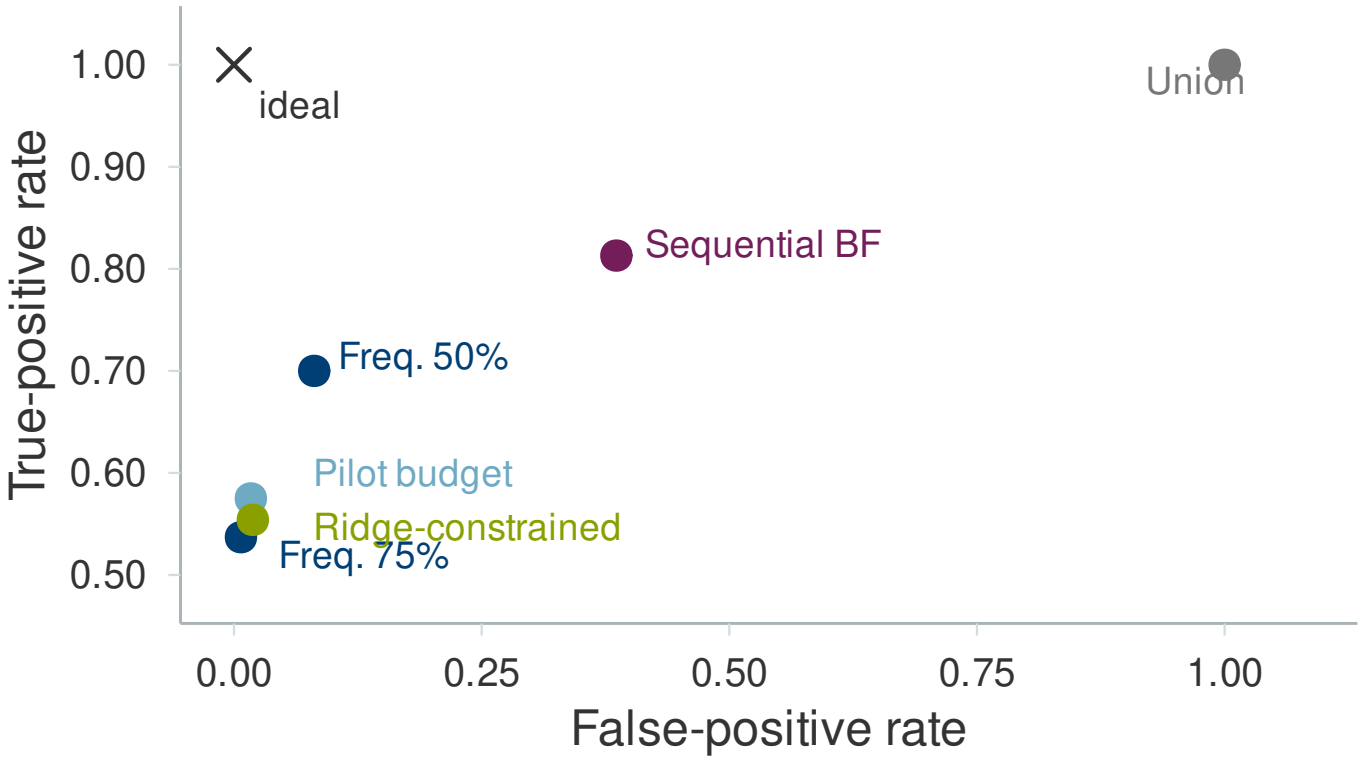
Conservative Variants Select 3–4 Controls



The baseline retains 21.4 controls on average; the pilot-budget and ridge variants retain 3.6.

