

Original Investigation

Evolution of Reporting *P* Values in the Biomedical Literature, 1990-2015

David Chavalarias, PhD; Joshua David Wallach, BA; Alvin Ho Ting Li, BHSc; John P. A. Ioannidis, MD, DSc

IMPORTANCE The use and misuse of *P* values has generated extensive debates.

OBJECTIVE To evaluate in large scale the *P* values reported in the abstracts and full text of biomedical research articles over the past 25 years and determine how frequently statistical information is presented in ways other than *P* values.




DESIGN Automated text-mining analysis was performed to extract data on *P* values reported in 12 821 790 MEDLINE abstracts and in 843 884 abstracts and full-text articles in PubMed Central (PMC) from 1990 to 2015. Reporting of *P* values in 151 English-language core clinical journals and specific article types as classified by PubMed also was evaluated. A random sample of 1000 MEDLINE abstracts was manually assessed for reporting of *P* values and other types of statistical information; of those abstracts reporting empirical data, 100 articles were also assessed in full text.

MAIN OUTCOMES AND MEASURES *P* values reported.

RESULTS Text mining identified 4 572 043 *P* values in 1 608 736 MEDLINE abstracts and 3 438 299 *P* values in 385 393 PMC full-text articles. Reporting of *P* values in abstracts increased from 7.3% in 1990 to 15.6% in 2014. In 2014, *P* values were reported in 33.0% of abstracts from the 151 core clinical journals (*n* = 29 725 abstracts), 35.7% of meta-analyses (*n* = 5620), 38.9% of clinical trials (*n* = 4624), 54.8% of randomized controlled trials (*n* = 13 544), and 2.4% of reviews (*n* = 71 529). The distribution of reported *P* values in abstracts and in full text showed strong clustering at *P* values of .05 and of .001 or smaller. Over time, the "best" (most statistically significant) reported *P* values were modestly smaller and the "worst" (least statistically significant) reported *P* values became modestly less significant. Among the MEDLINE abstracts and PMC full-text articles with *P* values, 96% reported at least 1 *P* value of .05 or lower, with the proportion remaining steady over time in PMC full-text articles. In 1000 abstracts that were manually reviewed, 796 were from articles reporting empirical data; *P* values were reported in 15.7% (125/796 [95% CI, 13.2%-18.4%]) of abstracts, confidence intervals in 2.3% (18/796 [95% CI, 1.3%-3.6%]), Bayes factors in 0% (0/796 [95% CI, 0%-0.5%]), effect sizes in 13.9% (111/796 [95% CI, 11.6%-16.5%]), other information that could lead to estimation of *P* values in 12.4% (99/796 [95% CI, 10.2%-14.9%]), and qualitative statements about significance in 18.1% (181/1000 [95% CI, 15.8%-20.6%]); only 1.8% (14/796 [95% CI, 1.0%-2.9%]) of abstracts reported at least 1 effect size and at least 1 confidence interval. Among 99 manually extracted full-text articles with data, 55 reported *P* values, 4 presented confidence intervals for all reported effect sizes, none used Bayesian methods, 1 used false-discovery rates, 3 used sample size/power calculations, and 5 specified the primary outcome.

CONCLUSIONS AND RELEVANCE In this analysis of *P* values reported in MEDLINE abstracts and in PMC articles from 1990-2015, more MEDLINE abstracts and articles reported *P* values over time, almost all abstracts and articles with *P* values reported statistically significant results, and, in a subgroup analysis, few articles included confidence intervals, Bayes factors, or effect sizes. Rather than reporting isolated *P* values, articles should include effect sizes and uncertainty metrics.

JAMA. 2016;315(11):1141-1148. doi:10.1001/jama.2016.1952
Corrected on May 12, 2016.

-  Editorial page 1113
-  Supplemental content at jama.com
-  CME Quiz at jamanetworkcme.com

Author Affiliations: Centre d'Analyse et de Mathématiques Sociales (CAMS), EHESS-CNRS UMR8557 and Complex Systems Institute of Paris Île-de-France (ISC-PIF, UPS3611), Paris, France (Chavalarias); Departments of Health Research and Policy and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California (Wallach); Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada (Li); Departments of Medicine, Health Research and Policy, and Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California (Ioannidis).

Corresponding Author: John P. A. Ioannidis, MD, DSc, Departments of Medicine, Health Research and Policy, and Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, 1265 Welch Rd, MSOB X306, Stanford, CA 94305 (jioannid@stanford.edu).

Many research fields in biomedicine and other disciplines use statistical testing methods that report *P* values to convey inferences about study results. There is increasing concern that *P* values are often misused, misunderstood, and miscommunicated.¹⁻⁶ Moreover, there is mounting evidence from diverse fields that reporting biases tend to preferentially select the publication and highlighting of results that are statistically significant, as opposed to “negative” results.⁶⁻¹² Such biases could have major implications for the reliability of the published scientific literature.

The aim of this study was to assess the reporting of *P* values across the biomedical literature over the past 25 years, evaluate the evolution of the use and reporting of *P* values over time in the overall biomedical literature and in specific types of articles, and determine how frequently statistical information is presented in other ways besides *P* values.

Methods

Eligible Articles for Text Mining

Automated text-mining analysis was performed on the entire MEDLINE database since 1990 and on a random sample of the PubMed Central (PMC) database. Both databases are freely accessible to the public. MEDLINE is the free bibliographic database of life sciences and biomedical information compiled by the US National Library of Medicine. PMC is a free repository of publicly accessible full-text scholarly articles from biomedical and life sciences journals.

For the MEDLINE text mining, all *P* value data were extracted from the MEDLINE archives from January 1, 1990, to June 4, 2015, for all items with article meta-data that have an abstract. The same methodology was applied to a random sample of the PMC database (downloaded March 20, 2015) consisting of full-text articles with an abstract.

