Detecting Researcher-Level p-Hacking: An Empirical Bayes Approach

Abel Brodeur* Karthik Tadepalli†

April 2, 2026

Abstract

Statistical methods typically lack power to detect p-hacking and publication bias among individual researchers, making it difficult to detect how prevalent p-hacking is. We propose a novel empirical Bayes method to estimate researcher-level p-hacking prevalence. Using data from top medical journals, we find strong evidence these practices are widespread: at least 85% of researchers over-reject null hypotheses, with a conservative lower bound of 73%. Our approach identifies 20% of researchers as over-rejecting, whereas conventional tests detect none. These findings demonstrate that p-hacking and publication bias are systemic rather than isolated misconduct, underscoring the need for structural reforms in scientific practice.

JEL Codes: B41, C12, I10

*University of Ottawa, Department of Economics and Institute for Replication. abrodeur@uottawa.ca

†University of California Berkeley, Department of Economics. karthikt@berkeley.edu

1 Introduction

Statistical significance testing remains the cornerstone of empirical inference in many scientific disciplines, but its misuse has generated increasing concern (Benjamin et al. (2018); Head et al. (2015); Simmons et al. (2011)). Two interrelated threats to research credibility—p-hacking and publication bias—have been shown to distort published findings, undermine replicability, and inflate effect sizes. In p-hacking, researchers deliberately or inadvertently manipulate data or modeling choices to achieve statistical significance (Ackley et al. (2025); Fitzpatrick et al. (2024); Masicampo and Lalande (2012); McCloskey and Michaillat (2024); Naguib (2025)); in publication bias, journals disproportionately publish statistically significant findings (Bartoš et al. (2025); Brodeur et al. (2023); Doucouliagos and Stanley (2013); Havránek et al. (2024); Ioannidis et al. (2017); Lakens (2015); Song et al. (2013); Van Aert et al. (2019)). These distortions undermine the reliability of research findings, inflates false positives, and can mislead future research, policy decisions, and public trust. While the presence of such distortions is well-documented at the level of entire literatures, little is known about how these behaviors are distributed across researchers.

This paper asks: is p-hacking (and, more broadly, over-rejection of null hypotheses) the product of a small subset of researchers, or are it driven by widespread behaviors shared across the research community? The distinction is not merely academic; it carries implications for scientific norms, research integrity policies, and editorial practice. Prior studies that test for excess significance near arbitrary thresholds typically lack statistical power to assess misconduct at the individual level (Brodeur et al. (2016); Elliott et al. (2022a); Gerber and Malhotra (2008); Simonsohn et al. (2014)), leading to an incomplete understanding of individual engagement in over-rejection. If p-hacking and publication bias are driven by a narrow set of researchers with consistently inflated results, targeted interventions may suffice. If, instead, most researchers engage in marginal over-rejection occasionally, systemic reform may be necessary. Furthermore, answering this question requires tools that can detect over-rejection even when individual researchers contribute only a small number of hypothesis tests—otherwise, it will only be possible to detect over-rejection in a tiny set of prolific researchers.

To address this, we develop an empirical Bayes framework that estimates the distribution of rejection rates across researchers, based on Kline and Walters (2021). We apply our method to a dataset of 2,796 articles published in The Lancet, The British Medical Journal (BMJ), and The New England Journal of Medicine (NEJM) between

2016 and 2022 (Brodeur et al. (2025)). From these papers, we extract 10,404 statistical tests reported in structured abstracts and focus on those near the 5% significance threshold. The novelty of our approach lies in leveraging cross-sectional information: although most researchers contribute only one or two marginal tests, combining information across hundreds of researchers allows us to estimate the prevalence and intensity of p-hacking and publication bias with high precision. We complement our main estimator with nonparametric alternatives and moment-matching bounds to assess robustness and minimize functional form assumptions.

Our findings reveal that over-rejection is widespread. The empirical Bayes estimator implies that at least 85% of researchers over-rejected marginal tests, while an alternative moment-matching approach suggests a lower bound of 73% of researchers over-rejecting. Based on this empirical Bayes estimate, we also identify approximately 20% of researchers as over-rejecting, based on the posterior intervals around their estimated rejection rates. This represents a marked increase in power over binomial tests for over-rejection, which cannot identify any researchers as over-rejecting. Together, these results suggest that the inflation of statistical significance is not the product of a few rogue actors, but rather a systemic pattern of behavior. Our contribution is both methodological and substantive: we argue that individual-level detection of statistical distortions is possible, and that the overwhelming majority of researchers in our sample exhibit behaviors consistent with p-hacking or publication bias.

Our paper contributes to the literature on p-hacking and publication bias (Andrews and Kasy (2019); Barnett and Wren (2019); Brodeur et al. (2020); Christensen and Miguel (2018); Franco et al. (2014); Havránek (2015); Vivalt (2019)) primarily through the focus on assessing the prevalence of over-rejection at the level of individual researchers. We develop an empirical Bayes approach to assess how many researchers over-reject marginal tests. We demonstrate that traditional tests lack power to detect over-rejection when applied to individual researchers (Elliott et al. (2022b)). Our method overcomes this limitation by leveraging cross-sectional variation in rejection rates among narrowly significant tests and estimating the underlying distribution of researcher-level rejection rates. This focus on researcher heterogeneity also distinguishes our paper from other work using deconvolution estimators to identify the presence of p-hacking (Kudrin 2022, 2024).

Focusing on medical articles—as opposed to those from economics or other social sciences—offers critical advantages when investigating whether p-hacking and publica-

tion bias are driven by a tail of extreme behavior, or by the vast majority of researchers over-rejecting marginal tests. First, medical research has immediate and tangible implications and articles in our sample frequently receive substantial attention from journalists, policymakers, and the general public. Second, medical journals employ highly standardized reporting formats, especially in the design of structured abstracts. Authors are required to follow stringent editorial guidelines regarding the presentation of results, often emphasizing statistical measures such as p-values and confidence intervals in consistent locations. This homogeneity minimizes coding ambiguity and ensures a more direct correspondence between reported findings and "main results", which can be ambiguous in economics journals. Third, conventions around authorship tend to be more standardized than in fields like economics. Specifically, the last-listed author is often the principal investigator or lab leader, while the first author typically reflects the primary contributor who led the work day-to-day. This enables us to make claims about the "author" of a paper that would be difficult in economics, where papers have multiple authors and no clear priority between them.

2 Framework

The conceptual framework for this paper is based on the *caliper test*. Caliper tests compare the distribution of p-values within a narrow range around a significance cutoff, to see how many more p-values there are just below the significance cutoff, compared to just above it. Define a rejection of test j by researcher i to be $R_{ij} = \mathbb{1}(p_j < 0.05)$ where p_j is the p-value.

Definition 1 *Test j is narrow if $p_j \in [0.05 - \delta, 0.05 + \delta]$, where the caliper width δ is small.*

Defining narrow tests allows us to make the key assumption for the caliper test:[1]

Assumption 1 *For narrow tests, rejections are Bernoulli trials with researcher-specific means:*

$$R_{ij} \overset{i.i.d}{\sim} Bernoulli(\pi_i)$$

With this assumption, and the focus on narrow tests, we can define p-hacking:

[1]The only substantive restriction in this assumption is that it prevents any kind of dependence between rejections; for example, it rules out that a researcher may have the leeway to publish a null result if they have published positive results in the past.