Simulation

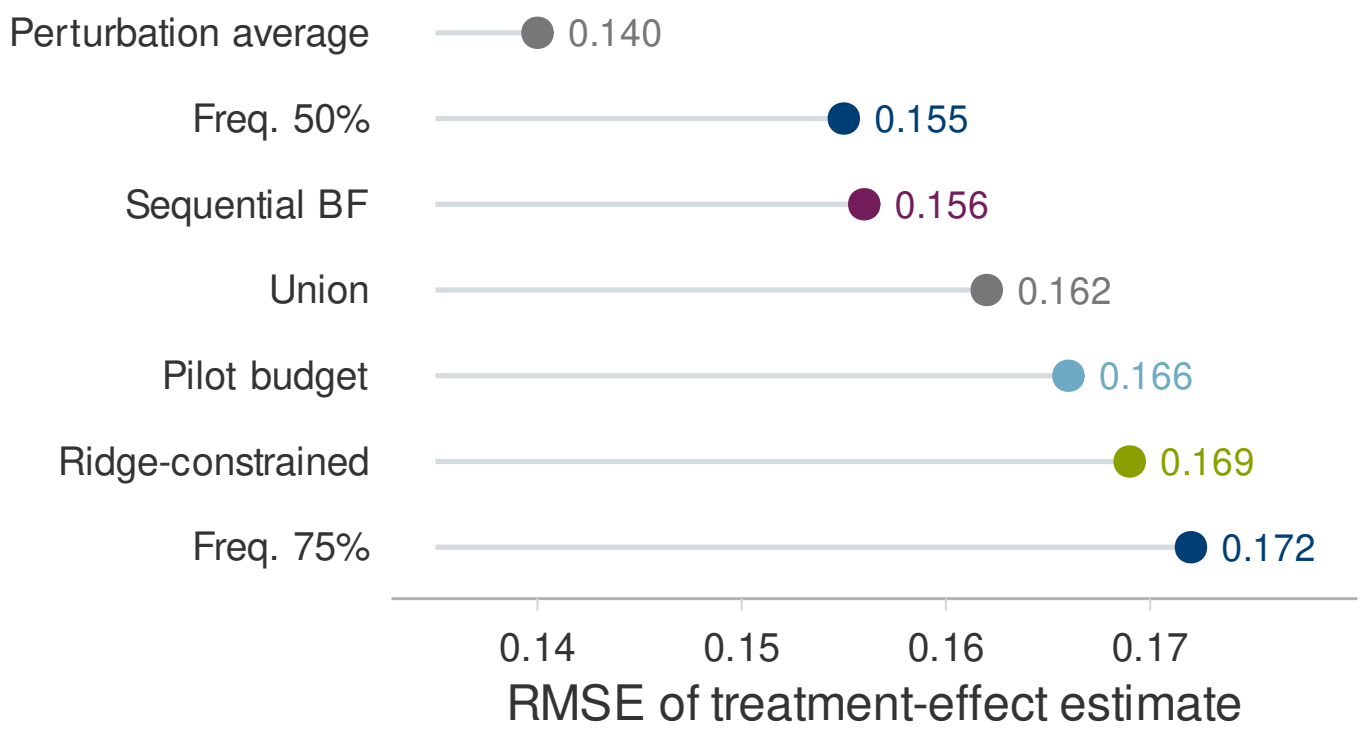
Sensitivity Comes with False Positives



The baseline gains sensitivity relative to frequency rules, but its false-positive rate rises to 0.386.

