**Taylor & Francis**
Taylor & Francis Group

# What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University and Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA

## ABSTRACT

*P* values linked to null hypothesis significance testing (NHST) is the most widely (mis)used method of statistical inference. Empirical data suggest that across the biomedical literature (1990–2015), when abstracts use *P* values 96% of them have *P* values of 0.05 or less. The same percentage (96%) applies for full-text articles. Among 100 articles in PubMed, 55 report *P* values, while only 4 present confidence intervals for all the reported effect sizes, none use Bayesian methods and none use false-discovery rate. Over 25 years (1990–2015), use of *P* values in abstracts has doubled for all PubMed, and tripled for meta-analyses, while for some types of designs such as randomized trials the majority of abstracts report *P* values. There is major selective reporting for *P* values. Abstracts tend to highlight most favorable *P* values and inferences use even further spin to reach exaggerated, unreliable conclusions. The availability of large-scale data on *P* values from many papers has allowed the development and applications of methods that try to detect and model selection biases, for example, p-hacking, that cause patterns of excess significance. Inferences need to be cautious as they depend on the assumptions made by these models and can be affected by the presence of other biases (e.g., confounding in observational studies). While much of the unreliability of past and present research is driven by small, underpowered studies, NHST with *P* values may be also particularly problematic in the era of overpowered big data. NHST and *P* values are optimal only in a minority of current research. Using a more stringent threshold, as in the recently proposed shift from $P < 0.05$ to $P < 0.005$, is a temporizing measure to contain the flood and death-by-significance. NHST and *P* values may be replaced in many fields by other, more fit-for-purpose, inferential methods. However, curtailing selection biases requires additional measures, beyond changes in inferential methods, and in particular reproducible research practices.

## 1. Introduction

Null hypothesis significance testing (NHST) and *P* value thresholds such as 0.05 have long been a mainstay of empirical work in the sciences. Increasingly, however, statisticians and other scientists concerned with learning from data have come to recognize major shortcomings in the way these methods are used. This paper, based on an invited plenary address to a recent ASA-sponsored workshop on statistical inference, summarizes recent empirical work on the use and misuse of *P* values and places in context what we have learnt towards solving this conundrum.

In what follows, Section 2 summarizes empirical results from a database of 13 million abstracts and 844 thousand full articles taken from PubMed Central between 1990 and 2015. Section 3 discusses how bias emerges from a multilayered selection process that leads to specific reported *P* values. Section 4 describes and discusses a variety of proposed remedies intended to address the problem of selection bias. These include: (4.1) alternative approaches to inference (effect sizes and confidence intervals, Bayesian methods, changing the *P* value threshold); (4.2) attempts to model the selection process (the *P* value curve and meta-analysis of publication selection); (4.3) examples of alternatives based on context and goals; and, finally (4.4) how reproducible research practices might offer the best solution. A concluding section offers some final thoughts.

## 2. Empirical Results: NHST is Widespread and Reliance on *P* Values Increases Over Time

There are over 100 million published articles in the scientific literature (Khabsa and Giles 2014), and a substantial proportion of them use data. Among those that use data an increasing proportion use also some tools of statistical inference beyond simple description. Different scientific fields use different statistical tools by tradition, but their traditions are not necessarily justified or fit-for-purpose. Convenience, inertia, poor quantitative and statistical training of scientists, and lack of initiative from journals and funding agencies may perpetuate the use and misuse of those tools (Szucs and Ioannidis 2017a).

In particular, the use and misuse of *P* values is, arguably, the most widely perpetrated misdeed of statistical inference across all of science (Chavalarias et al. 2016). NHST coupled with the use of *P* value thresholds dominates most fields in the biomedical and life sciences, social sciences, and physical sciences. Most fields use *P* value thresholds of 0.05 to differentiate in black-and-white fashion between "significant" and "nonsignificant" results. Exceptions do occur, for example, the use of *P* value thresholds of $3 \times 10^{-7}$ (5 sigma) in high-energy physics of $5 \times 10^{-8}$ (genome-wide significance) in genome epidemiology, but they are relatively uncommon when the scientific literature is seen in its total volume.