Definition 2 *Researcher i p-hacks if $\pi_i \neq 0.5$.*

The idea is that in a narrow range around an arbitrary cutoff, p-values should be just as likely to fall below as above it ($\pi_i = 0.5$). This definition mirrors how caliper tests are used to detect aggregate p-hacking in the literature.

As written, estimating π_i for each researcher seems like a hopeless endeavor. Most researchers have a small number of papers, and an even smaller number of papers where their tests are narrowly around the rejection threshold (which is necessary for our definition of p-hacking). For a researcher with four narrow tests, even if all four of them reject, this could still occur with probability $1/2^4 = 0.06$ under the null hypothesis of $\pi_i = 0.5$, and thus we could not say they were p-hacking. This limitation is the main reason that past papers have focused on detecting p-hacking in the literature as a whole, where there may be hundreds or thousands of narrow tests and so the caliper test is well-powered.

Solving this problem is our main innovation. We introduce an empirical Bayes framework to estimate the prevalence of p-hacking. The key insight is that while we have few tests per researcher, we can "borrow strength" from the entire population of researchers.

Assumption 2 *For every researcher i, π_i is drawn from a common distribution G:*

$$\pi_i \overset{iid}{\sim} G$$

G summarizes the distribution of rejection rates for narrow tests across researchers. This distribution has rich information about the prevalence of p-hacking. It tells us the share of researchers who are p-hacking. It also tells us how intensive p-hacking is; whether the aggregate p-hacking in the literature is driven by a large number of people p-hacking a little, or by a small number of people p-hacking a lot. Thus, our focus is on estimating G. The insight of empirical Bayes is that even with a small number of observations per researcher, as long as we have a large number of researchers, we can estimate G. And once we estimate G, we can use it as a powerful prior for a Bayesian test of researcher-level p-hacking. This is how we will proceed.

3 Data

We rely on the data from Brodeur et al. (2025) who assembled a comprehensive dataset comprising 2,796 original research articles published between 2016 and 2022 in three of the world's most influential medical journals: The Lancet, The British Medical Journal (BMJ), and The New England Journal of Medicine (NEJM). Each article in our sample reports empirical findings rooted in statistical inference. For each article, we extract all hypothesis tests presented within the results section of the structured abstract, yielding a total of 10,404 test statistics. The p-values are directly recorded where provided (30% of cases), and otherwise inferred via transformation of reported confidence intervals or test statistics to ensure consistency in measurement across the sample.

The distributional properties of the dataset reveal significant heterogeneity in the reporting and characteristics of test statistics. The median abstract reports three hypothesis tests, with 70% of test statistics derived from confidence intervals and 51% associated with randomized trials. Approximately half of the articles indicate pre-registration, and 11% pertain to COVID-19.

We matched papers to their first authors, while standardizing names to ensure that authors are matched to all of their papers in the data. We attribute each paper to the first author, justified by the medical context in which the first author is the lead researcher. We show throughout that our results are virtually identical when matching papers to their last-listed author (usually the PI in charge of the lab conducting the study).

Figures A1 and A2 illustrate the number of test statistics with p-values in the range $[0.02, 0.08]$ per first author and last-listed author respectively. Overall, about 500 first authors have one test in this range. Slightly less than 200 first authors have two tests, while approximately 100 first authors have three or more test statistics. We find a similar pattern for last-listed authors, although with less authors having only one test statistic.

Overall, this dataset provides a fertile empirical foundation to explore distributions in test statistics, particularly around the arbitrary threshold of $p < 0.05$.

4 Aggregate p-Hacking and Publication Bias

To explore the data further and motivate our approach, we start where most papers end—demonstrating the presence of p-hacking and/or publication bias in our sample. We focus specifically on the 5% significance level throughout. This threshold is commonly used to distinguish statistically significant from non-significant results and is either required or implicitly endorsed by the editorial guidelines of the three journals we examine.[2] Figure 1 plots the caliper test for p-values in the range $[0.02, 0.08]$. The presence of marginal p-hacking is visually apparent—78% of the p-values in this range are below 0.05, compared to only 50% under the null of no p-hacking.[3]

This finding begs the question—where does this inflation of significant results come from? Is it driven by a tail of extreme manipulation, or by the vast majority of researchers engaging in at least a little over-rejection? The aggregate version of the caliper test is silent on this question—it cannot distinguish between different researchers. This is why most papers that detect p-hacking/publication bias in a given literature cannot answer the question of how prevalent over-rejection is at the researcher level. This is the motivation behind our individual-level caliper test, given power through empirical Bayes.

5 G Estimation

This section takes our conceptual framework to the data and reports estimates of the prevalence of p-hacking and publication bias. Of note, we cannot disentangle p-hacking from publication bias because both behaviors produce similar distortions in reported results, notably the clustering of p-values just below conventional significance thresholds. Thus, we focus our framing on *over-rejection* of marginal tests as the main behavior that we are identifying with our method.

[2]The Lancet requires authors to report results using the 5% threshold, including confidence intervals and forest plots, across its various methodological guides. The BMJ does not mandate this threshold explicitly, but its worked examples consistently use it. In 2018, The New England Journal of Medicine updated its editorial policy to emphasize the 5% threshold more strongly, while also encouraging authors to limit the presentation of significance statistics to pre-specified analyses. The updated guidelines recommend replacing p-values with effect estimates and 95% confidence intervals unless multiplicity adjustments were pre-specified (Harrington et al. (2019)).

[3]It is of course not true that *exactly* 50% of p-values should fall below 0.05 under the null. Because the p-curve slopes downward, there will be slightly more p-values below 0.05 even in the absence of p-hacking. However, our estimates are quantitatively large enough that this bias is unlikely to be driving our results.

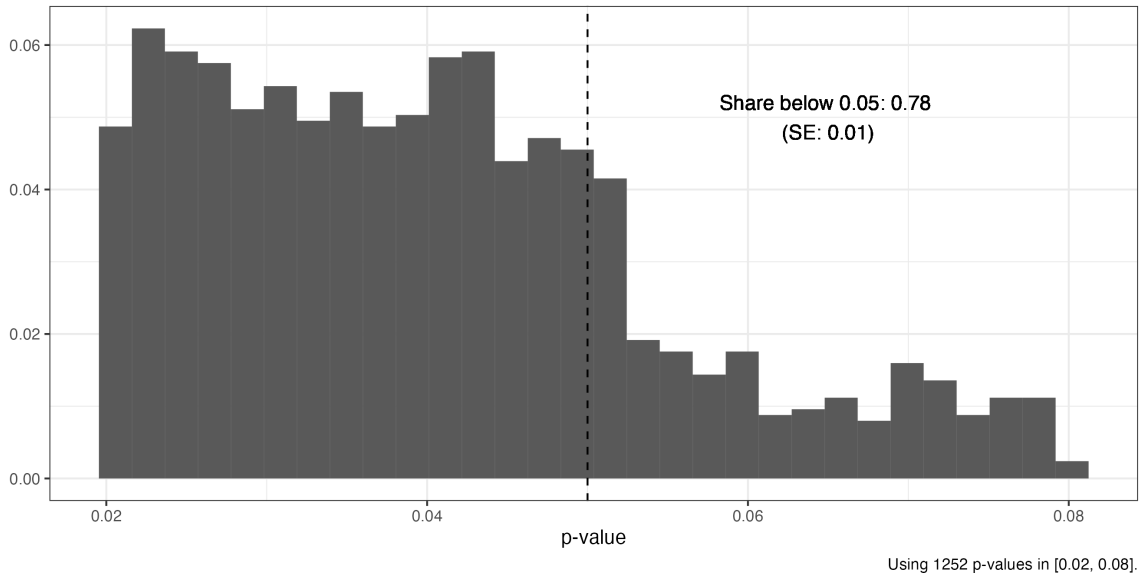

Figure 1: Aggregate over-rejection of marginal tests in our sample.