We defined a *P* value report as a string starting with either “p,” “P,” “p-value(s),” “P-value(s),” “P value(s),” or “p value(s),” followed by an equality or inequality expression (any combination of =, <, >, ≤, ≥, “less than,” or “of <”) and then by a value, which could include also exponential notation (for example, 10⁻⁴, 10(-4), E-4, (-4), or e-4). See the eAppendix in the Supplement for further details of the structure of the generated data sets of *P* values, the process followed to homogenize *P* value expressions, and technical issues and validation of the results yielded by the automated text mining vs in-depth manual extraction of information.

Evaluated Categories of Articles

Besides the analysis including all MEDLINE abstracts and PMC articles with an abstract, specific predetermined categories of articles that may be most important for clinical medicine were examined separately. These categories were the subset of the *Abridged Index Medicus* journals (a list of 151 English-language core clinical journals [<https://www.nlm.nih.gov/bsd/aim.html>]) as well as the articles included in the categories clinical trial, randomized controlled trial, meta-analysis, and review that were classified as such by PubMed. The “core clinical journals” category includes articles of various study designs that are all published in these journals. To avoid overlap in the results, data in

the clinical trials category exclude randomized clinical trials, and data in the reviews category exclude meta-analyses.

Main Analyses

The following characteristics of *P* value were evaluated: (1) the proportion of abstracts and full texts of articles that include *P* values and whether this is increasing over time; (2) the distribution of the reported *P* values, focusing in particular on the extent of reporting of very small *P* values ($\leq .001$) vs the conventional *P* value of .05 (long considered a threshold of formal statistical significance); (3) the evolution of the minimal (best, most statistically significant) and maximal (worst, least statistically significant) reported *P* values across abstracts and full-text articles; and (4) the number and proportion of abstracts and full-text articles that included at least 1 *P* value $\leq .05$.

In-depth Manual Assessment of Random Samples

Data were manually extracted in duplicate from a random sample of 1000 abstracts drawn from MEDLINE articles with abstracts. Two reviewers (J.D.W., A.H.T.L.) extracted data independently and then compared data extractions; persisting discrepancies were resolved through discussion with a third reviewer (J.P.A.I.). In each of these abstracts, we assessed reporting of any Bayes statistics, any *P* values (and, if so, how many), any statistically significant *P* values ($< .05$), any CI, any effect sizes (and, if so, what type, eg, odds ratio, hazards ratio), and any other information that would allow the calculation of effect sizes (eg, a comparison of proportions). The number of abstracts that had at least 1 effect size reported along with at least 1 corresponding *P* value or CI also was recorded. In addition, the frequency of qualitative statements about significance without reporting at least 1 corresponding *P* value, whether these statements were positive (eg, “was statistically significant”) or negative (eg, “was nonsignificant”), and whether there was any qualification of what type of significance was alluded to (statistical, clinical, biological, other) were recorded.

Of the 1000 abstracts, 796 were from articles reporting empirical data and, based on the abstract, reviewers felt that reporting of *P* values of effects in the full text could not be excluded; the others were from expert reviews or case reports where such reporting could reasonably be excluded. Of those 796, 100 were randomly selected and examined to determine whether the full text of the articles clearly specified the primary outcome(s) of interest and reported any *P* values, effect sizes, CIs, Bayesian methods or statistics, false-discovery rates (*q*-statistics), or sample size/power calculations.

Results

Reporting *P* Values in MEDLINE Abstracts

From January 1, 1990, to June 4, 2015, the MEDLINE archives included 16 013 338 items with article meta-data, of which 12 821 790 (80%) had an abstract, including 1 608 736 abstracts that reported *P* values. From these abstracts, a total of 4 572 043 *P* values were extracted.

The number of articles published in MEDLINE increased from 408 551 in 1990 to 1 189 664 in 2014, a relative increase of 4.5%

per year, and the number of abstracts reporting at least 1 *P* value increased from 20 769 in 1990 (7.3%) to 138 654 in 2014 (15.6%), a relative increase of 8.2% per year (eFigure 1 in the Supplement). However, there were differences in the reporting of *P* values among the categories of articles examined. In 2014, 33.0% (95% CI, 32.5%-33.5%) of the 29 725 abstracts of articles from the 151 core clinical journals reported *P* values. The proportion was 35.7% (95% CI, 34.5%-37.0%) in meta-analyses (*n* = 5620), 38.9% (95% CI, 37.5%-40.3%) in clinical trials (*n* = 4624, excluding randomized clinical trials), 54.8% (95% CI, 54.0%-55.6%) in randomized clinical trials (*n* = 13 544), and 2.4% (95% CI, 2.3%-2.5%) in reviews (*n* = 71 529, excluding meta-analyses) (Figure 1). The increase in the proportion of abstracts with *P* values over time pertains to all categories, with more prominent increases for meta-analyses (almost tripling in the last 2 decades).

Reporting *P* Values in Full-Text Articles

The random sample of the PMC database (downloaded March 20, 2015) included 843 884 full-text articles, of which 750 133 (89%) had an abstract, including 385 393 abstracts that reported *P* values. From these abstracts, 3 438 299 *P* values were extracted.

Based on the PMC sample of 750 133 full-text articles that had an abstract, 382 037 (50.9%) reported at least 1 *P* value in the full text, and 384 117 (51.2%) reported at least 1 *P* value either in the full text or in the abstract.

In the overall PMC sample, 25.3% (*n* = 97 463) of articles that reported any *P* values in the full text also reported at least 1 *P* value in the abstract (eFigure 2A in the Supplement). This proportion was steady over time (eFigure 2B in the Supplement), and in 2014 was 40.3% in core clinical journals, 44.7% in meta-analyses, 46.2% in clinical trials, 53.7% in randomized clinical trials, and 9.6% in reviews.