The remainder of this section first describes a database of published biomedical literature, then summarizes results obtained from text mining of this database (Chavalarias et al. 2016).

• *The database from PubMed and PubMed Central*

The results presented in this section are based on a survey of the entire biomedical literature published during the quarter-century from 1990 to 2015. Text mining was used to assess the presence of $P$ values in the abstracts of 16.2 million items (13.0 million of which had an abstract). Similar text mining was performed in PubMed Central (PMC) for 844,000 full-text articles. For details, see ref. 3. Across this large corpus of biomedical literature, when abstracts use $P$ values 96% of them have $P$ values of 0.05 or less. The same percentage (96%) applies for full-text articles. This is too good to be true. Dissecting the use and misuse or $P$ values may explain why.

• *The use of null hypothesis significance testing and P value thresholds is widespread*

The proportion of PMC papers with $P$ values in their abstract or text is 51.1% for all papers. However, this figure is an underestimate, because the text mining could not capture most $P$ values embedded in tables and figures. Manual evaluation of the full articles (including tables and figures) in a sample of 100 randomly selected articles from PubMed, found 55 that report $P$ values. The use of other tools of statistical inference is rare or nonexistent: of the 100 papers, four report confidence intervals for all the reported effect sizes, none uses Bayesian methods, and none uses methods based on false discovery rates.

• *The rate is higher in clinical journals and in meta-analyses*

Although the use of $P$ values is widespread, there are categories of studies for which the percentage reporting p-values is substantially higher than the overall rate. Among these categories are:

| | |
|---|---|
| Overall (all papers) | 51.1% |
| Articles published in core clinical journals | 78.4% |
| Meta-analyses | 82.8% |
| Randomized controlled trials | 76.0% |
| Other clinical trials (excluding randomized controlled trials) | 75.7% |

To the extent that $P$ values are misused, or are used in place of other more suitable methods, these high percentages are particularly concerning.

• *Reliance on P values is increasing over time*

The same survey (Chavalarias et al. 2016) suggests that the use of $P$ values has increased over the 25 years covered by the sample. For all Pubmed abstracts, the percentage of abstracts reporting $P$ values doubled from 8% in 1990 to 17% in 2015; the rate tripled for meta-analyses. For some types of designs such as randomized controlled trials, about 60% of articles currently have some $P$ value(s) in their abstracts. Of note, the proportion of articles that have $P$ values in the full-text is much larger (as shown above) than the proportion of those that have $P$ values in the abstract. Abstracts are, of course, highly prominent as they represent the façade of articles in terms of what they communicate.

• *The "typical size" of reported P values is 0.01, more or less*

Most of the reported $P$ values are modest. An exception is the tiny fraction (0.4%) of those presented with exponents of 10 (e.g., $2 \times 10{-8}$) or "EXP" of "E" notation, for which the mean –log($P$ value) is around 9, the other $P$ values have an average –log($P$ value) of 2, corresponding to $P = 0.01$.

Why should the widespread and growing reliance on $P$ values be such a concern? A main reason is selection bias. As described in the next section, selection operates at many levels, and the resulting bias substantially inflates the rate of false alarms in the published literature.

## 3. Selection Effects: The Typical Direction is Toward Claiming Greater Significance

"Selection" refers here to the collection of choices that lead from the planning of a study to the reporting of $P$ values. The premise of this section is that such selection occurs in many ways, at many steps in the analysis of a data set or study. At each step, there are choices to be made, and with each choice, there is an opportunity to shape the presentation of results. Section 3.1 offers simple empirical evidence of selection bias, namely, a tendency to choose smaller (more significant) $P$ values for inclusion in a paper's abstract. Expanding on this theme, Section 3.2 describes four expanding sets of choices, four layers of selection, as a frame for thinking about sources of bias in the analysis of data. These sets of choices are rarely reported in full transparency, and so remain hidden from the reader. However, the placement of $P$ values within the sections of an article (3.3) provides a way to track some of the selection bias. Cherry-picking is more pronounced in the most competitive journals (3.4). Big data sets, which offer greater scope for pattern searching, are correspondingly at greater risk for false positives (3.5).

