Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* **100**, 407–425.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

Francis, G. (2012). Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin and Review* **19**, 151–156.

French, C. (2012). *Precognition Studies and the Curse of the Failed Replications*. *Guardian* newspaper, 15 March. http://www.guardian.co.uk/science/2012/mar/15/precognition-studies-curse-failed-replications.

Garcia-Berthou, E. and Alcaraz, C. (2004). Incongruence between test statistics and *P* values in medical papers. *BMC Medical Research Methodology* **4**(1), 13.

Gelman, A. and Weakliem, D. (2009). Of beauty, sex, and power: statistical challenges in estimating small effects. *American Scientist* **97**, 310–316.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society A* **168**, 267–306.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine* **2**(8), e124.

Marcus, A. and Oransky, I. (2010–2013). *Retraction Watch Blog*. http://retractionwatch.wordpress.com.

Rosenbaum, P. R. (2010). *Observational Studies*, 2nd edition. New York: Springer.

Simmons, J., Nelson L. and Simonsohn U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359–1366.

Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P. and Ulmer, H. (2007). The use of statistics in medical research. *American Statistician* **61**, 47–55.

Yarkoni, T. (2011). *The Psychology of Parapsychology, or Why Good Researchers Publishing Good Articles in Good Journals Can Still Get it Totally Wrong*. Citation Needed blog, 10 January. http://www.talyarkoni.org/blog/2011/01/10/the-psychology-of-parapsychology-or-why-good-researchers-publishing-good-articles-in-good-journals-can-still-get-it-totally-wrong.

Biostatistics (2014), **15**, 1, *pp.* 23–27
doi:10.1093/biostatistics/kxt035
Advance Access publication on 25 September 2013

Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature

STEVEN N. GOODMAN

Stanford University, Stanford, CA 94305, USA

steve.goodman@stanford.edu

It is a pleasure to have the opportunity to comment on this contribution by Jager and Leek (2013), perhaps the first broad-based empirical attempt to examine the now 8-year-old claim that most conclusions in the

medical literature are false (Ioannidis, 2005). The latter paper has had considerable traction, reportedly being the most downloaded article in the history of PLoS Medicine. Its potential positive impact was more wariness about blithely accepting the results of published studies, better understanding of the factors that lead to misleading research, and more awareness of the domain of meta-research. The potential negative impact of the claim was an unwarranted degree of skepticism, hopefully not cynicism, about truth claims in medical science. So, exploring that hypothesis further using empirical data was much needed. But before commenting on Jager and Leek's contribution, it is useful to understand the assumptions and methodology that supported the original claim, outlined in detail in 2007 (Goodman and Greenland, 2007).

Ioannidis's argument was based on simple Bayesian mathematics, the same kind used in diagnostic studies. However, two elements were incorporated that had the mathematical effect of nullifying research evidence. First, the information from an experiment used in his model was simply $P < 0.05$, not the exact P-value. Thus, an epidemiologic study generating a main finding with a $P < 0.0001$ was treated in that model identically to one with $P = 0.04$. Quantitatively, for a study with 80% power, this reduced the maximum Bayes factor—the degree to which an experiment raises the prior odds of the alternative hypothesis—from infinity to 16.

The second step involved the introduction of a "bias" factor. This was defined as the probability that even if a finding was not significant, a significant result is reported. There is little doubt that P-values are misrepresented in the literature. However, the values assigned to the bias factor in this model virtually eliminated the remaining evidential weight of findings. The bias factors assigned to all non-randomized controlled trial (RCT) designs lowered the maximum achievable Bayes factor from 16 to 2.6 for "moderate" bias (in well-powered epidemiologic studies), an 89% additional reduction in evidential effect, and to 1.1 for "high bias" studies, a 99.3% additional reduction.

The claim that "more than half of scientific results are false" followed inevitably from these numbers, since the model rendered most scientific studies almost informationless. Bayes factors of 2.6 and 1.1 barely budge prior probabilities. Ioannidis assigned a maximum of 25% prior probability to hypotheses tested in non-randomized studies. With the model capping Bayes factors at 2.6, an impossibly low maximum, if the prior probability of truth was 25%, the probability that a hypothesis was true after a non-randomized study, no matter how powerful the result, could not exceed 46%.

