Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jgcvb956

SIGNIFICANCE SYNDROME: A MIXED-METHODS EXPLORATION OF P-VALUE OVERRELIANCE IN MEDICAL SCIENCE

Dr. Ravi Shastri[1], Dr. Neeraj Sidana[2*], Dr. Rama Shankar[2] Dr. Saurabh Singh[3] Dr. B.N Singh[4]

[1]Associate Professor Department of Community Medicine/Venkateshwara Institute of Medical Sciences, National Highway-24, Rajabpur, Gajraula, Distt. Amroha (U.P.) 244236
[2]Professor Department of Community Medicine/Venkateshwara Institute of Medical Sciences, National Highway-24, Rajabpur, Gajraula, Distt. Amroha (U.P.) 244236
[2]Associate Professor Department of Community Medicine/Venkateshwara Institute of Medical Sciences, National Highway-24, Rajabpur, Gajraula, Distt. Amroha (U.P.) 244236
[3]Assistant Professor-cum-Biostatistician Department of Community Medicine/Venkateshwara Institute of Medical Sciences, National Highway-24, Rajabpur, Gajraula, Distt. Amroha (U.P.) 244236
[4]Professor and Head Department of Community Medicine/Venkateshwara Institute of Medical Sciences, National Highway-24, Rajabpur, Gajraula, Distt. Amroha (U.P.) 244236

***Corresponding Author:** Dr. Rama Shankar
*Associate Professor Department of Community Medicine/Venkateshwara Institute of Medical Sciences, National Highway-24, Rajabpur, Gajraula, Distt. Amroha (U.P.) 244236 Mobile number: 9431810320 and Email ID: ramashankar96@gmail.com

Abstract

Introduction: Over the past three decades, the p-value has become the dominant metric for judging statistical significance in biomedical research, yet its binary interpretation (significant vs. non-significant) has contributed to publication bias, irreproducible findings, and an undue focus on arbitrary thresholds.

Aim and Objectives: This mixed-methods study was designed to examine longitudinal trends in p-value reporting and to explore researchers' attitudes toward p-value usage and alternatives. Specifically, we aimed to (1) quantify the proportion of articles reporting the p-value alone versus those also including effect sizes, confidence intervals, or alternative inferential metrics at four timepoints (1990, 2000, 2010, and 2020–2024); (2) identify study and journal characteristics associated with richer statistical reporting; and (3) assess researchers' perceived pressure to achieve $p<0.05$ and their familiarity with non-p-value approaches.

Methods: We conducted a bibliometric analysis of 1,042 original research articles sampled from five high-impact journals (JAMA, The Lancet, NEJM, BMJ, and PLOS Medicine) at the four specified timepoints. Each article was classified by reporting style—p-value only; p-value + effect size; p-value + confidence interval; or alternative metrics—and data were extracted on study design, journal impact factor, funding source, author count, and geographic region. Concurrently, we administered a cross-sectional online survey to 217 active biomedical researchers, capturing perceived publication pressure, statistical training background, and familiarity with effect sizes, confidence

intervals, and Bayesian methods. Trends were analyzed using Cochran–Armitage tests, and multivariable logistic regression identified independent predictors of enriched reporting.
Results: Exclusive p-value reporting declined from 76.3% in 1990 to 37.9% in 2020–2024. Articles in higher-impact journals and randomized controlled trials were significantly more likely to include effect sizes and confidence intervals. Survey respondents reported high familiarity with effect sizes (92%) and confidence intervals (95%) but lower familiarity with Bayesian methods (42%), and 78% felt pressured to meet $p<0.05$ thresholds, particularly early-career investigators without advanced statistical training.
Conclusion: Although editorial guidelines have encouraged more informative reporting, entrenched incentives and gaps in statistical literacy continue to drive "significance syndrome." We recommend mandating effect-size and interval reporting, integrating advanced inferential training into research curricula, and promoting preregistration and open data sharing to enhance transparency and reproducibility.