### 3.1. P Values in the Abstracts are More "Significant" than P Values in the Full Text

Abstracts offer authors the best opportunity to say they have something important to present. If there is a selection bias at work in the choice of which $P$ values to highlight, we would expect to find that bias to show up in a comparison of $P$ values in the abstract with $P$ values in the full text. Specifically, we would expect an author to select for the abstract some of the more impressive of the $P$ values reported in the full text. As a measure of the selection effect, we use the ratio of papers/abstracts reporting $P$ values at 0.05 to those reporting $P$ values ≤ 0.001. For the papers in the full text sample (Chavalarias et al. 2016), the number of $P$ values at 0.05 exceeds by 11% the number of $P$ values at 0.001 or less. However, the opposite in seen in the abstracts of these papers, where $P$ values of 0.05 are 41% fewer than $P$ values of 0.001 or less. Clearly, there is conscious or subconscious selection of more impressive $P$ values in the abstracts.

The selection gradient is more steep in the Core Clinical Journals category where in the abstracts $P$ values of 0.05 are 73% fewer than the $P$ values of = <0.001, while in the full text they are only 16% fewer.

The comparisons here are based on observable data, but they are merely the visible manifestation of a multilayered selection process.

### 3.2. Layers of Selection for P Values: Which P Values Get Reported

One can think of layers of selection that are applied in the presentation and highlighting of results through *P* values.

- The universe of all *P* values obtained in all the analyses conducted during a scientific process. Unless everything is rigorously prespecified (an uncommon scenario) there can be many trail-and-error efforts at different analyses, a "garden of forking paths" as Andrew Gelman has characterized the process (Gelman 2014). With few exceptions, these iterations and forking paths are not yet documented anywhere. The next layer:
- All the *P* values that other analysts that other authors may obtain, if the original authors can make the data and script/code for their analysis available. If only the data are available without the guidance of a specific analysis plan, one can explore still more options:
- All possible analyses that might be run, for example, using different modeling choices, different adjustments for multivariable models, or different definitions for variables of interest. This can give a sense of the magnitude of the "vibration of effects," that is, by how much results can vary depending on the endorsed exact analytical choices (Patel et al. 15). This layer of variability is of course conditional on the data made available, and does not take into account whatever tailoring the authors may have done to arrive at the version of the data made available to others. Thus, there is a final layer:
- All possible results from pre-processing of the data, for example, trying multiple covariates but only making available in public those included in the "nicest" model (the one presented in the paper). In the absence of full preregistration (Chambers 2013), there is no obstacle to such an approach.

Empirical evaluations (Patel et al. 2015) of the vibration of effects (obtained with different analyses) has shown that if there is sufficient data and degrees of freedom for choices of models almost any result can be obtained. This results in the "Janus phenomenon" where totally opposite results are possible to obtain routinely provided there are sufficient degrees of freedom (Patel et al. 2015).

Most of the time data and script/code are not available, so these selection dilemmas are hidden from an outsider examining a report of a study. However, what can still be visible is the extent of selection within different sections of a published paper.

### 3.3. Selection Within Sections of a Paper

Section 3.1 compared the set of *P* values reported in the full text of an article with the set of *P* values reported in the abstract. The main finding was that *P* values chosen for the abstract tended to show greater significance than those reported in the text, and that the gradient was more pronounced in some types of journals and types of designs. It is useful to extend this approach by defining a hierarchy of prominence for the location of reported *P* values within an article.

- (a) Tables and figures tend to offer the most comprehensive recording of results, although even these may be cherry-picked from among several analyses of the data. Unfortunately, as noted before, text mining is not consistently able to recognize *P* values reported in tables and figures.
- (b) *P* values chosen for discussion in the text constitute a subset of all *P* values, and are typically chosen for discussion either because of interest, but occasionally because of some anomaly.
- (c) *P* values chosen for the abstract are even more likely to be chosen for impressive significance (Chavalarias et al. 2016) and are sometimes accompanied by implausible effect sizes (Gøtzsche 2006).
- (d) Finally, at the top of the hierarchy, are the *P* values taken most seriously by the authors in reaching their conclusions. Conclusions have been documented empirically to depend very often on "spin" (Boutron et al. 2014). With spin, results that fail to register formally as statistically significant can still be taken as "significant."