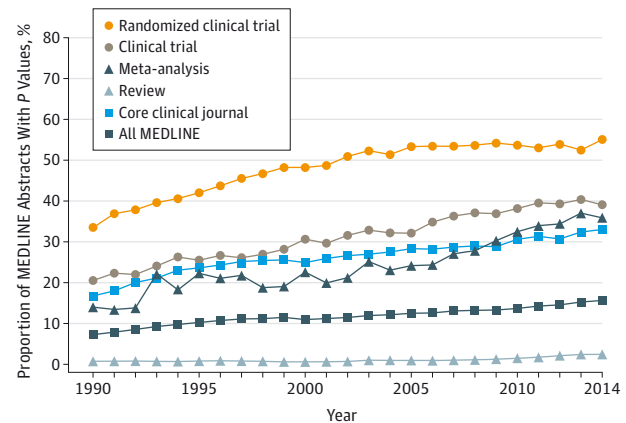
The proportions of PMC articles that reported at least 1 *P* value in either the full text or the abstract were 69.7% for core clinical journals (*n* = 6410), 82.7% for meta-analyses (*n* = 4688), 75.5% for clinical trials (*n* = 7731), 75.9% for randomized clinical trials (*n* = 14 646), and 22% for reviews (*n* = 47 191). When limited to the period 2011-2015, the proportions were 68.0% for core clinical journals, 83.7% for meta-analyses, 79.0% for clinical trials, 75.4% for randomized clinical trials, and 22.1% for reviews.

Distribution of *P* Values in Abstracts and Full Text

The distribution of reported *P* values, both in abstracts and in full text, showed strong clustering around some specific rounded *P* values, most commonly *P* values of .05 and of .001 or smaller, with less prominent clustering for values of .01. This distribution was similar for *P* values reported in full-text articles that had abstracts (Figure 2), in PMC abstracts (eFigure 3 in the Supplement), and in PMC full-text articles (eFigure 4 in the Supplement), although more “strongly” significant results (ie, *P* values of .001 or smaller) were reported more commonly in abstracts than in the PMC full-text articles. For example, in the PMC abstracts, *P* values of .05 were much less frequent (0.59-fold [*n* = 50 084]) than *P* values of .001 or smaller (*n* = 85 195), whereas in PMC full-text articles *P* values of .05 were slightly more frequent (1.10-fold [*n* = 935 627]) than *P* values of .001 or smaller (*n* = 837 761).

The distribution of reported *P* values varied across different categories of articles. The reporting of *P* values in meta-analyses,

Figure 1. Proportion of MEDLINE Abstracts Reporting at Least 1 *P* Value in the Period 1990-2015



Article Category	Abstracts, No.	Abstracts Reporting <i>P</i> Values, No. (range per year)
Randomized clinical trial	332 262	165 968 (2056-12 101)
Clinical trial	197 921	60 317 (736-3700)
Meta-analysis	52 639	16 018 (30-2960)
Review	1 396 965	17 125 (176-2245)
Core clinical trial	796 103	210 045 (2907-10 090)
All MEDLINE	12 821 790	1 608 736 (20 709-138 218)

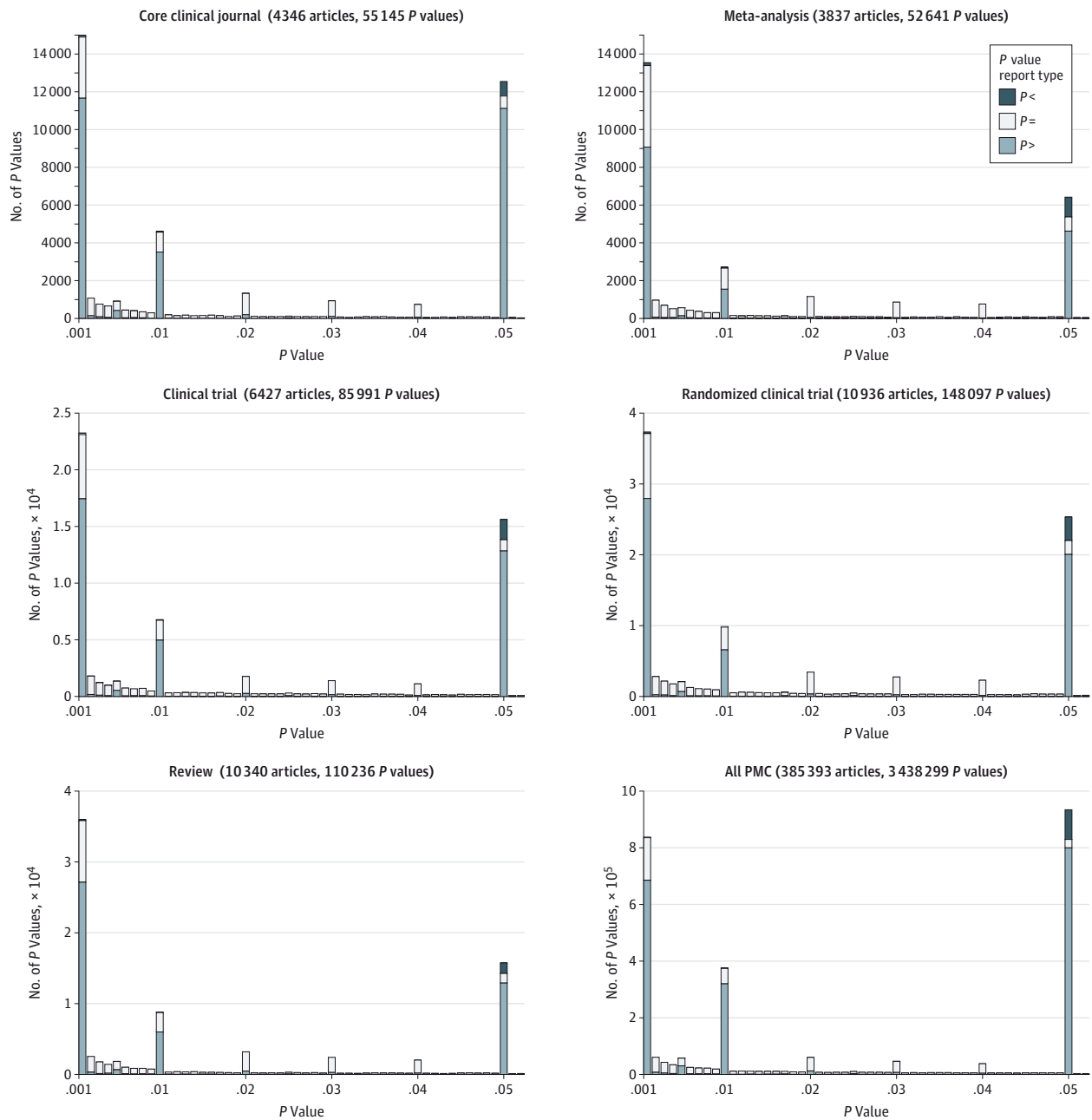
reviews, randomized clinical trials, clinical trials, and in articles in core clinical journals had showed more strongly statistically significant results (ie, *P* values of .001 or less) in abstracts compared with full-text PMC articles. For example, when limited to core clinical journals, *P* values of .001 or smaller were only slightly more frequent (1.2-fold [*n* = 15 001]) than values of .05 (*n* = 12 513) in the full-texts, but values of .001 or smaller were far more frequent (3.7-fold [*n* = 2782]) than values of .05 in the abstracts (*n* = 759). Similar enrichment in small *P* values was also seen for all study design types (Figure 2; eFigures 3 and 4 in the Supplement). The relative proportion of abstracts of PMC articles reporting highly statistically significant results (*P* values of .001 or less) was even greater, especially for reviews, meta-analyses, and articles in core journals.