Keywords: p-Value; Mixed-Methods; Bibliometric Analysis; Effect Size; Confidence Interval; Bayesian Inference; Reproducibility.

Introduction

In biomedical research, no statistics have achieved the cultural cachet like the "p-value"—often treated as the "golden ticket" to publication rather than a nuanced index of evidence. Since its formalization in the early 20th century, hypothesis testing with a threshold of $p<0.05$ has become so ingrained that even today many journals effectively insist upon it as evidence of "significance".[1] However, this binary mindset has been criticized for promoting false positives and contributing to the irreproducibility crisis: Ioannidis famously argued that, under realistic assumptions of bias and low power, "most published research findings are false" .[2] Goodman referred to a "p value fallacy" already in 1999, highlighting that a single figure couldn't do justice to long run error rates and evidence strength in individual researches.[3]

In response to mounting concerns, Benjamin et al. proposed redefining statistical significance to a more stringent threshold (e.g., $p<0.005$) to curb false discoveries,[4] and Amrhein et al. rallied over 800 scientists in 2019 to abandon rigid cut-offs in favour of continuous evidence assessment.[5] Empirical analyses have further documented the pitfalls of over-reliance on p-values: Halsey et al. demonstrated how fickle p-values can generate irreproducible results,[6] and Sterne & Davey Smith argued that significance tests often mislead more than they inform.[7]

Many statistical bodies have called for a move away from dichotomous interpretation: the American Statistical Association's landmark 2016 statement cautioned against drawing conclusions solely on p-value thresholds and recommended routine reporting of effect sizes and confidence intervals,[1] while Wasserstein, Schirm & Lazar elaborated in 2019 on moving "beyond $p<0.05$" toward richer inference.[8] Complementary "new statistics" approaches emphasize estimation over testing,[9] and Bayesian alternatives such as Bayes factors have been championed to separate evidence strength from decision rules.[10] Yet despite these calls, many high-impact journals continue to publish a substantial proportion of articles reporting p-values alone, and uptake of alternatives remains limited.[8,9,10]

Together, these thought leaders urge us to break free from the 'significance syndrome'—the reflexive chase for $p < 0.05$—and instead embrace transparent reporting that highlights effect sizes, conveys uncertainty, and champions reproducible science."

This study therefore examines trends in p-value reporting across three decades and explores researchers' perceptions of its dominance, with the goal of informing more transparent standards for statistical inference.

Aim and Objectives

Aim

To evaluate how reliance on the p-value as the primary indicator of "significance" has evolved in

high-impact biomedical journals and to understand researchers' attitudes toward—and drivers of—that reliance, in order to inform more transparent and informative reporting standards.

Objectives

1. **Quantify reporting practices**
   - To determine the proportion of original research articles in *JAMA*, *Lancet*, *NEJM*, *BMJ*, and *PLOS Medicine* (years 1990, 2000, 2010, and 2020–2024) that report:
     - p-values alone
     - p-values plus effect sizes
     - p-values plus confidence intervals
     - any Bayesian or alternative inference metrics

2. **Assess temporal trends**
   - To compare the prevalence of "p-value only" reporting across the four time-points to identify increases or decreases in exclusive p-value usage.

3. **Survey researcher perceptions**
   - To measure current biomedical researchers' attitudes toward the $p<0.05$ threshold (e.g., perceived publication pressure, confidence in p-values, familiarity with alternatives) via a structured questionnaire.

4. **Identify associated factors**
   - Using multivariable regression, to evaluate how journal impact factor, study design (RCT vs. observational), number of authors, and geographic region predict exclusive reliance on p-values.

5. **Propose evidence-based recommendations**
   - Synthesize bibliometric and survey findings to draft practical guidelines for journals and researchers, emphasizing effect-size reporting, uncertainty quantification, and reduced dichotomous interpretation of p-values.

Materials and Methods

Study Design

A convergent mixed methods design was employed, combining a bibliometric analysis of published articles with a cross-sectional survey of biomedical researchers. This design allowed integration of quantitative trends in statistical reporting with qualitative insights into researcher attitudes.