The typical direction of bias is towards claiming more significance as one moves through these selection steps. Of course, there can be exceptions to this rule, as in some cases nonsignificant *P* values are more attractive, for example, in noninferiority studies, but these tend to be the minority. Also, the exact use of and selection bias on *P* values at these different steps may depend on the type of discipline and the journal where research is published.

### 3.4. Cherry Picking in the More Competitive Basic Science Journals

For most journals, the tabulated results are likely to be more complete and less selective than the one or few *P* values highlighted in the abstract. However, for extremely competitive basic science journals such as Nature, Science, and PNAS, we have observed (Cristea and Ioannidis 2018) that when authors use *P* values in a Figure or Table (something they do increasingly over time) these *P* values are almost uniformly statistically significant. This uniformity of significant *P* values suggests that cherry-picking has already happened at the step where results are tabulated. It is reasonable to infer that "artificial scarcity"—the availability of very limited print space in prestigious journals—creates pressure to report impressive results (Young et al. 2008).

### 3.5. P Values, Big Data, and False Positives

An emerging compounding problem is the increasing availability of massive databases that can be analyzed in many scientific fields. While the typical challenge for most scientific work to date has been the conduct of underpowered studies (Szucs and Ioannidis 2017b), "big data" is bringing the reverse challenge of overpowered studies. Massive data sets expand the number of analyses that can be performed, and the multiplicity of possible analyses combines with lenient *P* value thresholds like 0.05 to generate vast potential for false positives. As just one extreme example, an analysis of the entire Swedish population might conclude—if results are taken at face value using lenient *P* value thresholds—that three quarters of medication classes are associated with cancer risk: obviously an impossible result (Patel et al. 2016).

**Table 1.** Is NHST a good choice for various research applications?

| Research application | Is NHST a good choice? |
| --- | --- |
| Developing a prognostic score for CVD? | No, selection of variables should not use NHST or use very lenient Type 1 error |
| Assessing a diagnostic test for depression? | No, absolute magnitude of improvement in diagnostic performance matters more |
| Evaluating medical therapies in randomized trials? | Mostly no, the "2 trials with $P < 0.05$" rule has modest discriminating performance |
| Mining electronic health records? | No, specificity of $P < 0.05$ in searching for genuine effects is very low |
| Mining big data from metabolomics? | No, except for screening, and only with multiplicity-corrected $P$ values thresholds |
| Assessing whether to exclude women athletes | No, magnitude of the competitive advantage |
| with high testosterone should from the Olympics? | Is what matters |

Unfortunately, it is easier to document the problem than to offer a simple, effective solution. We next consider some proposed remedies.

## 4.  Some Proposed Remedies

This section summarizes four sets of proposed remedies: alternative approaches to inference (4.1); examples of fit-to-purpose measures (4.2); attempts to model the selection process (4.3); and standards for reproducible research (4.4).

### 4.1.  Alternatives Approaches to Inference and Complements to P Values

We mention here well-known alternatives to null hypothesis testing via $P$ values at the 5% level: effect sizes, confidence intervals, methods based on false discovery rates, Bayesian methods, and a change to far more stringent thresholds for $P$ values.

Within the frequentist framework, some have proposed more extensive and routine use of effect sizes and confidence intervals as alternatives or complements to null hypothesis testing using $P$ values. Surely, such a change could help many papers become more understandable and less often misleading for both experts and users, especially for decision-making in medical applications. Across the biomedical literature the proportion of abstracts (11%) that report at least one effect size is a bit less but roughly the same as the proportion of abstracts (12.5%) that report at least one $P$ value (Chavalarias et al. 2016). For large-scale inference, methods based on the false discovery rate may be more appropriate in many if not most papers that currently use $P$ values, and, of course, Bayesian methods offer yet another approach. However, it is not clear that greater use of these alternative approaches would substantially diminish bias from selective reporting of the sort described in Section 3, because similar biases can be present regardless of the approach to inference used.