Minimal and Maximal Reported *P* Values

In MEDLINE abstracts, the most statistically significant (minimal/lower/“best”) reported *P* value became more prominently significant over time, with a more rapid change in the period 1990-1995 and a slower change since then (eFigure 5 in the Supplement). At the end of 2014, the average $-\log_{10}$ best reported *P* value was 2.48 overall (corresponding to *P* = .003) and ranged between 2.3 and 2.7 (corresponding to *P* = .002-.005) for the different categories of articles examined, except for meta-analyses, for which the $-\log_{10}$ best reported *P* value exceeded 3.1; ie, the average meta-analysis reported at least 1 *P* value below .0008.

The least statistically significant (maximal/higher/“worst”) *P* value in MEDLINE abstracts became slightly less prominently significant over time (eFigure 6 in the Supplement). At the end of 2014 the average $-\log_{10}$ worst *P* value was 1.63 overall (corresponding to *P* = .02), 1.35 for randomized controlled trials (corresponding to *P* = .0437), and 1.52 to 1.63 for all other categories

Figure 2. Distribution of P Values in 385 393 PMC Full-Text Articles That Have Abstracts



Numerical values not shown (>.05) represent 17.41% of the total (598 611 P values). There are 50 bins shown, each with width .001.

(corresponding to $P = .02-.03$). Thus, even the “worst” reported P values still remained mostly within the range of nominally statistically significant results ($P < .05$). In the PMC full-text articles, at the end of the study period the average $-\log_{10}$ best P value reached approximately 2.57 overall, ie, $P = .0027$ (eFigure 7 in the Supplement).

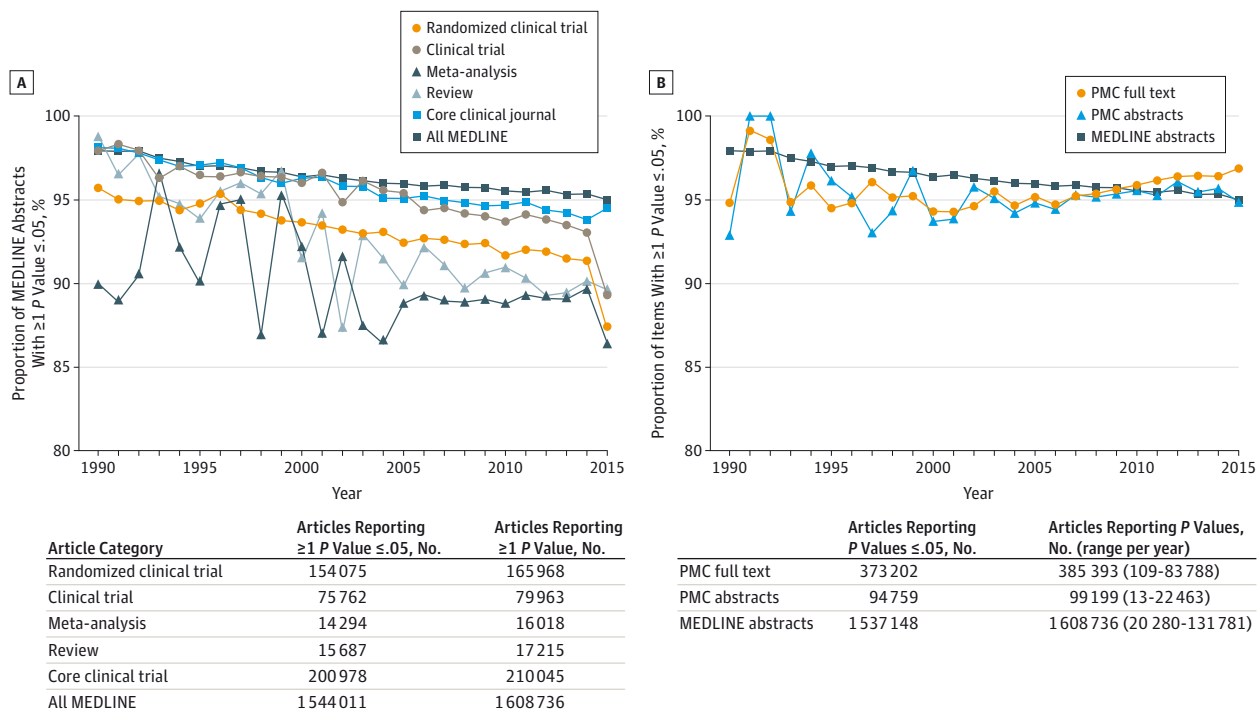
In addition, the proportion of P values reported in MEDLINE abstracts as inequalities (eg, “ $P <$ ” or “ $P \leq$ ”) decreased over time (a larger percentage of “ $P =$ ” values were reported, eFigure 8 in the Supplement). When analyses were limited to precise P values (“ $P =$ ”), at the end of the study period, across MEDLINE abstracts