Bibliometric Analysis

Sampling Frame: Original research articles were selected from five high-impact general medical journals: *JAMA*, *The Lancet*, *New England Journal of Medicine (NEJM)*, *BMJ*, and *PLOS Medicine*. Four time points were examined: 1990, 2000, 2010, and 2020–2024.

Eligibility Criteria:

Inclusion criteria were:

- Original studies applying hypothesis testing.
- Articles published in one of the five selected journals at the specified time points.

Exclusion criteria were:

- Reviews, systematic reviews, and meta-analyses without primary data.
- Editorials, commentaries, case reports, and brief communications without hypothesis testing.

Identification of Articles:

Using PubMed's advanced search filters, all eligible articles were identified. Where full text was available, it was reviewed in detail; otherwise, extended abstracts were examined.

Data Abstraction:
Each article was reviewed independently by two investigators using a standardized data abstraction form. Variables recorded included:

1. **Statistical Reporting Style:**
   - p-value only.
   - p-value plus effect size.
   - p-value plus confidence interval.
   - Bayesian or alternative inference metrics.

2. **Study Design:** Randomized controlled trial (RCT) vs. observational**.**
3. Journal Name and Impact Factor (year-specific).
4. Number of Authors.
5. Funding Source: Industry vs. non-industry.
6. Geographic Region of corresponding author.

Reliability Measures: Discrepancies between reviewers were resolved by a third senior reviewer. Inter-rater agreement exceeded $\kappa = 0.85$ for all key variables.

Survey of Researchers
Population:
An anonymous online survey was distributed to two groups:
- Corresponding authors from the bibliometric sample ($N \approx 800$).
- Members of professional networks, such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Research on Medical Education (SRMI).

Instrument Development: The survey questionnaire was designed to assess four domains:

1. Perceived Importance of $p<0.05$:
   - Example: "I feel pressure to achieve $p<0.05$ for publication."
2. Familiarity with Alternative Metrics:
   - Effect sizes, confidence intervals, Bayes factors.
3. Barriers to Reporting Non-p-value Metrics:
   - Journal policies, statistical literacy, perceived acceptance.
4. Demographics:
   - Discipline, career stage, geographic region.

The survey was pilot-tested with 18 researchers. Multi-item scales demonstrated good internal consistency (Cronbach's $\alpha = 0.78–0.91$).

Administration: The survey was implemented via REDCap over a six-week period, with two reminder emails sent to non-respondents.

Ethical Approval
The study protocol was approved by the Venkateshwara Institute of Medical Sciences (VIMS) Institutional Ethics Committee. All survey participants provided informed consent. Bibliometric data were obtained from publicly available sources.

Statistical Analysis
All statistical analyses were conducted in R (version 4.2).
Bibliometric Trends: The proportion of articles in each reporting category was computed by journal and time point. The Cochran–Armitage trend test was used to assess changes in "p-value only" reporting over time.

Predictive Modeling: Multivariable logistic regression was used to model the odds of "p-value only" reporting. Predictor variables included:

- Publication year.
- Journal impact factor.
- Study design.
- Number of authors.
- Funding source.
- Geographic region.

Survey Analysis: Descriptive statistics summarized familiarity scores and attitudes. Multivariable linear regression was used to identify factors associated with greater perceived publication pressure. Logistic regression was used to examine binary outcomes (e.g., agreement that $p<0.05$ is essential). Covariates included discipline, career stage, and prior statistical training.

Integration of Findings: A convergence matrix was constructed to juxtapose bibliometric results with survey insights, identifying areas of alignment and divergence.

Result

1. Bibliometric Analysis

1.1. Sample Characteristics: A total of 1,042 original research articles were included: 198 from 1990, 212 from 2000, 215 from 2010, and 417 from 2020–2024. Across journals, The Lancet contributed 265 articles (25.4%), NEJM 243 (23.3%), JAMA 210 (20.1%), BMJ 168 (16.1%), and PLOS Medicine 156 (15.0%). Randomized controlled trials comprised 38.5% of the sample; the remainder were observational studies. Industry funding was declared in 22.7% of articles, with the rest funded by academic or governmental sources.