Meanwhile the widespread and expanding use of $P$ values suggests the urgency of our need for change. The recent proposal (Benjamin et al. 2017) to lower the traditional threshold for declaring significance from 0.05 to 0.005 should be seen mostly as a temporizing measure, a dam to contain the flood. Many caveats exists for such an approach, most of them raised in the original paper (Benjamin et al. 2017); their discussion is beyond the scope of the current paper. If applied across the biomedical literature of 1990–2015 surveyed in (Chavalarias et al. 2016), the proposed threshold of 0.005 will change the characterization of about one third of the $P$ values that are reported in the abstract and considered statistically significant. To the extent that the large majority of these $P$ values reflect spuriously significant associations due to selection bias, a change to the more stringent

threshold is likely to do more good than harm. The benefit may apply both to the (generally more appropriate) interpretation of past literature and the generation and reporting of new studies (Ioannidis 2018). All the same, changing the threshold cannot directly address the threat of selection bias.

So far, we have considered only broad-brush changes: greater use of effect sizes and confidence intervals, methods based on false discovery rates, Bayesian methods, and more stringent thresholds for declaring a result significant. For a great many applied problems there is a fit-to-purpose measure that is more suitable than the observed significance level for NHST.

### 4.2.  Specific Examples Based on Context and Goals

In most fields and with most types of study designs, NHST should not be the default choice for analysis. Table 1 lists a (nonrandom) sample of some common questions that arise in biomedical research.

None of these applications seems to be a good fit to for NHST:
- A prognostic score should be developed either without using statistical significance for choosing variables to include, or else using a very lenient Type I error rate such as alpha = 0.2 or even higher rather than 0.05.
- For estimating diagnostic performance metrics, $P$ values from testing against the null are not meaningful. The magnitude of the improvement in sensitivity and specificity is what matters, not whether the null hypothesis of no improvement can be rejected.
- Randomized trials have used $P$ values routinely, but simulations suggest (van Ravenzwaaij and Ioannidis 2017) that they are suboptimal and the rule of "Two trials with $P < 0.05$" for licensing is problematic.
- In the big data environment of electronic health records of omics, $P < 0.05$ makes no sense: It has negligible specificity and can lead to myriad false positive results.
- Finally, a recent consultation concerned the question, "Should women athletes with high testosterone be excluded from the Olympics?" The proponents of this exclusion use results from a paper (Bermon and Garnier 2018) that shows a barely significant difference between women with high and low testosterone (details omitted). However, the magnitude of the difference is tiny. Taking 99% confidence intervals into account, the possibility of a 10% advantage (the disqualifying limit) can be clearly excluded.
- Neither the broad-brush changes of 4.1 nor the more narrowly tailored statistical inference tools address directly the threat of selection bias. We turn next to proposals for modeling the selection process.

### 4.3. Attempts to Model the Selection Process

The large-scale availability of *P* values that can be readily extracted has led to interesting models that try to differentiate (a) distributions of *P* values in a body of literature commensurate with bias (e.g., p-hacking for passing traditional thresholds of statistical significance (Szucs 2016) from (b) those distributions that are most compatible with genuine discoveries of non-null associations and effects in the absence of such bias. The assumptions of each of these modeling approaches need to be carefully considered. An increasingly popular approach are P-curves, curves that plot the distribution of *P* values in a set of studies. It is speculated that a P-curve analysis (Simonsohn et a. 2014) may differentiate between bias and genuine discoveries. Such P-curves (Simonsohn et al. 2014) may indeed work quite well for randomized experimental studies with no other sources of bias. However, P-curves are sensitive to even tiny bias, corresponding to distortions of the effect sizes by 0.01 standard deviations in a setting of observational studies with confounding. Such a bias, through tiny, can generate a spurious P-curve that resembles genuine discoveries when in fact it is a mere artifact of confounding and omitted variable bias (Bruns and Ioannidis 2016).