the mean $-\log_{10}$ best reported P value was 2.2 (corresponding to $P = .006$) and the mean $-\log_{10}$ worst reported P value was 1.45 (corresponding to $P = .035$), whereas the mean $-\log_{10}$ best reported P value in PMC full-text articles was 2.42 (corresponding to $P = .004$) (eFigures 9-11 in the Supplement).

Frequency of Reporting of at Least 1 P Value of .05 or Less

Across the 1 608 736 MEDLINE abstracts with any P value reported, 96.0% reported at least 1 P value that was .05 or less, with a slight decrease over time from 97.9% in 1990 to 95.0% in 2014 (Figure 3A). Similarly high proportions of P values of .05 or less

Figure 3. Evolution of P Values Reported in the Period 1990-2015



A, Proportion of items that have at least 1 P value of .05 or less in MEDLINE abstracts, 1990-2015. B, Proportion of articles that have at least 1 P value of .05 or less in PubMed Central abstracts and full-text articles and MEDLINE abstracts, 1990-2015.

were reported across all specific categories of articles examined, including, in 2014, 93.8% in core clinical journals ($n = 9193$), 89.7% in meta-analyses ($n = 1802$), 93.1% in clinical trials ($n = 1714$), 91.4% in randomized clinical trials ($n = 6784$), and 90.1% in reviews ($n = 1564$). Similarly, among the 385 393 PMC articles with P values reported in the full text, 96.8% reported at least 1 P value that was .05 or less, and this proportion remained stable over time (Figure 3B).

In-depth Manual Assessment

Among a random sample of 1000 MEDLINE abstracts, 796 reported empirical data. Among these 796, none of the abstracts reported any Bayes factor (0% [95% CI, 0.0%-0.5%]). A total of 125 (15.7% [95% CI, 13.2%-18.4%]) abstracts reported at least 1 P value; of these, 118 (94.4% [95% CI, 88.8%-97.7%]) reported at least 1 statistically significant P value (Table 1). Of the 378 P values identified in these abstracts, 332 (87.8% [95% CI, 84.1%-91.0%]) were .05 or lower, including 83 (22.0% [95% CI, 17.9%-26.5%]) that were .001 or lower.

Only 18 (2.3% [95% CI, 1.3%-3.6%]) abstracts reported at least 1 CI. Of the 49 CIs reported in the abstracts, 37 (75.5% [95% CI, 61.1%-86.7%]) were statistically significant (entirely on the same side of the null). Almost always (47/49; 95.9% [95% CI, 86.0%-99.5%]), these CIs were clearly specified as 95% CIs.

From the 179 (22.5% [95% CI, 19.6%-25.5%]) abstracts with at least 1 effect size, with information allowing for the calculation of at least 1 effect size, or both, 111 abstracts (13.9% [95% CI, 11.6%-16.5%]) reported at least 1 effect size and 99 (12.4% [95% CI, 10.2%-14.9%]) included information allowing for the

calculation of at least 1 effect size. Among the 179 abstracts, the majority (94/179; 52.5% [95% CI, 44.9%-60.0%]) did not include any P values, and the vast majority (163/179; 91.1% [95% CI, 85.9%-94.8%]) did not include any CI.

A wide variety of metrics were used to report effect sizes (Table 2). Among 269 reported effect sizes, only 66 (24.5% [95% CI, 19.5%-30.1%]) had a corresponding P value and only 40 (14.9% [95% CI, 10.8%-19.7%]) had a corresponding CI. Certain effect metrics, in particular ratios (relative risk or reduction, odds ratio, hazard ratio), were likely to be accompanied by P values, CIs, or both, whereas P values were not commonly reported for other metrics, such as fold difference or change, percent difference or change, and correlation coefficients.

Overall, 181 of the 1000 abstracts (18.1% [95% CI, 15.8%-20.6%]) included at least 1 qualitative statement (eg, "however, the difference of both parameters was not significant"; "no significant difference was found comparing Group 2 with Group 3") about significance without reporting at least 1 corresponding P value (included only positive statements: 113 [62.4%; 95% CI, 54.9%-69.5%]; only negative statements: 42 [23.2%; 95% CI, 17.3%-30.0%]; both positive and negative statements: 26 [14.4%; 95% CI, 9.6%-20.3%]). Effect sizes accompanying these statements were uncommon ($n = 26$ [14.4%; 95% CI, 9.6%-20.3%]), and measures of uncertainty were rarely presented ($n = 3$ [1.7%; 95% CI, 0.3%-4.8%]). Few abstracts ($n = 16$ [8.8%; 95% CI, 5.1%-14.0%]) had at least 1 statement with "statistical," "statistically," or a similar phrase accompanying the significance statement (Table 1). Only 1 abstract mentioned clinical significance.