Simulation

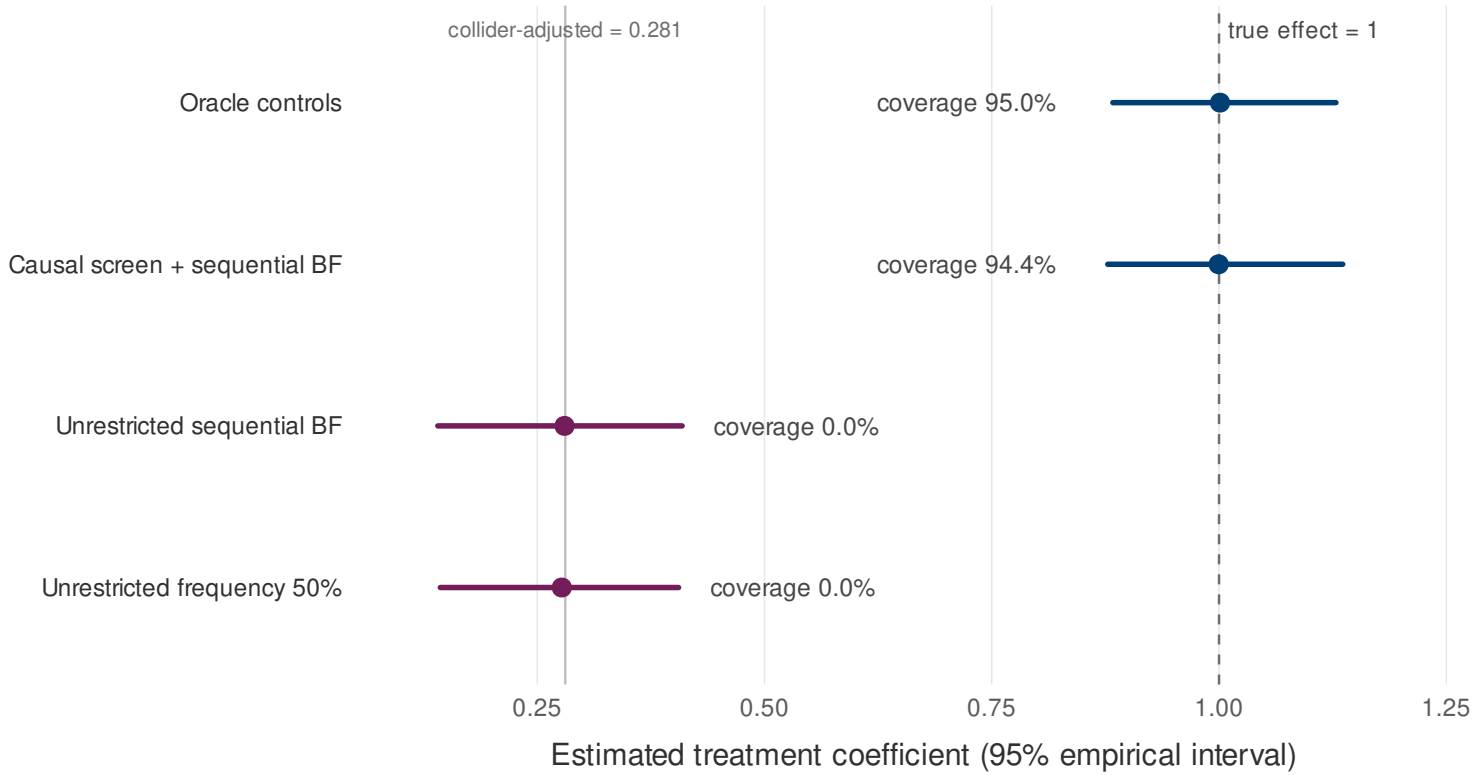
Treatment-Effect RMSE Is Similar



Sequential BF is close to the frequency and union rules. Its 95% interval coverage is 0.867, so RMSE similarity does not imply nominal inference.