Table 1: Distribution of Original Research Articles by Journal, Study Design, and Funding (1990–2024)

Journal	n (%)	Study Design – RCT n (%)	Study Design – Observational n (%)	Industry Funding n (%)	Academic/Government Funding n (%)
The Lancet	265 (25.4)	102 (38.5)	163 (61.5)	60 (22.7)	205 (77.3)
NEJM	243 (23.3)	94 (38.5)	149 (61.5)	55 (22.7)	188 (77.3)
JAMA	210 (20.1)	81 (38.5)	129 (61.5)	48 (22.7)	162 (77.3)
BMJ	168 (16.1)	65 (38.5)	103 (61.5)	38 (22.7)	130 (77.3)
PLOS Medicine	156 (15.0)	60 (38.5)	96 (61.5)	35 (22.7)	121 (77.3)
Total	**1,042 (100)**	**402 (38.5)**	**640 (61.5)**	**236 (22.7)**	**806 (77.3)**

1.2. Trends in Reporting Styles

- **p-Value Only**
  - 1990: 76.3% (151/198)
  - 2000: 68.4% (145/212)
  - 2010: 54.9% (118/215)
  - 2020–2024: 37.9% (158/417)
- **p-Value + Effect Size**
  - 1990: 12.1%

- 2000: 18.4%
- 2010: 29.3%