5.1 Empirical Bayes

Our main task is to estimate G, the distribution of rejection rates for narrow tests across researchers. This is a *deconvolution* task—to infer the distribution of unknown parameters that generated a known distribution. In principle, G can be an arbitrary distribution—we have no reason to believe it takes some specific functional form. However, fully nonparametric approaches to estimating G are extremely data-intensive, requiring both large numbers of researchers and many tests per researcher. Thus, we use a parametric estimator, which assumes that G belongs to a smooth exponential family, given arbitrary flexibility through a spline basis.[4] We test the null hypothesis that all researchers are fair—in other words, the null hypothesis that G is a degenerate distribution with all mass at $\pi_i = 0.5$.

Figure 2 shows the CDF of our estimate $\hat{G}$, and compares it to simulated deconvolutions of this fair distribution (to test the null hypothesis). Our results show that the deconvolution from the empirical distribution of rejection rates is far more extreme than any deconvolutions from a distribution with no p-hacking. Indeed, we can rule out that the mass at $\pi_i \leq 0.5$ is more than 15% of researchers, implying a staggering 85% of researchers are engaged in p-hacking. The median researcher rejects 80% of narrow tests. By all metrics, our estimates show that p-hacking and/or publication

[4]See Narasimhan and Efron (2020) for more details on the estimation of G.

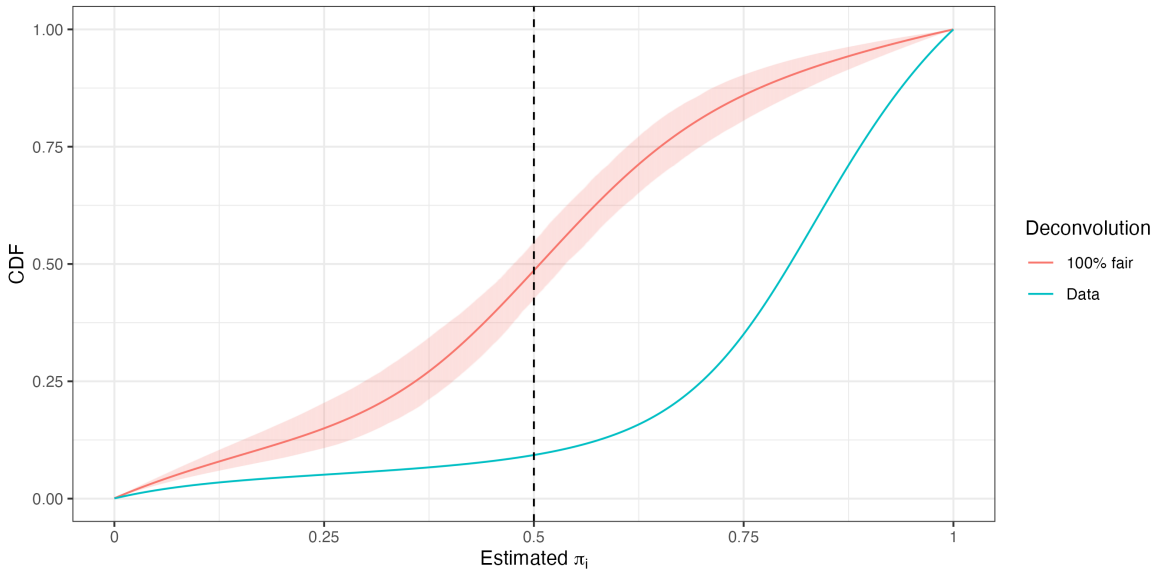

Figure 2: $\hat{G}$ estimated by deconvolution, compared to deconvolutions from a fair distribution ($\pi_i = 0.5$). Confidence intervals generated by bootstrapping over deconvolutions from fair distributions.

bias are staggeringly prevalent among medical researchers.

Our estimated $\hat{G}$ is also robust to using a tighter caliper range ($[0.03, 0.07]$ rather than $[0.02, 0.08]$) (Figure A4), and to attributing papers to their last authors rather than their first authors (Figure A3).

5.2 Moment-Matching Bounds

As a robustness exercise, we bound the prevalence of p-hacking in a different way that doesn't rely on empirical Bayes assumptions. The idea is to find the distribution $\bar{G}$ with the *lowest* prevalence of p-hacking and publication bias, while still being consistent with the data, as in Kline and Walters (2021). Specifically, $\bar{G}$ is found as the solution to the following optimization problem:

$$\begin{aligned} \max_{G} \quad & dG(0.5) \\ \text{s.t.} \quad & E[G^k] = \frac{1}{N}\sum_{i=1}^{N} r_i^k \qquad \forall k \leq K \end{aligned}$$

Here, a distribution is chosen to maximize $dG(0.5)$ (i.e., the mass at $\pi_i = 0.5$, corresponding to no p-hacking). $E[G^k]$ is the k-th moment of this chosen distribution, and $m_k \equiv 1/N \sum_{i=1}^{N} r_i^k$ is the corresponding uncentered moment of the *empirical* distribution of rejection rates (r_i being researcher i's observed rejection rate for narrow tests). K is the number of empirical moments we wish to match, a free parameter.[5]

To understand this exercise, consider a simple case where $K = 1$, i.e., we are only trying to match the mean rejection rate across researchers. Suppose the mean rejection rate across researchers is $m_1 = 0.8$ (as in our data). What is the lowest prevalence of p-hacking/publication bias that can rationalize this? It cannot be 0; the excess rejection rate (0.8 vs 0.5) must be coming from somewhere. The most conservative approach is to assume that some fraction θ are p-hacking completely ($\pi_i = 1$) while the remainder are not ($\pi_i = 0.5$). Then the mean rejection rate is

$$\begin{aligned} \theta \cdot 1 + (1 - \theta) \cdot 0.5 &= m_1 \\ \implies 0.5 + 0.5 \cdot \theta &= 0.8 \\ \implies \theta &= 0.6 \end{aligned}$$

Thus, if the mean rejection rate is 0.8, then at least 60% of researchers are engaged in p-hacking/publication bias. This example shows that moment-matching can generate lower-bound estimates on the prevalence of p-hacking/publication bias, independently of the empirical Bayes estimators.

To actually solve this optimization problem, we represent each distribution in a discretized form $\{\lambda_j\}$, where j represents grid points spaced between 0 and 1, and λ_j is the probability mass on value j. Since both our objective and our constraints are linear in probability mass, this can be solved by linear programming. Appendix Table A1 reports the results of this estimation with $K = 2$, and shows that at least 73% of researchers must be engaged in p-hacking in order to match just the first two moments of the empirical distribution of rejection rates.