Simulation

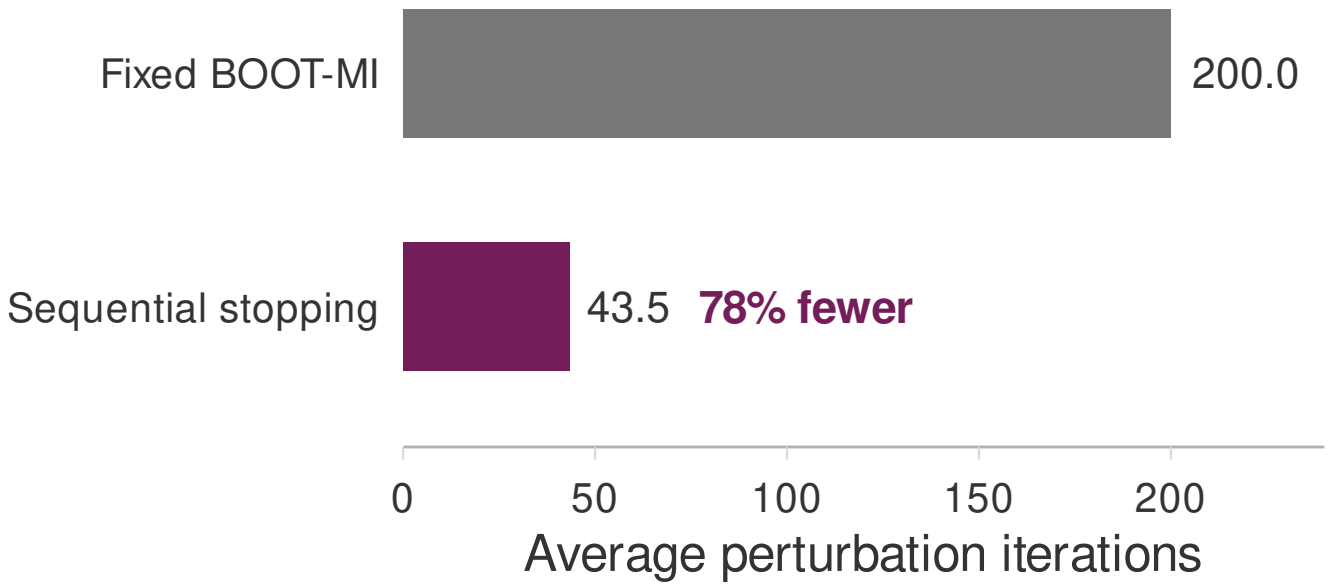
A Stable Collider Still Biases the Treatment Effect



With 500 replications, excluding the collider before selection restores centering and coverage; both unrestricted aggregation rules select it persistently.

Simulation

Stopping Uses 78% Fewer Iterations



These are perturbation counts, not method-specific wall-clock runtimes.

Simulation
Simulation Takeaways

- The baseline calibration yields the highest TPR among selective rules, at a substantial FPR cost.
- Pilot-budget and ridge-constrained variants provide conservative operating points.
- Treatment-effect estimation is competitive, not uniformly dominant.
- Coverage evaluates the complete pipeline under the simulated design; aggregate coverage is 0.867 rather than 0.95.
- The clearest operational gain is a 78% reduction in perturbation iterations.