Table 1. Reporting Characteristics of 1000 Abstracts

Abstract Reporting Characteristics	No. (%) [95% CI]
P Values (n = 796 Articles)	
At least 1 P value	125 (15.7) [13.2-18.4]
At least 1 statistically significant P value ($\leq .05$)	118 (14.9) [12.5-17.6]
At least 1 statistically nonsignificant P value ($> .05$)	28 (3.5) [2.3-5.0]
Confidence Intervals (CIs) (n = 796 Articles)	
At least 1 CI	18 (2.3) [1.3-3.6]
At least 1 95% CI	17 (2.1) [1.2-3.4]
Effect Sizes (n = 796 Articles)	
At least 1 effect size	111 (13.9) [11.6-16.5]
At least 1 effect size and 1 P value for at least 1 of the effect sizes	37 (4.7) [3.3-6.4]
At least 1 effect size and 1 95% CI for at least 1 effect size	14 (1.8) [1.0-2.9]
Effect Sizes That Can Be Calculated (n = 796 Articles)	
Abstracts for which at least 1 effect size can be calculated	99 (12.4) [10.2-14.9]
Abstracts for which at least 1 effect size can be calculated and 1 P value is reported for at least 1 effect size that can be calculated	43 (5.4) [3.9-7.2]
Abstracts for which at least 1 effect size can be calculated and 1 CI is reported for at least 1 effect size that can be calculated	0 (0.0) [0.0-0.5]
Qualitative Statements About Significance (n = 1000 Articles)	
Including at least 1 statement about significance	181 (18.1) [15.8-20.6]
Including at least 1 statement about significance, with at least 1 effect size or for which 1 effect size can be calculated	26 (2.6) [1.7-3.8]
Including at least 1 statement about significance, with at least 1 effect size or for which 1 effect size can be calculated and at least 1 CI for at least 1 effect size	3 (0.3) [0.1-0.9]

Table 2. Types of Effect Sizes^a

Type of Effect Size	Total Occurrences of Effect Sizes, No. (n = 269)	No. (%)		
		With P Value	With CI	With P Value or CI
Percent difference or change ^b	86	14 (16.3)	4 (4.7)	18 (20.9)
Correlation coefficient	58	18 (31.0)	0	18 (31.0)
Fold difference or change	37	1 (2.7)	0	1 (2.7)
Absolute difference or change	34	12 (35.3)	8 (23.5)	18 (52.9)
Odds ratio	15	8 (53.3)	12 (80.0)	14 (93.3)
Hazard ratio	12	5 (41.7)	7 (58.3)	12 (100.0)
Relative risk or risk ratio	8	5 (62.5)	3 (37.5)	8 (100.0)
Beta coefficient	8	3 (37.5)	0	3 (37.5)
Relative risk reduction	6	0	6 (100.0)	6 (100.0)
Interclass correlation coefficient	4	0	0	0
Variance explained	1	0	0	0

^a Table does not include 215 occurrences for which information was presented that could allow calculation of an effect size from the presented numbers; P values accompanied 83 of the 215 (38.6%).

^b Includes absolute risk reduction for binary variables expressed as percentages.

Of the 100 full-text articles with empirical data that were randomly selected for manual review and data extraction, 1 could not be retrieved. Among the remaining 99 full-text articles, 55 (55.6% [95% CI, 45.2%-65.5%]) reported at least 1 P value, 12 (12.1% [95% CI, 6.4%-20.2%]) reported at least 1 CI, 46 (46.5% [95% CI, 36.4%-56.8%]) reported at least 1 effect size, 0 (0% [95% CI, 0%-3.7%]) reported Bayesian methods, and 1 (1.0% [95% CI, 0.0%-5.5%]) reported false-discovery rate methods. Only 4 articles (4% [95% CI, 0.1%-10.0%]) reported all effect sizes with corresponding CIs. Only 3 articles (3% [95% CI, 0.6%-8.6%]) included some sample size/power calculations in their Methods (eAppendix in the Supplement). Only 5 articles (5.1% [95% CI, 1.7%-11.4%])

specifically mentioned their primary outcome(s) of interest; of these, only 2 articles provided an effect size and P value for primary outcome(s) of interest (eAppendix in the Supplement).

Discussion

This evaluation of abstracts in the entire MEDLINE database from 1990 to 2015 and in a large number of recent full-text PMC articles showed an increasing prevalence of P values reported in the biomedical literature. Moreover, P values reported in abstracts were in general lower (showing greater statistical significance) than P values reported in the full text. The use of P values was even more common in core clinical journals and in influential articles such as randomized trials and meta-analyses. The selection of more statistically significant P values in the abstracts was prominent also in randomized trials and meta-analyses. In-depth manual analysis of a sample of 1000 abstracts and 100 full-text articles demonstrated that Bayesian methods and false-discovery rate methods were almost entirely absent, and use of CIs was seldom reported and provided mostly for risk metrics. Effect sizes were reported in a sizeable proportion of abstracts but almost always without information that would allow conveying their uncertainty. Furthermore, besides the substantial proportion of abstracts that report P values, a larger proportion of abstracts included qualitative statements about significance, mostly without any other quantitative information.

There is a long-standing debate about the use of P values. Many authors have recognized the limitations and problems of reliance on P values alone.¹⁻⁶ P values do not provide a direct estimate of how likely a result is true or of how likely the null hypothesis ("there is no effect") is true. Moreover, they do not convey whether a result is clinically or biologically significant. P values depend not only on the data but also on the statistical method used, the assumptions made, and the appropriateness of these assumptions.