Ioannidis has been a leader in empirically assessing the reliability of the biomedical literature (e.g. Ioannidis *and others*, 2001; Ioannidis and Tzoulaki, 2012; Kyzas *and others*, 2005; Panagiotou and Ioannidis, 2012; Pereira *and others*, 2012), but this theoretical model was not empirically based, except perhaps via the informal estimation of prior probabilities. We stated that "the claim that half the medical literature is false depends in part on the distribution of P-values..." (Goodman and Greenland, 2007). Therefore, it is gratifying to see Jager and Leek's use of an empirical distribution of P-values to estimate the reliability of significant findings. The P-value distribution Jager and Leek find is striking, with a high proportion of quite small P-values, although a back-of-the-envelope calculation shows this should not be surprising. If a study has 90% power to detect a given effect size, if that effect size is in fact observed, the P-value will equal about 0.001 (Goodman, 1992). The same calculation for a study with 80% power shows that, for an observed effect equal to the detectable effect, the P-value will be about 0.005. In other words, if the study is designed to detect a likely effect size or a smaller one, as good studies should be, and the effect exists, quite small P-values are expected.

It is interesting, albeit probably fortuitous, that their estimate for the reliability of significant findings, 86%, is identical to the maximum posterior probability of the alternative hypothesis with 50% prior probability after a $P = 0.05$, under a Gaussian model (Berger and Sellke, 1987; Goodman, 1999). It has long been known among Bayesians that P-values in the 1–5% range generate conclusions with far less than 95% certainty if the underlying hypotheses are not already likely to be true. It would be valuable to learn from Jager and Leek's data not what the reliability is of conclusions based on $P < 0.05$, but what P-value cutoff was needed to produce a false discovery rate (FDR) of 5%. (Bayesian calculations suggest it should

be near 0.01.) This would be a very useful benchmark for an empirically based downward revision of the conventional threshold of significance for clinical and epidemiologic research. Even RA Fisher said a $P < 0.05$ only indicated that the experiment should be repeated, i.e. not that the observed relationship be regarded as true (Fisher, 1926).

Jager and Leek also quantitatively model the effect of certain kinds of biases in P-value reporting. This is worth serious attention both because the bias factor had such a large effect in Ioannidis's model, and because properly modeling prevalent biases is critical to assessing the literature's credibility. As previously noted, Ioannidis defined bias as the reporting of a significant P-value when the "actual" P-value was non-significant. This can happen in a multitude of ways, some of which are listed below:

(1) *Changing primary endpoints*: When the primary endpoint is non-significant and a secondary or surrogate endpoint is significant, the authors report both, but emphasize the significant one, maybe even falsely claiming it was primary.
(2) *Selective reporting*: The primary endpoint is non-significant and secondary or surrogate endpoints are significant, or the primary endpoint is significant only in subgroups, or the authors report only the significant results arising from multiple analyses without revealing which analyses have been performed.
(3) *Publication bias*: The results are not to the liking or interest of authors or editors (often as a result of non-significance), resulting in nonpublication or delayed publication of a study.
(4) *Fraud*: The primary endpoint is non-significant, but is reported falsely to be significant, either through outright falsification, or by manipulation of the analysis to make the $P < 0.05$.

Each of these scenarios has profoundly different effects, and must be modeled differently. Not all are necessarily even biases. That depends critically on whether one regards each study as exploring only one "global null" hypothesis regardless of endpoint (e.g. "the treatment does not affect anything"), or if each exposure–outcome relationship constitutes a different hypothesis. For example, suppose that a study of an anti-hypertensive drug is conducted that has a pre-specified primary endpoint of mortality, and a secondary endpoint of blood pressure. The results show a non-significant, negligible effect on mortality in spite of a sizable effect on blood pressure. In the paper, the authors highlight the anti-hypertensive effect, perhaps being silent or deceptive about which endpoint was pre-specified as primary. But regardless of this pre-specification, it would be perfectly reasonable to conclude that the anti-hypertensive treatment had an effect on blood pressure but not mortality. The failure of surrogates or intermediate endpoints to predict clinical endpoints is a widely recognized problem (Fleming and DeMets, 1996), and they are rarely regarded as interchangeable. So highlighting the blood pressure effect instead of the primary mortality effect does not necessarily produce a false conclusion, because there is no false reporting, no selective reporting and no multiplicity with respect to the same hypothesis; an effect on mortality or an effect on blood pressure are two distinct hypotheses. Changing primary endpoints produces the illusion of bias only if by pre-specifying one outcome as primary and the other as secondary somehow makes one outcome take precedence over the other in determining whether a treatment works.