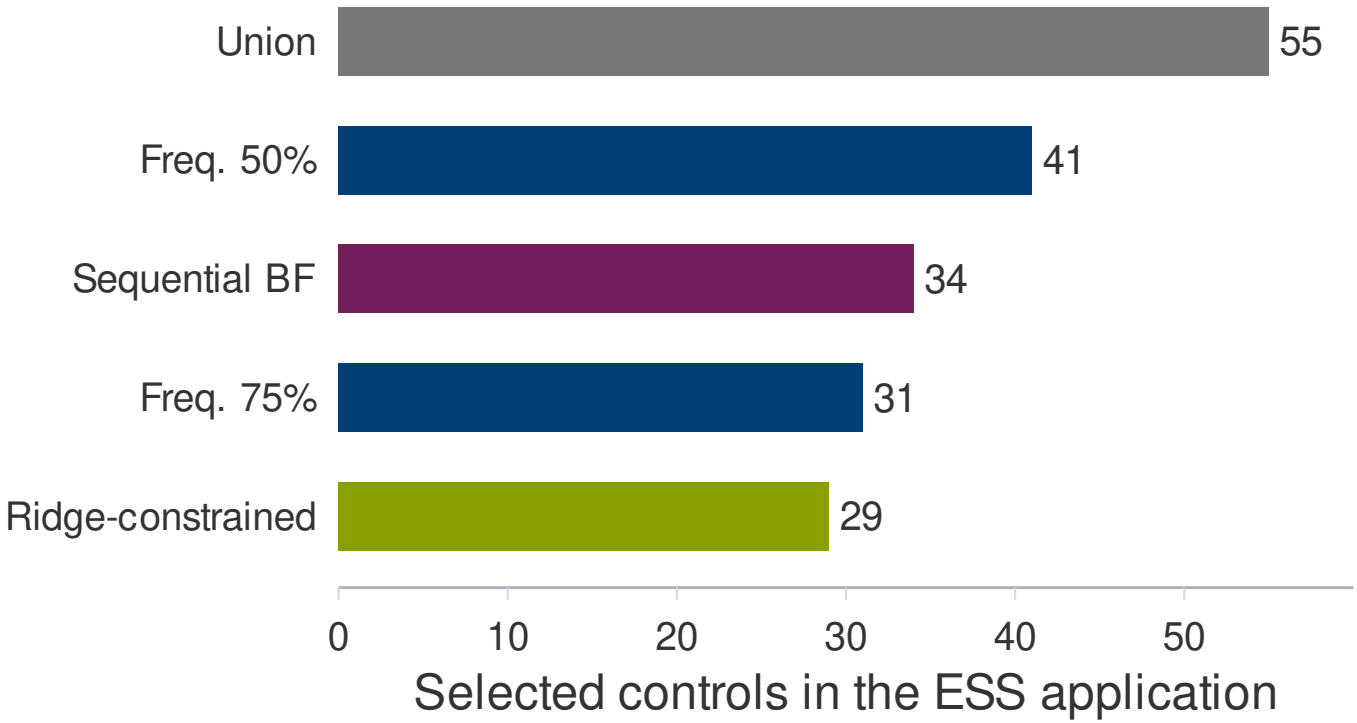
Empirical Illustration

ESS Application

- Data: European Social Survey, Round 10.
- Analytic sample: women with nonmissing employment and household-child indicators.
- Remaining covariates are allowed to be missing and handled by imputation.
- Sample size after restrictions: $n = 8{,}543$.
- Main empirical threshold: $c = \log(1000)$ for conservative classification.