Nevertheless, the use of *P* values is not necessarily a problem. The wide use of *P* values in the literature may largely signify that more articles are using some sort of statistical analysis and that reporting results of these analyses using *P* values is the norm. The higher prevalence of *P* values in abstracts over time also may be associated in part with wider use of structured abstracts.¹³ Problems arise when *P* values are frequently misinterpreted about what they mean, when they are selectively reported, and when they are not accompanied by estimates of effect size and uncertainty.^{14,15} Effect sizes and measures of uncertainty are key to understanding the results and making decisions. Other useful statistical approaches that may be more directly interpretable include Bayesian methods^{16,17} that focus on calculating posterior probabilities based on prior beliefs and observed data and false-discovery approaches that aim to estimate the chance that a “discovery” is false.

In this study, there was strong clustering at specific rounded *P* values, especially at the value .05, both in the abstracts and in the full text. This is probably a consequence of the widespread use of this value in hypothesis testing. The common approach of reporting *P* values simply as $P < .05$ (or any other threshold) is inferior to providing precise values. These data indicate that over time more *P* values were reported more precisely (“ $P =$ ”), but a large proportion were still presented with inequality thresholds.

In this study, the full texts of PMC articles had more *P* values of .05 as compared with values of .001 or less that are often labeled “highly statistically significant.” Conversely, abstracts have a selection bias favoring results with very small *P* values.^{18,19} Thus, in the MEDLINE abstracts examined in this study, *P* values of .001 or less were far more commonly reported than values of .05. However, *P* values reported in the text also may be lower on average than the larger sample of those that are reported and typically were not automatically extracted by our screening tools. Furthermore, the *P* values in the full text or tables may be a select set of all *P* values obtained in all the analyses (published and unpublished) performed in a study.^{20,21} If so, the *P* values reported in abstracts may represent an even more highly selected sample. This suggests that abstracts may appear to provide a somewhat distorted picture of the evidence. However, many readers focus primarily on the abstracts of the articles. The low *P* values (ie, .001 or less) were reported in abstracts for meta-analyses and reviews and for abstracts from articles in core medical journals. These segments of the literature are influential in clinical medicine and practice,²² and core medical journals are most influential in both clinical practice and research.

We also documented reporting of more statistically significant results (ie, lower *P* values) on average over time in abstracts, with a decrease in the average “best” (minimal) *P* value reported in an abstract. This pattern may be genuine to some extent; ie, results may become more statistically significant on average in more recent research. This may reflect an increasing number of measured variables and respective analyses (with more opportunities to obtain lower *P* values) or an increase in the sample size of studies performed in some fields. However, in our detailed review of 99 articles, only 3 included sample size calculations.

Alternatively, this phenomenon of lower *P* values reported over time may reflect a combination of increasing pressure to deliver (even more) significant results in the competitive, publish-

or-perish scientific environment as well as the recent conduct of more studies that test a very large number of hypotheses and thus can reach lower *P* values simply by chance. Some fields, such as genetics, have adopted far more stringent rules of claiming statistical significance; eg, $P < 10^{-8}$ is a standard threshold for genome-related studies.^{23,24} In the present study, the pervasive presence of *P* values less than .05 in almost all abstracts that reported *P* values suggests that this threshold has lost its discriminating ability for separating false from true hypotheses; more stringent *P* value thresholds are probably warranted across scientific fields.^{25,26}

There was also a small, but slowly increasing, number of reports that included only statistically nonsignificant results in the abstract. This is a welcome change and may suggest an increasing niche for the publication of “negative” studies. However, it is unclear whether these nonsignificant results are also interpreted as such by their authors. For example, there is evidence for spin effects, whereby investigators interpret nonsignificant findings as if they were significant.²⁷ Moreover, the majority of abstracts did not report a single statistically nonsignificant *P* value. Beyond biomedicine, other data suggest that “negative” results are disappearing from many scientific fields and from research conducted in many countries,²⁸ and there are spuriously too many statistically significant results,²⁹ while all results should be communicated in an unbiased fashion regardless of their statistical significance.³⁰

The biomedical literature contains more publications than any other scientific discipline. Analyses conducted in diverse biomedical and scientific fields may give comparative insights on the dynamics of using *P* values.^{28,29,31-34} An evaluation³¹ of abstracts with the Scopus search engine as a proxy (the Scopus database also includes other sciences) also showed a substantial increase over time in the number of abstracts reporting “ $P > .05$ ” and a 10-fold increase between 1990 and 2013 in the number of *P* values in the range of .041 to .049. Although more nonsignificant *P* values are reported in absolute numbers, their proportion among all *P* values reported may not necessarily increase and may even decrease in many disciplines.²⁸ Moreover, several theoretical and empirical investigations^{29,33,34} have suggested that concentration of *P* values in the .041 to .049 range is a sign of potential *P* value hacking, whereby investigators manipulate their analyses until they cross the desirable threshold of nominal statistical significance.

This study has some limitations. First, our mining of full-text articles excluded tables and other display items. Thus, if anything, the automated approach underestimates the prevalence of *P* values in the biomedical literature. Based on our manual in-depth evaluation, this underestimation is small. Second, no automated data extraction can verify the accuracy of the extracted information. However, our manual assessment also suggested that inaccuracies were not frequent. Third, some of the abstracts and articles (eg, case reports and reviews) may require no statistical tests.