Overall, this bounding approach shows that even with the limitations of the data, the quantitative estimates of p-hacking prevalence from empirical Bayes estimators are preserved under a more conservative approach (85% from the estimated $\hat{G}$, and 73%

[5] An important caveat is that the formula above is simplified for exposition; in reality, because the empirical means r_i are themselves measured with sampling variation, the higher empirical moments are upward-biased. Thus, we apply a recursive bias-correction to each empirical moment when estimating $\bar{G}$, following Kline and Walters (2021).

from our moment-matching lower-bound approach).

6 Posterior Estimation

6.1 Posterior Means and Intervals

With estimates of $\hat{G}$ in hand, we can compute the posterior distribution of rejection rates for each researcher. This gives us a Bayesian test for p-hacking: whether the 95% posterior interval for each researcher includes $\pi_i = 0.5$ (no p-hacking) or not. Figure 3 plots the posterior means and 95% intervals for all first-listed authors in our sample, using the parametric estimate $\hat{G}$ to construct the posterior. (See Figures A6 and A7 for the posterior means and intervals for all the last-listed authors and for p-values in the range $[0.03, 0.07]$ respectively.)

We can see that a large number of researchers share the same posterior mean and interval. This makes sense, because in our framework, a researcher is defined simply by having a certain number of narrow tests and a certain number of narrow rejections. Since we are dealing with a coarse number of tests and rejections—the modal author in this sample has 1 narrow test and 1 narrow rejection—this will yield a coarse distribution of posterior means. We can also see that for 20% of first authors, the posterior interval for π_i does *not* include 0.5, meaning that we can conclude that those authors are engaged in p-hacking and/or publication bias.

This might seem like a surprisingly conservative statement compared to before, when we estimated the prevalence of p-hacking to be 85%. But this is a natural consequence of moving from statements about G to statements about *individual researchers*. Even if the prevalence of p-hacking is 85%, we should not expect to be able to identify which individual researchers are in that group. For example, if 100 researchers each had a single narrow test and 99 of them rejected that test, we would conclude that a very large fraction of them were engaged in p-hacking. But we couldn't identify any individual researcher as engaged in p-hacking, since they would be identical to each other. Tests for p-hacking at the individual level are more conservative than tests for the prevalence of p-hacking—as they should be. Nevertheless, our method represents a substantial improvement in the power to identify over-rejection at the researcher level, relative to classical tests.

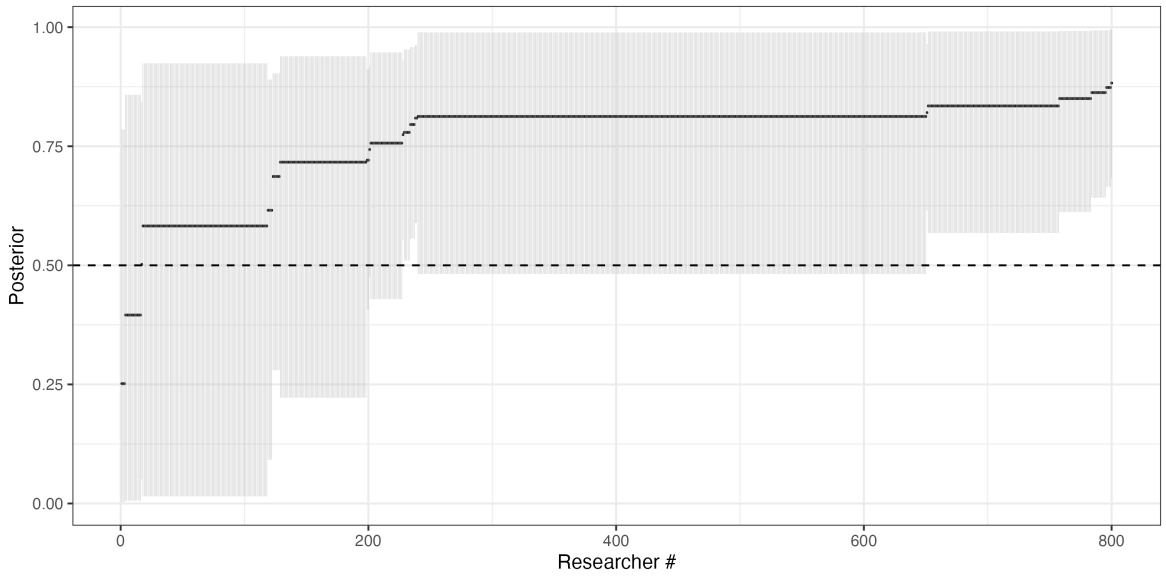

Figure 3: Distribution of posterior means and intervals across first authors

6.2 $\hat{G}$ vs Individual Rejection Rates

To understand these posterior estimates better, it helps to understand how much these conclusions are driven by our estimate of $\hat{G}$ compared to each researcher's own rejection rate r_i. In other words, how much are the conclusions about each researcher based on that researcher's behavior, vs the population behavior?

One way to do this is to estimate researcher-specific weights on the prior vs data. The posterior mean for π_i can be expressed as a weighted average of the empirical rejection rate r_i and the prior mean $E[\hat{G}]$. Knowing all three of these quantities, we can back out the implied weights on $E[\hat{G}]$. Figure 4 plots the distribution of these weights. For basically all researchers, more than half of the weight comes from the prior. The median researcher's estimate has 75% weight on the prior, and only 25% weight on the data.

Thus, the estimated $\hat{G}$ is in fact the dominant influence on the posterior estimates, and the main reason why we can detect over-rejections by 20% of researchers. Intuitively, because of the small number of tests per researcher, the researcher-specific rejection rates cannot "speak for themselves", and so we have to draw our conclusions about individual researchers from the large prevalence of over-rejections demonstrated in our estimated $\hat{G}$.

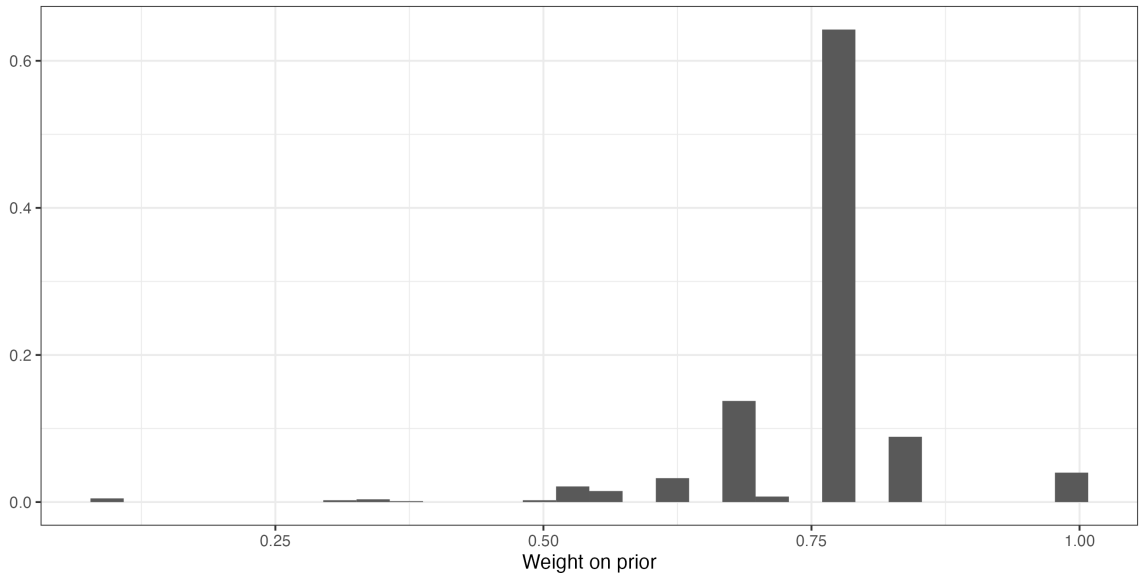

Figure 4: Distribution of researcher-specific weights on $\hat{G}$ compared to data.