Empirical Illustration

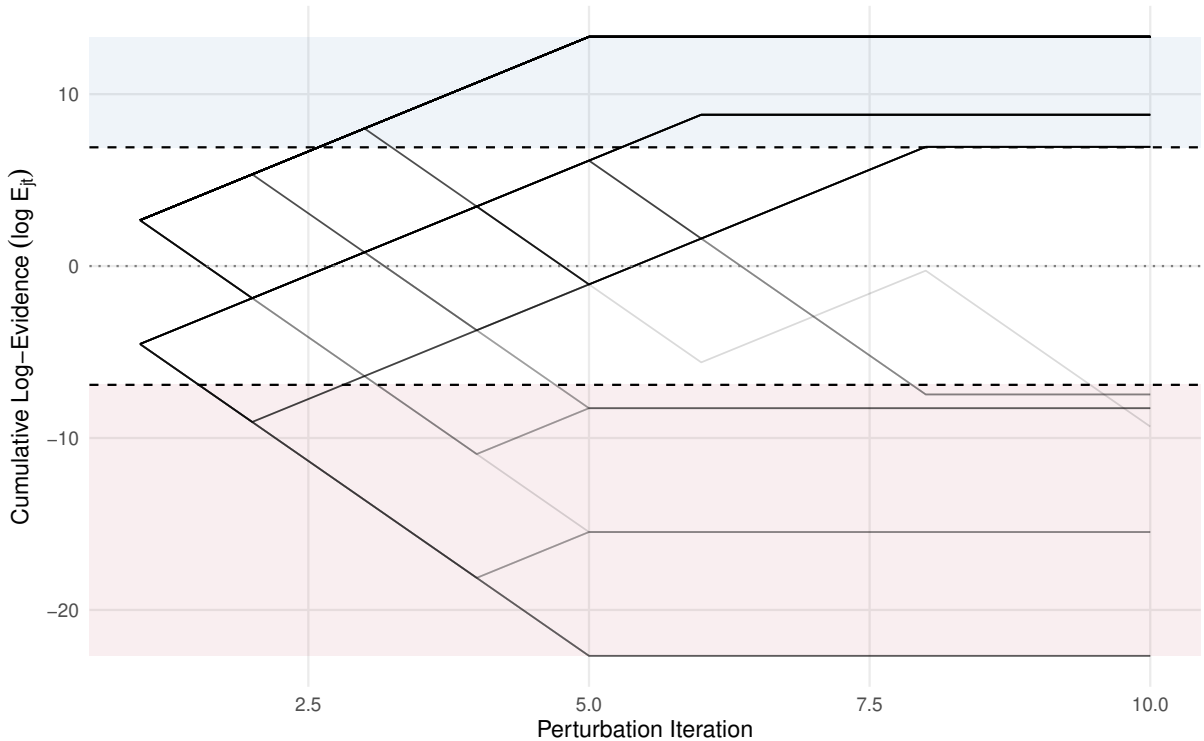
ESS Models Range from 29 to 55 Controls



The sequential BF model lies between the 50% and 75% frequency rules; ridge calibration is the most selective.