Overall, we do not recommend that *P* values should be abandoned. Alternative statistics such as Bayes factors may also be warranted and helpful to consider in many cases, but even if there were a change from *P* values to Bayes factors or false-discovery rates, this would not necessarily reduce the problem of selective reporting and lack of reporting of important information on

effect sizes, such as **absolute and relative risk measures** and mean differences. The transparency, accuracy, and information content of **the biomedical literature would benefit** from increased reporting of **both effect sizes** and **measures of uncertainty** or at least both effect size and *P* value in abstracts. Qualitative statements about significance in the absence of quantitative information are difficult to interpret and may be **misleading because statistical, biological, and clinical significance are different concepts** and subjective interpretation of significance may be incorrect. Therefore, such isolated qualitative statements should be avoided. By default, isolated reporting of *P* values also should be avoided, unless a cogent argument can be made that effect size is not relevant (eg, in some genomic studies). In addition, journals should en-

courage investigators to report in their abstracts the quantitative findings of their main analyses and not necessarily those that were nominally statistically significant.

Conclusions

In this analysis of *P* values reported in MEDLINE abstracts and in PMC articles from 1990-2015, more MEDLINE abstracts and articles reported *P* values over time, and almost all abstracts and articles with *P* values reported statistically significant results. Rather than reporting isolated *P* values, articles should include effect sizes and uncertainty metrics.

ARTICLE INFORMATION

Correction: This article was corrected online on May 12, 2016, to add links to the full dataset.

Author Contributions: Drs Chavalarias and Ioannidis had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Chavalarias, Wallach, Ioannidis.

Acquisition, analysis, or interpretation of data: All authors.

Drafting of the manuscript: Chavalarias, Wallach, Ioannidis.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Chavalarias, Wallach, Ioannidis.

Administrative, technical, or material support: Chavalarias.

Study supervision: Ioannidis.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Funding/Support: The Meta-Research Innovation Center at Stanford (METRICS) is supported by a grant by the Laura and John Arnold Foundation. The work of Dr Chavalarias is supported by the Complex Systems Institutes of Paris Ile-de-France (ISC-PIF), the Région Ile-de-France and a grant from the CNRS Mastodons program. Mr Li was supported by a Canadian Institute for Health Research Doctoral Scholarship with a Michael Smith Foreign Study Supplement. The work of Dr Ioannidis is supported by an unrestricted gift by Sue and Bob O'Donnell to Stanford Prevention Research Center.

Role of the Funders/Sponsors: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Full Dataset: The full records of *P* values extracted from PubMed and PMC, with their metadata (doi:10.7910/DVN/6FMFT3), are available at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6FMFT3>.

REFERENCES

- Goodman SN. Toward evidence-based medical statistics, 1: the *P* value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
- Goodman S. A dirty dozen: twelve *p*-value misconceptions. *Semin Hematol*. 2008;45(3):135-140.
- Gelman A. *P* values and statistical practice. *Epidemiology*. 2013;24(1):69-72.
- Cohen J. The earth is round ($p < .05$). *Am Psychol*. 1994;49:997-1003.
- Gigerenzer G, Krauss S, Vitouch O. The null ritual: what you always wanted to know about significance testing but were afraid to ask. In: Kaplan D, ed. *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage; 2004:391-408.
- Fanelli D. "Positive" results increase down the hierarchy of the sciences. *PLoS One*. 2010;5(4):e10068.
- Berlin JA, Begg CB, Louis TA. An assessment of publication bias using a sample of published clinical trials. *J Am Stat Assoc*. 1989;84:381-392.
- Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867-872.
- Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. *JAMA*. 1992;267(3):374-378.
- Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
- Young SS, Karr A. Deming, data, and observational studies. *Significance*. 2011;8:116-120.
- Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*. 1995;90:419-435.
- Hartley J. Current findings from research on structured abstracts. *J Med Libr Assoc*. 2004;92(3):368-371.
- Gardner MJ, Altman DG. Confidence intervals rather than *P* values. *Br Med J (Clin Res Ed)*. 1986;292(6522):746-750.
- Gardner MJ, Altman DG. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines*. London, United Kingdom: BMJ Publishing;1989.
- Lehmann HP, Goodman SN. Bayesian communication. *J Am Med Inform Assoc*. 2000;7(3):254-266.
- Goodman SN. Toward evidence-based medical statistics, 2: the Bayes factor. *Ann Intern Med*. 1999;130(12):1005-1013.
- Göttsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. *BMJ*. 2006;333(7561):231-234.
- Ioannidis JP. Discussion: why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics*. 2014;15(1):28-36.
- Tsilidis KK, Panagiotou OA, Sena ES, et al. Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biol*. 2013;11(7):e1001609.
- Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4(3):245-253.
- Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. *JAMA*. 2005;293(19):2362-2366.
- Xu C, Tachmazidou I, Walter K, et al. Estimating genome-wide significance for whole-genome sequencing studies. *Genet Epidemiol*. 2014;38(4):281-290.
- Hoggart CJ, Clark TG, De Iorio M, et al. Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol*. 2008;32(2):179-185.
- Johnson VE. Revised standards for statistical evidence. *Proc Natl Acad Sci U S A*. 2013;110(48):19313-19317.
- Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *BMJ*. 2001;322(7280):226-231.
- Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. 2010;303(20):2058-2064.
- Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics*. 2011;90:891-904.
- Simonsohn U, Nelson LD, Simmons JP. *P*-curve: a key to the file-drawer. *J Exp Psychol Gen*. 2014;143(2):534-547.
- van Assen MA, van Aert RC, Nuijten MB, Wicherts JM. Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS One*. 2014;9(1):e84896.
- de Winter JC, Dodou D. A surge of *p*-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*. 2015;3:e733.
- Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*. 2014;15(1):1-12.
- Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of *p*-hacking in science. *PLoS Biol*. 2015;13(3):e1002106.
- Simonsohn U, Nelson LD, Simmons JP. *p*-Curve and effect size: correcting for publication bias using only significant results. *Perspect Psychol Sci*. 2014;9(6):666-681.