Other approaches that can benefit from large-scale availability of *P* value data aim to model the publication selection process over time. When data from large collections of studies or from hundreds or thousands of meta-analyses are available one can assess the average pattern of selection. Consider, for example, the potential strength of publication bias for initial studies, early replication, and later replications (Pfeiffer et al. 2011). The selection forces may depend on the circumstances and the availability of prior evidence on the same question, as, for example, in the "Proteus phenomenon" (Ioannidis and Trikalinos 2005), where once a highly significant result is prominently published, there is a window of opportunity in the next year or two to publish a result that is totally opposite to the original. Furthermore, one can model average biases in sets of multiple studies, but, unfortunately, it is not possible to apply the averages to correct the results of any one particular study.

Meta-analyses can fix only a part of the problem of selective reporting. Sometimes different selection effects will have opposite directions and may cancel out, but more frequently, they may become more prominent. In this sense, meta-analyses may be useful in that a large body of literature can show the bias in sharper relief (Fanelli, Costas, and Ioannidis, 2017).

None of the proposed changes considered so far in this section can offer a head-on challenge to the threat of selection bias. In long term, the only direct protection must come from standards for reproducible research.

### 4.4. Reproducible Research is Key to Addressing Selection Effects Head-On

Given the shortcomings of proposed remedies and the vulnerability to selection bias present in all approaches to inference, that bias cannot be prevented even by requiring authors to make available both their data and the script or code used for the analysis. Unless this script and code were preregistered (Chambers 2013) it is not possible to tell whether the analysis plan was pre-specified or that it represents the final step of an extreme data exploration that remains unshared. Selection biases may be manageable mostly with improvements in reproducible research practices, such as better transparency, pre-registration, availability of all raw data and software code, greater collaboration and openness among scientists, and the adoption of rewards and incentives that can facilitate such behavior (Munafò et al. 2017).

## 5. Concluding Thoughts

In conclusion, the use of *P* values has become an epidemic affecting the majority of scientific disciplines. Decisive action is needed both from the statistical and wider scientific community (Wasserstein and Lazar 2016). Strong selection biases can make almost everything (seem) statistically significant and it is very likely that these biases do operate in many, probably most scientific fields that use *P* values, especially with lenient $P < 0.05$ thresholds for claiming success. Implausibly, 96% of the biomedical literature that uses *P* values in the abstract or in the full text claims statistically significant results (Chavalarias et al. 2016). Empirical data combined with plausible argument show that selection effects occur at multiple steps in the process of analyzing data and presenting the results, and that these strongly bias the selection of *P* values in the direction of greater significance. It has even been argued (Fanelli 2010) that fields with the highest proportion of significant claims may be least reliable, and that this ecological relationship can serve as the basis for a hierarchy of scientific fields.

NHST and *P* values are inherently most suitable/optimal for only a minority of current research. Using a more stringent threshold is a temporizing measure to avoid death-by-significance. NHST and *P* values may be replaced in many fields by other inferential methods that will be more fit for reading the results, understanding what they mean, and (when needed) acting on them. However, curtailing selection biases will still require additional drastic measures rather than just a change in inferential method. Changes in the choice of inferential methods do not necessarily address the threat of selection bias head-on. The only direct protection against selection bias is to embrace reproducible research practices, including careful choice and layout of study design and hypotheses with specified and registered in advance methods and analyses, whenever appropriate.