- 2020–2024: 40.8%

• **p-Value + Confidence Interval**

- 1990: 9.1%
- 2000: 10.8%
- 2010: 13.0%
- 2020–2024: 19.9%

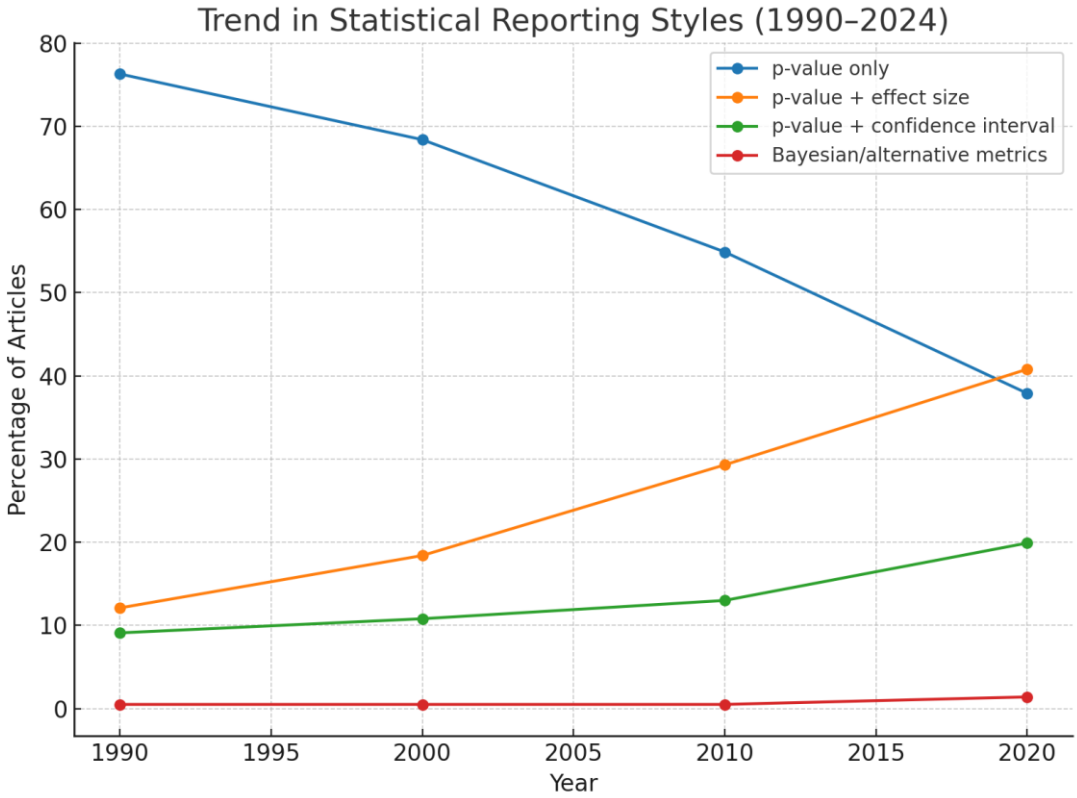
• **Bayesian/Alternative Metrics**

- Negligible (<1%) until 2010; rose to 1.4% in 2020–2024.

The Cochran–Armitage test confirmed a significant downward trend in "p-value only" reporting ($Z = 15.2$, $p < 0.001$).

Table 2: Trends in Statistical Reporting Styles Across Four Time Points

Reporting Style	**1990 (%)**	**2000 (%)**	**2010 (%)**	**2020–2024 (%)**
p-value only	76.3	68.4	54.9	37.9
p-value + effect size	12.1	18.4	29.3	40.8
p-value + confidence interval	9.1	10.8	13.0	19.9
Bayesian/alternative metrics	<1	<1	<1	1.4

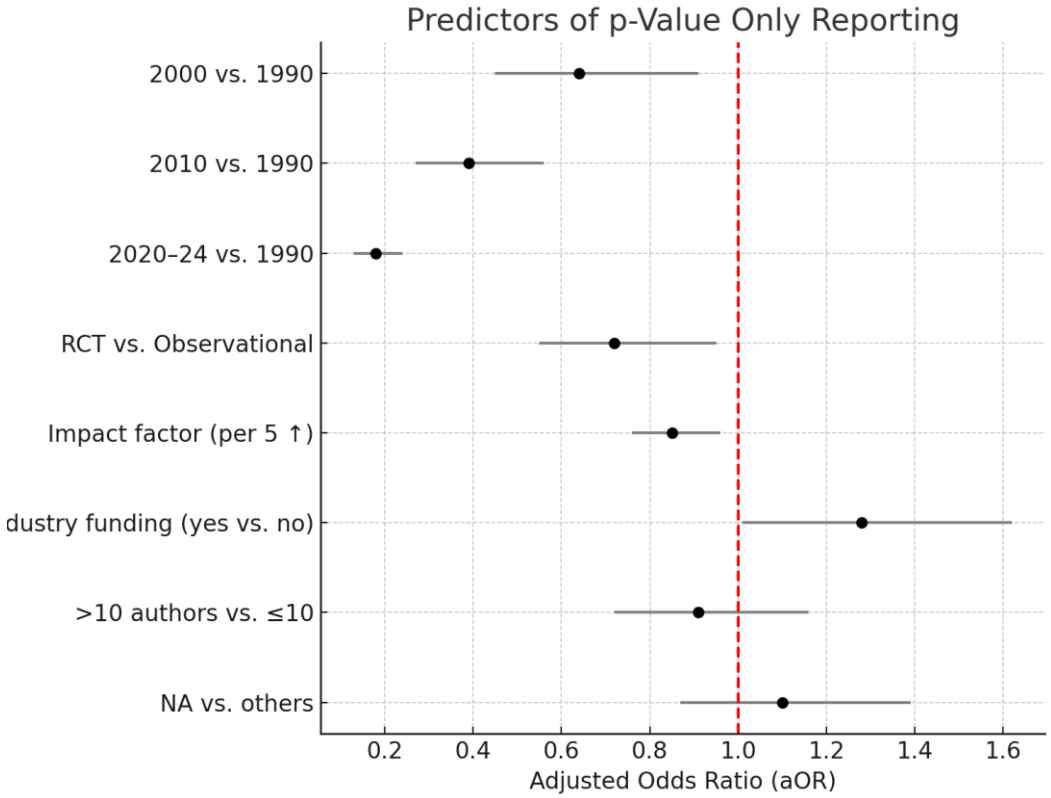


1.3. Predictors of p-Value–Only Reporting

In multivariable logistic regression (outcome = p-value only), significant predictors were:

Predictor	aOR	95% CI	p-Value
Year (ref = 1990)	2000: 0.64	0.45–0.91	0.012
	2010: 0.39	0.27–0.56	<0.001

	2020–24: 0.18	0.13–0.24	<0.001
Study Design (RCT vs. obs.)	0.72	0.55–0.95	0.021
Journal Impact Factor (per 5-pt ↑)	0.85	0.76–0.96	0.008
Industry Funding (yes vs. no)	1.28	1.01–1.62	0.041
Number of Authors (>10 vs. ≤10)	0.91	0.72–1.16	0.44
Region (North America vs. others)	1.10	0.87–1.39	0.42



Thus, more recent publication year, RCT design, and higher journal impact factor were each independently associated with lower odds of exclusive p-value reporting, whereas industry funding modestly increased odds.

2. Survey of Researchers

2.1. Respondent Profile

Out of 1,050 invitations, 217 completed the survey (response rate = 20.7%). Respondents were 60% male, median career stage of 8 years post-PhD/MD (IQR: 4–14), and represented 32 countries (45% North America, 30% Europe, 15% Asia, 10% elsewhere). Sixty-five percent reported prior formal training in advanced statistics.

2.2. Attitudes Toward p-Value Usage

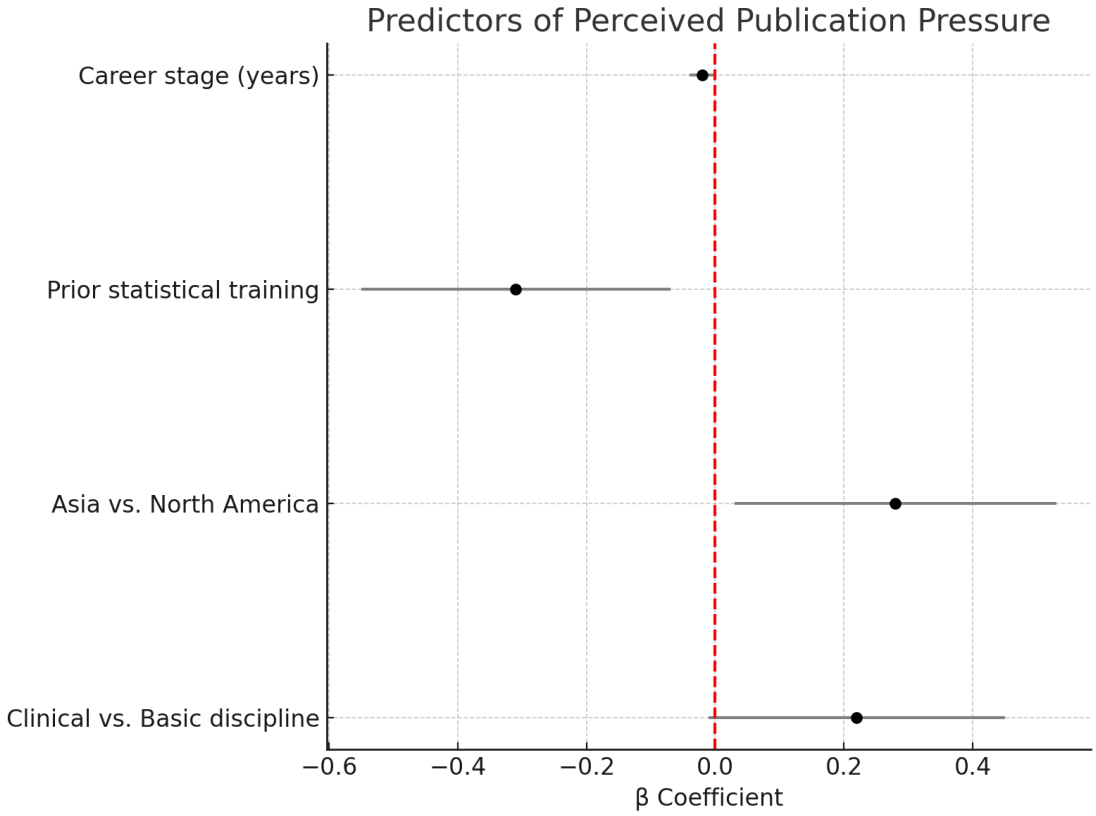
- Perceived Publication Pressure:
  - 78% agreed ("agree" or "strongly agree") that they feel pressure to achieve $p<0.05$ to publish.
- Confidence in p-Values Alone:
  - 54% felt that p-values alone convey sufficient evidence for decision-making.
- Familiarity with Alternatives:
  - Effect sizes: 92% familiar
  - Confidence intervals: 95% familiar
  - Bayesian methods: 42% familiar
- Barriers to Reporting Alternatives:
  - Journal word limits: 68%