6.3 Comparison to Likelihood Test

To understand how useful the empirical Bayes method of detecting over-rejection is, we can compare its results to the likelihood test conducted on the same population of researchers. The likelihood test for researcher i checks whether the number of rejections for i falls between the 2.5th and 97.5th percentiles of a binomial distribution with probability 0.5 and a number of trials equal to the researcher's number of narrow tests. In typical frequentist fashion, we are estimating the probability of observing researcher i's empirical rejection rate given the null of $\pi_i = 0.5$ and their number of narrow tests.

Figure A8 shows the comparison between the likelihood test and the posterior interval approach. In contrast to the posterior interval approach which identifies over-rejection among 20% of researchers, the likelihood test cannot detect over-rejection in virtually any researchers. This is a consequence of small sample sizes per researcher, which completely defeats the likelihood test, but which the empirical Bayes test is designed to overcome. Thus, the empirical Bayes approach significantly increases power over the classical test conducted separately for each individual.

7 Conclusion

This paper introduces an empirical Bayes framework to estimate the prevalence of p-hacking and publication bias at the researcher level. Unlike existing approaches, which typically focus on detecting distortions in the aggregate distribution of test statistics, our method enables inference about individual researchers' rejection rates—even when each researcher has only a small number of tests. Applying this framework to a dataset of over 10,000 hypothesis tests drawn from top medical journals, we provide the first systematic evidence on the distribution of over-rejecting null hypotheses across researchers.

We find that over-rejection of null hypotheses is not confined to a small group of outliers. Instead, our estimates suggest that it is pervasive: at least 73 percent of researchers in our sample over-reject null hypotheses, with our main specification indicating a prevalence rate of 85 percent. Moreover, we are able to statistically identify roughly 20 percent of researchers as over-rejecting, while classical likelihood-based tests fail to identify any researchers as engaged in over-rejection. These findings highlight the utility of empirical Bayes techniques in settings where individual-level data is sparse but the population is large.

The distinction between over-rejection concentrated among a few researchers and widespread marginal over-rejection is critical. Our results suggest the latter is more consistent with the data, implying that systemic incentives—not just individual bad actors—may underlie the distortion in published findings. If over-rejection were primarily driven by a small subset of researchers persistently engaging in misconduct, interventions could be focused on individual accountability and punitive actions, potentially through audits and retractions. However, our findings indicate that over-rejection is diffuse and systemic, with a publication culture that rewards statistical significance and treats null results as uninteresting or unpublishable (Chopra et al. (2024)). Thus, it is unlikely that actions targeted at individuals will be sufficient to make research more credible. Promising approaches to address these systemic issues include expanding the use of pre-analysis plans and registered reports (Arpinon and Espinosa (2023); Burlig (2018); Fang et al. (2015); Scheel et al. (2021)), and encouraging the publication of null findings (Blanco-Perez and Brodeur (2020)). Requiring transparency through mandatory data and code sharing can also deter selective reporting by increasing the risk of detection (Askarov et al. (2023); Brodeur et al. (2024)).

References

Ackley, S. F., Andrews, R. M., Seaman, C., Flanders, M., Chen, R., Wang, J., Lopes, G., Sims, K. D., Buto, P., Ferguson, E. et al.: 2025, Trends in the distribution of p values in epidemiology journals: Decreased p hacking or increased power? medRxiv.

Andrews, I. and Kasy, M.: 2019, Identification of and correction for publication bias, *American Economic Review* **109**(8), 2766–94.

Arpinon, T. and Espinosa, R.: 2023, A Practical Guide to Registered Reports for Economists, *Journal of the Economic Science Association* **9**(1), 90–122.

Askarov, Z., Doucouliagos, A., Doucouliagos, H. and Stanley, T. D.: 2023, The Significance of Data-Sharing Policy, *Journal of the European Economic Association* **21**(3), 1191–1226.

Barnett, A. G. and Wren, J. D.: 2019, Examination of CIs in Health and Medical Journals from 1976 to 2019: An Observational Study, *BMJ Open* **9**(11), e032506.

Bartoš, F., Maier, M., Stanley, T. and Wagenmakers, E.-J.: 2025, Robust Bayesian Meta-Regression: Model-Averaged Moderation Analysis in the Presence of Publication Bias, *Psychological Methods* .

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al.: 2018, Redefine Statistical Significance, *Nature Human Behaviour* **2**(1), 6–10.

Blanco-Perez, C. and Brodeur, A.: 2020, Publication Bias and Editorial Statement on Negative Findings, *The Economic Journal* **130**(629), 1226–1247.

Brodeur, A., Carrell, S., Figlio, D. and Lusher, L.: 2023, Unpacking p-Hacking and Publication Bias, *American Economic Review* **113**(11), 2974–3002.

Brodeur, A., Cook, N. and Heyes, A.: 2020, Methods matter: P-hacking and publication bias in causal analysis in economics, *American Economic Review* **110**(11), 3634–3660.

Brodeur, A., Cook, N. and Heyes, Anthony, W. T.: 2025, Media stars: Statistical significance and research impact. IZA Discussion Paper 18034.

Brodeur, A., Cook, N. and Neisser, C.: 2024, P-hacking, data type and data-sharing policy, *The Economic Journal* **134**(659), 985–1018.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y.: 2016, Star Wars: The Empirics Strike Back, *American Economic Journal: Applied Economics* **8**(1), 1–32.

Burlig, F.: 2018, Improving Transparency in Observational Social Science Research: A Pre-Analysis Plan Approach, *Economics Letters* **168**, 56–60.

Chopra, F., Haaland, I., Roth, C. and Stegmann, A.: 2024, The null result penalty, *The Economic Journal* **134**(657), 193–219.

Christensen, G. and Miguel, E.: 2018, Transparency, Reproducibility, and the Credibility of Economics Research, *Journal of Economic Literature* **56**(3), 920–80.

Doucouliagos, C. and Stanley, T. D.: 2013, Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity, *Journal of Economic Surveys* **27**(2), 316–339.

Elliott, G., Kudrin, N. and Wüthrich, K.: 2022a, Detecting p-hacking, *Econometrica* **90**(2), 887–906.

Elliott, G., Kudrin, N. and Wüthrich, K.: 2022b, The power of tests for detecting p-hacking.

Fang, A., Gordon, G. and Humphreys, M.: 2015, Does registration reduce publication bias? evidence from medical sciences. Unpublished manuscript, New York, NY: Columbia University.

Fitzpatrick, B. G., Gorman, D. M. and Trombatore, C.: 2024, Impact of Redefining Statistical Significance on p-Hacking and False Positive Rates: An Agent-Based Model, *PLoS One* **19**(5), e0303262.