Empirical Illustration

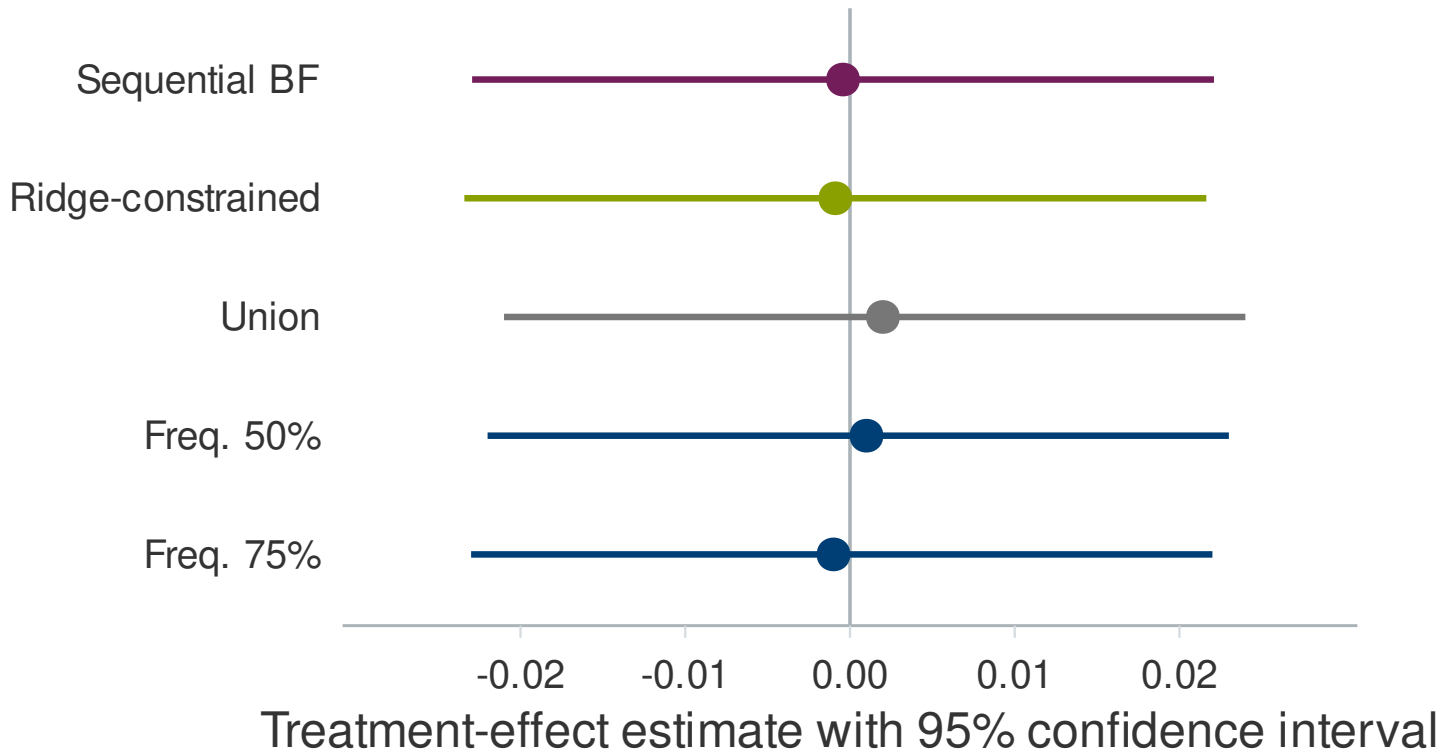
Empirical Evidence Paths



In the ESS application, all candidate variables are classified within 10 perturbation iterations under the conservative threshold.

Empirical Illustration

ESS Estimates Are Indistinguishable from Zero



All confidence intervals overlap closely; aggregation changes model size, not the empirical conclusion.

Conclusion
What the Method Adds

- A structured alternative to union and frequency aggregation.
- A working likelihood-ratio scale for inclusion, exclusion, and stopping.
- Explicit boundary-crossing and classification results under the working model.
- A BF-guided pilot budget that also uses within-iteration evidence strength.
- Applicability beyond PDS whenever a selection method yields binary detection indicators.

Conclusion
Limitations

- Stability cannot identify confounders or detect colliders; causal admissibility must be supplied before selection.
- The Bernoulli detection model is a working approximation; perturbation outcomes are not truly independent.
- Calibration relies on a short pilot phase and should be checked in applications.
- Working-model boundary control is not structural-support FWER without an additional domination condition.
- Treatment-effect coverage is an empirical diagnostic; coefficients on selected controls have no post-selection coverage guarantee here.

Conclusion

Takeaway

Main Message

Causal design defines what may be selected; sequential evidence makes the remaining high-dimensional decision transparent and finite.

- Treatment-effect performance is competitive, not uniformly superior; conservative variants recover sparser operating points.
- Coverage is an empirical diagnostic, not an automatic consequence of stable selection.
- The stopping rule uses 78% fewer perturbation iterations on average.

Computations: bwHPC infrastructure (DFG INST 35/1597-1 FUGG).
Collaboration: HERMES seed grant program.

References

Belloni, A., Chernozhukov, V. and Hansen, C.: 2014a, High-dimensional methods and inference on structural and treatment effects, *Journal of Economic Perspectives* **28**(2), 29–50.

Belloni, A., Chernozhukov, V. and Hansen, C.: 2014b, Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies* **81**(2), 608–650.

Chen, Q. and Wang, S.: 2013, Variable selection for multiply-imputed data with application to dioxin exposure study, *Statistics in Medicine* **32**(21), 3646–3659.

Efron, B.: 1979, Bootstrap methods: Another look at the jackknife, *The Annals of Statistics* **7**(1), 1–26.

Efron, B. and Tibshirani, R. J.: 1994, *An Introduction to the Bootstrap*, 1 edn, Chapman and Hall/CRC.

Hernán, M. A. and Robins, J. M.: 2020, *Causal Inference: What If*, Chapman & Hall/CRC, Boca Raton, FL.

References

Little, R. J. A. and Rubin, D. B.: 2019, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Wiley.

Meinshausen, N. and Bühlmann, P.: 2010, Stability selection, *Journal of the Royal Statistical Society: Series B* **72**(4), 417–473.

Pearl, J.: 2009, *Causality: Models, Reasoning, and Inference*, 2 edn, Cambridge University Press, Cambridge.

Rubin, D. B.: 1987, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

Schwerter, J., Bleher, J., Doebler, P. and McElvany, N.: 2025, Metropolitan, urban, and rural regions: How regional differences affect elementary school students in germany, *AERA Open* **11**(1), 1–26.
URL: *https://doi.org/10.1177/23328584251331453*