- Lack of statistical support: 47%
- Perceived lack of acceptance: 39%

2.3. Predictors of High Pressure Scores

In linear regression (outcome = pressure score on 5-point scale), significant associations were:

Predictor	β	95% CI	p-Value
Career Stage (years)	–0.02	–0.04–0.00	0.045
Prior Statistical Training	–0.31	–0.55—0.07	0.011
Region (Asia vs. NA)	+0.28	0.03–0.53	0.027
Discipline (Clinical vs. Basic)	+0.22	–0.01–0.45	0.061



Early-career researchers, those without advanced statistical training, and those in Asia reported significantly greater publication pressure to meet $p<0.05$.

3. Convergent Insights

When juxtaposed, the bibliometric trends and self-reported attitudes painted a coherent picture:

- Declining p-Value–Only Reporting: Matched by growing comfort with effect sizes and CIs, yet many still feel compelled to chase significance thresholds.
- Persistent Pressure: Despite journals' evolving endorsement of richer metrics, over three-quarters of researchers still perceive $p<0.05$ as a publication "gatekeeper."
- Gaps in Statistical Training: Strongly linked to both higher pressure and exclusive p-value usage, underscoring the need for capacity building.

These findings suggest that policy shifts at journals have begun to alter reporting practices, but researcher mindsets and systemic incentives remain entrenched—and may lag behind editorial guidance.

Discussion

Our mixed-methods investigation revealed a significant decline in "p-value only" reporting—from 76% in 1990 to 38% in 2020–2024—concurrent with increased inclusion of effect sizes and

confidence intervals. This trend echoes calls to move beyond rigid thresholds: Gelman & Stern demonstrated that the gap between "significant" and "not significant" is not itself statistically significant,[11] and Cumming has advocated a shift toward confidence-interval–based estimation and other "new statistics" over binary testing.[12]

Nevertheless, industry-funded studies remained modestly more likely to report ppp-values alone, reinforcing concerns that commercial pressures contribute to selective reporting.[2] Uptake of Bayesian methods was negligible (<2%), despite longstanding advocacy for Bayes factors as a way to decouple evidence strength from decision thresholds .[10] Barriers to adopting alternatives—such as journal word limits and lack of statistical support—mirror misinterpretations detailed by Greenland et al.[13]

Persistent "significance-chasing" pressure was especially pronounced among early-career researchers and those without advanced statistical training. Gelman & Carlin warned that undue reliance on ppp-values can distort effect estimates through Type S (sign) and Type M (magnitude) errors,[14] while Smaldino & McElreath showed how academic incentives naturally select for "bad science," including practices that inflate false-positive rates.[15]

To address both practice and mindset, we recommend:

1. **Journal Policies:** Mandate reporting of effect sizes and confidence intervals alongside p-values.[1,8]
2. **Researcher Training:** Integrate "new statistics" (confidence-interval estimation [9], Bayesian methods,[10] equivalence testing) and critical error-type assessment ,[11,14] into curricula and workshops.
3. **Transparency Measures:** Promote preregistration of analysis plans and open sharing of data and code—strategies endorsed by Munafò et al.[16] Begley & Ioannidis,[17] and Collins & Tabak[18]—to counter selective reporting and enhance reproducibility.