Franco, A., Malhotra, N. and Simonovits, G.: 2014, Publication Bias in the Social Sciences: Unlocking the File Drawer, *Science* **345**(6203), 1502–1505.

Gerber, A. S. and Malhotra, N.: 2008, Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?, *Sociological Methods & Research* **37**(1), 3–30.

Harrington, D., D'Agostino Sr, R. B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S.-L. T., Drazen, J. M. and Hamel, M. B.: 2019, New Guidelines for Statistical Reporting in the Journal, *New England Journal of Medicine* **381**(3), 285–286.

Havránek, T.: 2015, Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting, *Journal of the European Economic Association* **13**(6), 1180–1204.

Havránek, T., Irsova, Z., Laslopova, L. and Zeynalova, O.: 2024, Publication and Attenuation Biases in Measuring Skill Substitution, *Review of Economics and Statistics* **106**(5), 1187–1200.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. and Jennions, M. D.: 2015, The extent and consequences of p-hacking in science, *PLoS Biology* **13**(3), e1002106.

Ioannidis, J. P., Stanley, T. D. and Doucouliagos, H.: 2017, The Power of Bias in Economics Research, *Economic Journal* **127**(605), F236–F265.

Kline, P. and Walters, C.: 2021, Reasonable doubt: Experimental detection of job-level employment discrimination, *Econometrica* **89**(2), 765–792.

Kudrin, N.: 2022, Robust caliper tests. Working paper, University of California San Diego.

Kudrin, N.: 2024, Testing for and evaluating the extent of selective reporting.

Lakens, D.: 2015, On the challenges of drawing conclusions from p-values just below 0.05, *PeerJ* **3**, e1142.

Masicampo, E. and Lalande, D. R.: 2012, A peculiar prevalence of p values just below. 05, *Quarterly Journal of Experimental Psychology* **65**(11), 2271–2279.

McCloskey, A. and Michaillat, P.: 2024, Critical Values Robust to p-Hacking, *Review of Economics and Statistics* pp. 1–35.

Naguib, C.: 2025, P-hacking and significance stars. Mimeo University of Bern.

Narasimhan, B. and Efron, B.: 2020, deconvolver: A g-modeling program for deconvolution and empirical bayes estimation, *Journal of Statistical Software* **94**, 1–20.

Scheel, A. M., Schijen, M. R. and Lakens, D.: 2021, An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports, *Advances in Methods and Practices in Psychological Science* **4**(2), 25152459211007467.

Simmons, J. P., Nelson, L. D. and Simonsohn, U.: 2011, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, *Psychological Science* **22**, 1359–1366.

Simonsohn, U., Nelson, L. D. and Simmons, J. P.: 2014, P-Curve: A Key to the File Drawer, *Journal of Experimental Psychology: General* **143**, 534–547.

Song, F., Hooper, L. and Loke, Y. K.: 2013, Publication Bias: What Is it? How do we Measure it? How do we Avoid it?, *Open Access Journal of Clinical Trials* pp. 71–81.

Van Aert, R. C., Wicherts, J. M. and Van Assen, M. A.: 2019, Publication Bias Examined in Meta-analyses from Psychology and Medicine: A Meta-Meta-Analysis, *PloS one* **14**(4), e0215052.

Vivalt, E.: 2019, Specification Searching and Significance Inflation Across Time, Methods and Disciplines, *Oxford Bulletin of Economics and Statistics* **81**(4), 797–816.

A Appendix Figures and Tables

Range	Mass
$p = 0.5$	0.203
$p < 0.5$	0.112
$p > 0.5$	0.685

Table A1: Moment-matching lower bound on p-hacking, with $K = 2$ moments matched.