## References

Benjamin, D. J., et al. (2018), "Redefine Statistical Significance," *Nature Human Behaviour,* 2, 6–10. [23]

Bermon, S., and Garnier, P. Y. (2018), "Serum Androgen Levels and their Relation to Performance in Track and Field: Mass Spectrometry Results from 2127 Observations in Male and Female Elite Athletes," *British Journal of Sports Medicine*, 51, 1309–1314. [23]

Boutron, I., Altman, D. G., Hopewell, S., Vera-Badillo, F., Tannock, I., and Ravaud, P. (2014), "Impact of Spin in the Abstracts of Articles Reporting Results of Randomized Controlled Trials in the Field of Cancer: the SPIIN Randomized Controlled Trial," *Journal of Clinical Oncology*, 32, 4120–4126. [22]

Bruns, S. B., and Ioannidis, J.P (2016), "p-Curve and p-Hacking in Observational Research," *PLoS One*, 11, e0149144. [24]

Chambers, C. D. (2013), "Registered Reports: a New Publishing Initiative at Cortex," *Cortex*, 49, 609–610. [22,24]

Chavalarias, D., Wallach, J. D., Li, A. H., and Ioannidis, J. P. (2016), "Evolution of Reporting P Values in the Biomedical Literature, 1990–2015," *JAMA*, 315, 1141–1148. [20,21,22,23,24]

Cristea, I. A., and Ioannidis, J. P. (2018), "P-values in Display Items are Ubiquitous and Almost Invariably Significant: A Survey of Top Science Journals," *PLoS ONE*, 13, e0197440. [22]

Fanelli, D. (2010), ""Positive" Results Increase Down the Hierarchy of the Sciences," *PLoS One*, 5, e10068. [24]

Fanelli, D., Costas, R., and Ioannidis, J. P. (2017), "Meta-assessment of Bias in Science," *Proceedings of the National Academy of Sciences U S A*, 114, 3714–3719. [24]

Gelman, A. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–65. [22]

Gøtzsche, P. C. (2006), "Believability of Relative Risks and Odds Ratios in Abstracts: Cross Sectional Study," *BMJ*, 333, 231–234. [22]

Ioannidis, J. P. (2018), "The Proposal to Lower P-value Thresholds to .005," *JAMA*, 319, 1429–1430. [23]

Ioannidis, J. P., and Trikalinos, T. A. (2005), "Early Extreme Contradictory Estimates may Appear in Published Research: the Proteus Phenomenon in Molecular Genetics Research and Randomized Trials," *Journal of Clinical Epidemiology*, 58, 543–549. [24]

Khabsa, M., and Giles, C. L. (2014), "The Number of Scholarly Documents on the Public Web," *PLoS One*, 9, e93949. [20]

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017), "A Manifesto for Reproducible Science," *Nature Human Behaviour*, 1, 0021. [24]

Patel, C. J., Burford, B., and Ioannidis, J. P. (2015), "Assessment of Vibration of Effects due to Model Specification can Demonstrate the Instability of Observational Associations," *Journal of Clinical Epidemiology*, 68, 1046–1058. [22]

Patel, C. J., Ji, J., Sundquist, J., Ioannidis, J. P., and Sundquist, K. (2016), "Systematic Assessment of Pharmaceutical Prescriptions in Association with Cancer Risk: a Method to Conduct a Population-wide Medication-wide Longitudinal Study," *Scientific Reports*, 10, 31308. [22]

Pfeiffer, T., Bertram, L., and Ioannidis, J. P. (2011), "Quantifying Selective Reporting and the Proteus Phenomenon for Multiple Datasets with Similar Bias," *PLoS One*, 6, e18362. [24]

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. [24]

Szucs, D., and Ioannidis, J. P. (2017a), "When Null Hypothesis Significance Testing is Unsuitable for Research: a Reassessment," *Frontiers in Human Neuroscience*, 11, 390. [20]

Szucs, D., and Ioannidis, J. P. (2017b), "Empirical assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature," *PLoS Biology*, 15, e2000797. [22]

Szucs, D. (2016), "A Tutorial on Hunting Statistical Significance by chasing N," *Frontiers in Psychology*, 7, 1444. [24]

van Ravenzwaaij, D., and Ioannidis, J. P. (2017), "A Simulation Study of the Strength of Evidence in the Recommendation of Medications Based on Two Trials with Statistically Significant Results," *PLoS One*, 12, e0173184. [23]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [24]

Young, N. S., Ioannidis, J. P., and Al-Ubaydli, O. (2008), "Why Current Publication Practices may Distort Science," *PLoS Medicine*, 5, e201. [22]