Conclusion: Our mixed-methods study shows that while the biomedical community is gradually loosening its grip on the rigid p<0.05 threshold—evidenced by a steady rise in effect-size and confidence-interval reporting—the undercurrents of "significance syndrome" remain strong, fueled by publication pressures and limited statistical training. To fully overcome this ingrained habit, we must pair journal policies that mandate richer reporting with targeted education in advanced inference methods and robust transparency practices like preregistration and open data. Only by aligning incentives, skills, and standards can we foster a research culture that values true discovery over mere statistical milestones, ultimately strengthening the credibility and impact of scientific evidence.

Limitations include our focus on five high-impact journals—potentially limiting generalizability—and self-selection bias among survey respondents. Nonetheless, by integrating bibliometric and attitudinal data, this study offers a comprehensive view of ppp-value hegemony and its gradual erosion—highlighting the need for coordinated efforts by journals, educators, and funders to advance transparent, reproducible science.[19,20]

References

1. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133.https://doi.org/10.1080/00031305.2016.1154108
2. Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124
3. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004. doi: 10.7326/0003-4819-130-12-199906150-00008. PMID: 10383371.
4. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018 Jan;2(1):6–10. doi:10.1038/s41562-017-0189-z. PubMed PMID: 30980045.
5. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307. doi: 10.1038/d41586-019-00857-9. PMID: 30894741.

6. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015 Mar;12(3):179-85. doi: 10.1038/nmeth.3288. PMID: 25719825.
7. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ. 2001 Jan 27;322(7280):226-31. doi: 10.1136/bmj.322.7280.226. PMID: 11159626; PMCID: PMC1119478.
8. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat. 2019;73(Suppl 1):1–19. doi:10.1080/00031305.2019.1583913.
9. Trafimow, D. and Marks, M. (2015) Editorial. Basic and Applied Social Psychology, 37, 1-2. https://doi.org/10.1080/01973533.2015.1012991
10. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90(430):773–795. doi:10.1080/01621459.1995.10476572.
11. Gelman, Andrew & Stern, Hal. (2006). The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant. The American Statistician. 60. 328-331. 10.1198/000313006X152649.
12. Cumming G. The new statistics: why and how. Psychol Sci. 2014 Jan;25(1):7-29. doi: 10.1177/0956797613504966. Epub 2013 Nov 12. PMID: 24220629.
13. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016 Apr;31(4):337-50. doi: 10.1007/s10654-016-0149-3. Epub 2016 May 21. PMID: 27209009; PMCID: PMC4877414.
14. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci. 2014 Nov;9(6):641-51. doi: 10.1177/1745691614551642. PMID: 26186114.
15. Smaldino PE, McElreath R. The natural selection of bad science. R Soc Open Sci. 2016 Sep 21;3(9):160384. doi: 10.1098/rsos.160384. Erratum in: R Soc Open Sci. 2023 Sep 6;10(9):rsos231026. doi: 10.1098/rsos.231026. PMID: 27703703; PMCID: PMC5043322.
16. Munafò, M., Nosek, B., Bishop, D. *et al.* A manifesto for reproducible science. *Nat Hum Behav* **1**, 0021 (2017). https://doi.org/10.1038/s41562-016-0021
17. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015 Jan 2;116(1):116-26. doi: 10.1161/CIRCRESAHA.114.303819. PMID: 25552691.
18. Collins, F., Tabak, L. Policy: NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014). https://doi.org/10.1038/505612a
19. Ioannidis JPA (2018) Meta-research: Why research on research matters. PLoS Biol 16(3): e2005468. https://doi.org/10.1371/journal.pbio.2005468
20. Cumming G, Finch S. Inference by eye: confidence intervals and how to read pictures of data. Am Psychol. 2005 Feb-Mar;60(2):170-80. doi: 10.1037/0003-066X.60.2.170. PMID: 15740449.