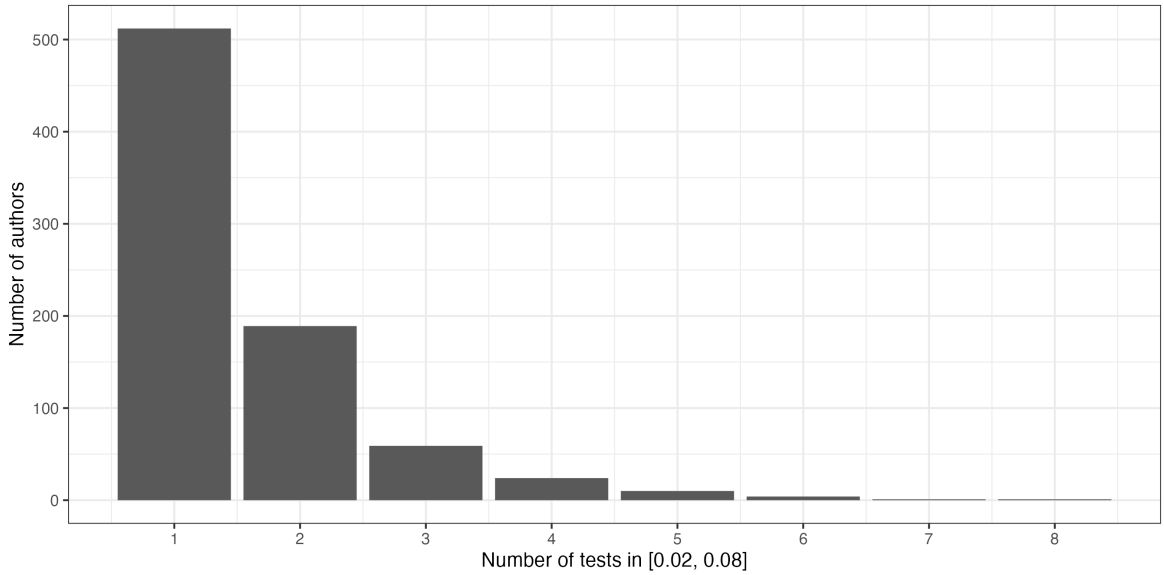

Figure A1: Number of tests per first author with p-values in $[0.02, 0.08]$

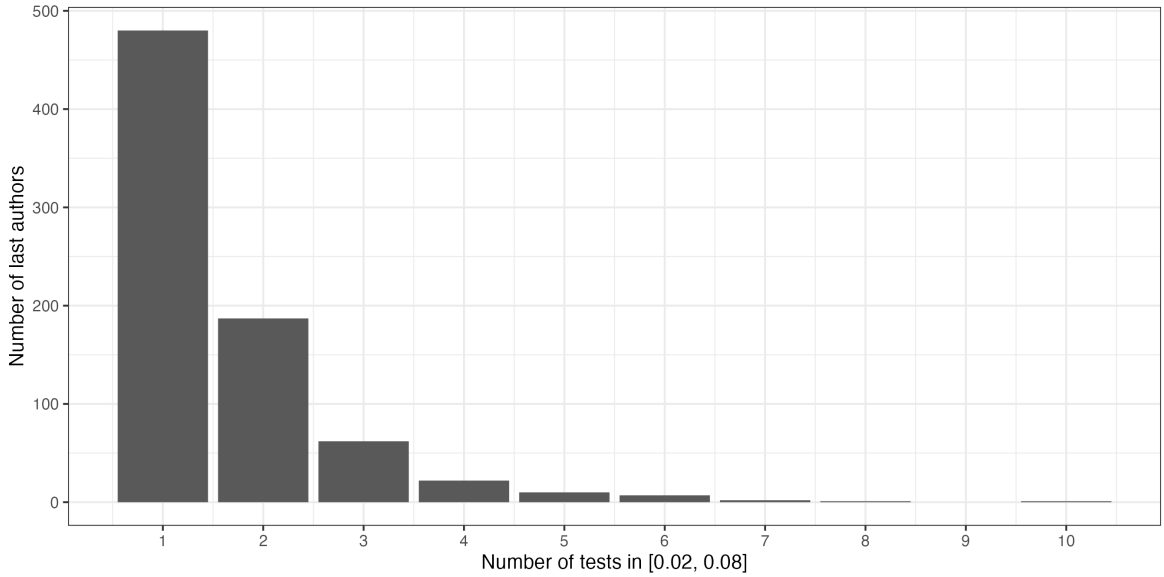

Figure A2: Number of tests per last author with p-values in $[0.02, 0.08]$

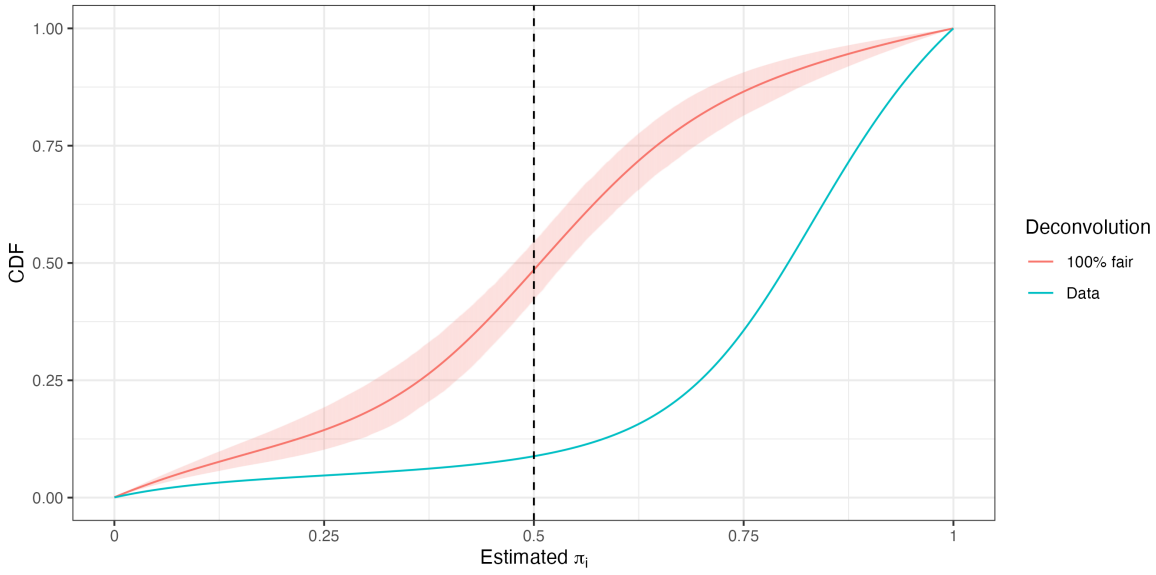

Figure A3: $\hat{G}$ estimated by deconvolution for last-listed authors, compared to deconvolutions from a fair distribution.

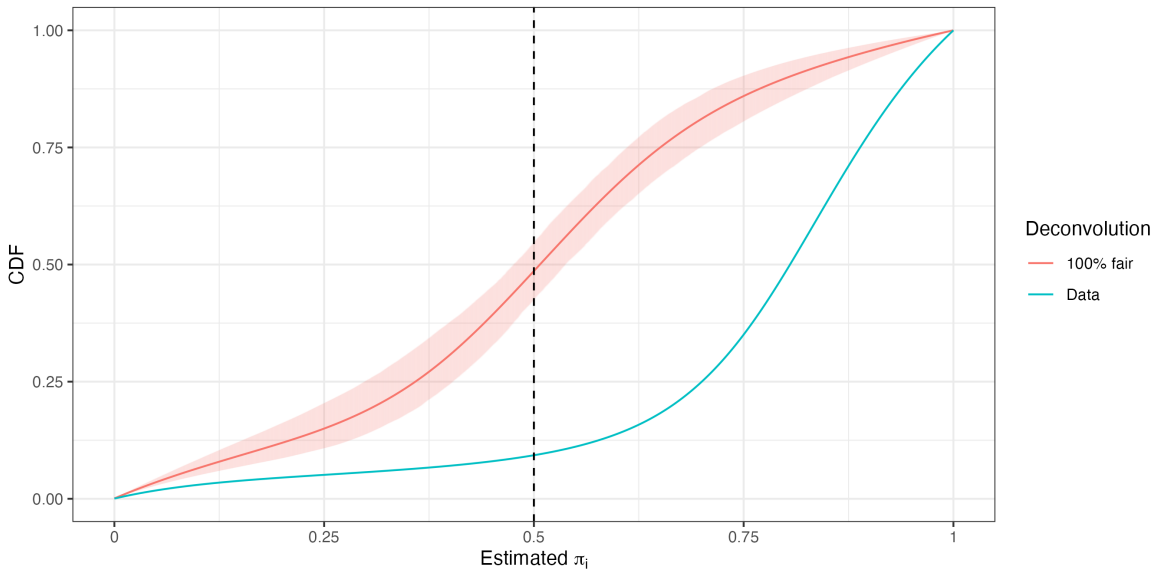

Figure A4: $\hat{G}$ estimated by deconvolution for p-values in the range $[0.03, 0.07]$

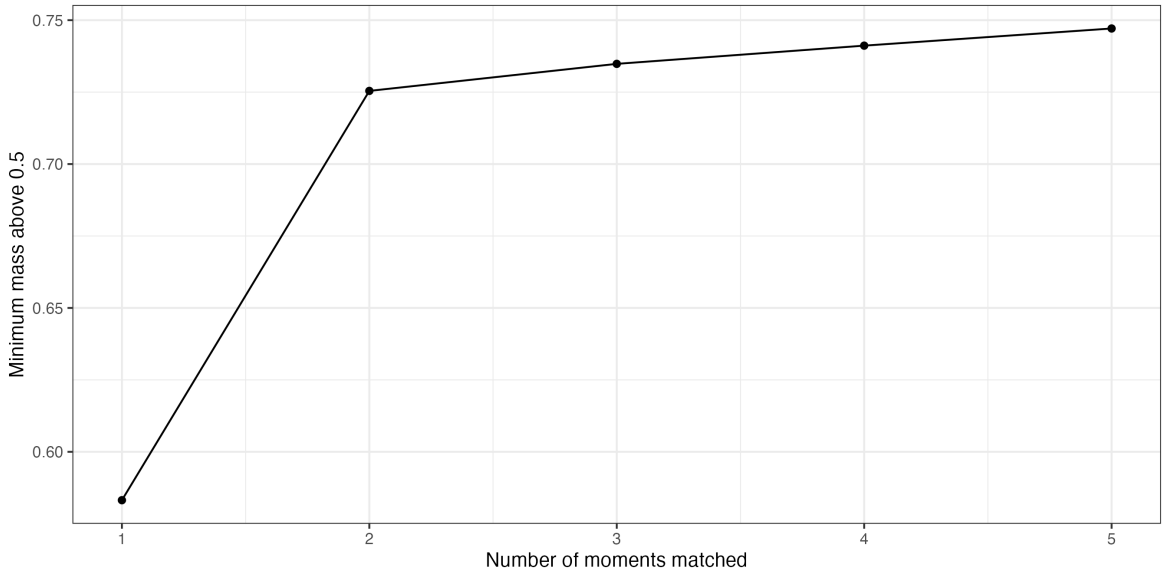

Figure A5: Sensitivity of lower bounds to the number of moments matched.

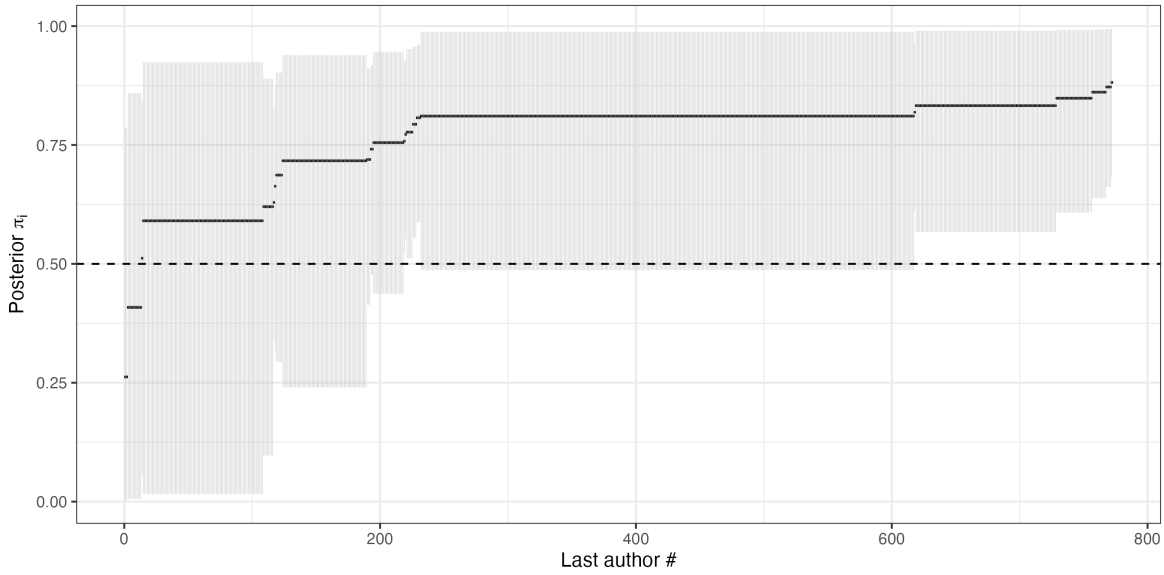

Figure A6: Distribution of posterior means and intervals across last-listed authors

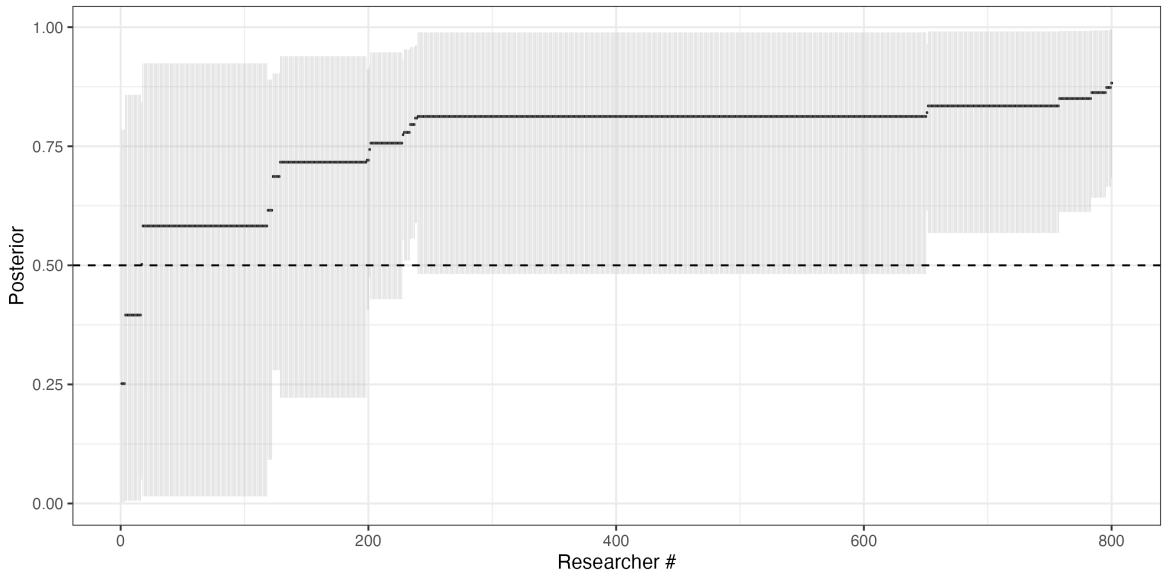

Figure A7: Distribution of posterior means and intervals across first authors for p-values in the range $[0.03, 0.07]$

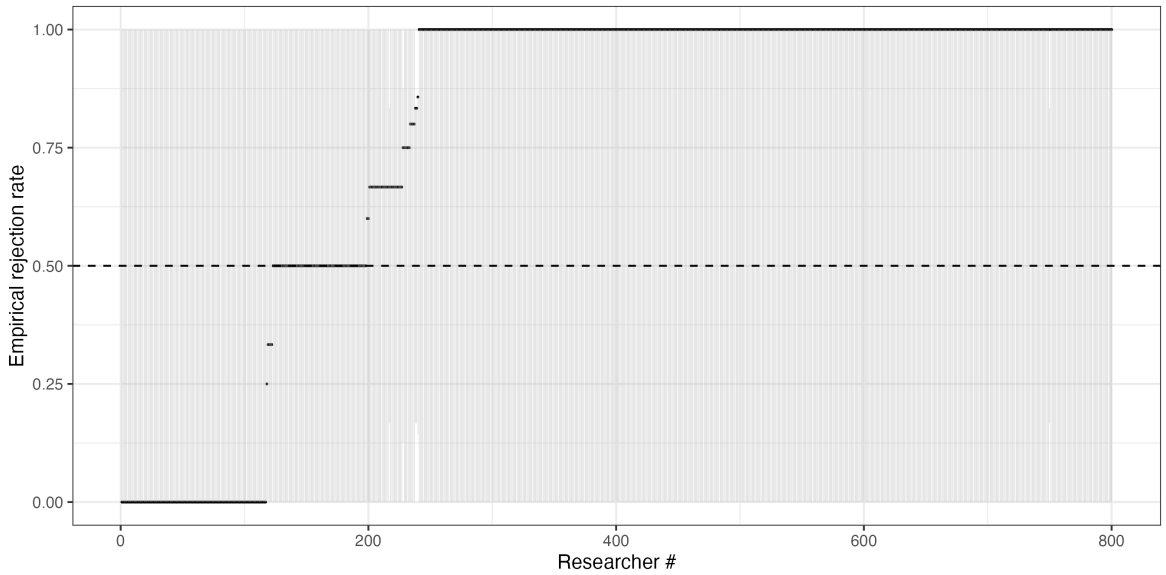

Figure A8: Distribution of 95% confidence intervals for empirical rejection rates under the null of $\pi_i = 0.5$.