Selective reporting is another matter, producing bias via evidence suppression, and this is the bias that Jager and Leek model. They assume that an author will report only the minimum of 20 hypothesis tests, which is fairly extreme for a clinical or epidemiologic study. Their simulation results in Figure 5(B) suggests that this would increase their FDR estimate from 14% to only about 20%. In theory, this degree of hidden multiplicity increases the type I error rate from 5% to 64%, so it is surprising to see such a modest increase in the FDR. Perhaps, this is because they are simulating actual P-values instead of binary significance verdicts, but more insight into that phenomenon would be useful. But there are still many ways in which such bias modeling would need to be adjusted if we are to rely on analyses like these.

First, there is no information on study design. Different designs, particularly randomized ones, have differing susceptibility to bias, and their results are to be viewed with different degrees of credibility. While the mathematics of FDR may allow this to be ignored (except perhaps through the shapes of the null and alternative distributions), scientific theory, and a substantial literature tell us that design makes a difference. As part of a test of the validity of their method, it would have been illuminating for Jager and Leek to have stratified their results not by journal, but by design. The fact that the American Journal of Epidemiology had a similar FDR to studies in clinical journals could be either an interesting finding or a red flag.

Second, the P-values reported in these abstracts are typically highly correlated, often associated with closely related or identical hypotheses. For example, in their abstract #10 on endocarditis, we see that each P-value refers to the same exposure–outcome analysis, albeit adjusted for different covariates. The randomized controlled trial (RCT) in example #6 reports results at 3 months and 18 months of follow-up, clearly correlated. An analysis assuming different degrees of within-study correlation for these P-values would be welcome.

Third, some percentage of the P-values did not refer to main findings, but rather ancillary observations, such as covariate differences between groups (e.g. the $P < 0.001$ in Abstract #2). The number of these P-values unrelated to study hypotheses, and the impact of their inclusion, are important to assess.

A fourth issue is the representativeness of this sample, which the authors acknowledge. While these are indeed some of the most highly cited general medical journals, they represent only a tiny fraction of the journal universe. Highly cited, influential papers are published in the leading specialty journals as well, and medical journals below this top tier publish many more papers, almost certainly with less rigorous designs and analyses. Methodologic review of papers in smaller journals is done less often than in the larger ones (Goodman *and others*, 1998). Jager and Leek's estimate of the FDR for these top journals is thus probably lower than for the medical literature as a whole.

A fifth issue is the use of abstracts as the source of the primary data. While this study probably could not have been done at this scale in any other way, abstracts are known to be an imperfect reflection of actual findings (Berwanger *and others*, 2009; Ghimire *and others*, 2012; Pitkin *and others*, 1999). Whether this would have materially affected these results would be interesting to explore in a subsample.

Sixth is the issue of false-negative findings. If statements are to be made that some percentage of scientific results are false, then it is as important to assess the percentage of results with $P > 0.05$ that represent real effects. While false-positive discovery might be the dominant error and concern in genomics, in which the FDR is most frequently applied, in the clinical literature there is significant concern about missing effective treatments, and the effect of underpowered or poorly designed studies that might produce false negatives. Global claims about the reliability of the medical literature should ideally model both errors.

Any estimate of the reliability of the medical literature must incorporate, as a primary result and not just as a sensitivity analysis, some estimate of the effect of biases, either in reporting, design, or conduct. There can no longer be any doubt that their effect is substantial. The selective reporting model used here is a good start, but more sophisticated, empirically based models need to be developed.

Whether this kind of content-free, design-free approach is the best way to assess the veracity of medical research claims is an open question, but it will serve as a valuable complement to other ways of addressing this issue. While I believe that this empirically based estimate of 86% of the reliability of statistically significant findings in the scientific literature is closer to the truth than the earlier assumption-based estimate of $<50\%$, the real truth is probably somewhere between the two, and likely varies importantly by domain of investigation and by design.

Jager and Leek's attempt to bring a torrent of empirical data and rigorous statistical analyses to bear on this important question is a major step forward, its weaknesses are less important than its positive contributions, and it should serve as a guidepost for further work. I congratulate them on moving the field forward with this contribution, and look forward to further progress.

Acknowledgments

Conflict of Interest: None declared.

References

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of *P*-values and evidence. *Journal of the American Statistical Association* **82**, 112–122.

Berwanger, O., Ribeiro, R. A., Finkelsztejn, A., Watanabe, M., Suzumura, E. A., Duncan, B. B., Devereaux, P. J. and Cook, D. (2009). The quality of reporting of trial abstracts is suboptimal: survey of major general medical journals. *Journal of Clinical Epidemiology* **62**(4), 387–392.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33**, 503–513.

Fleming, T. R. and DeMets, D. L. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine* **125**(7), 605–613.

Ghimire, S., Kyung, E., Kang, W. and Kim, E. (2012). Assessment of adherence to the CONSORT statement for quality of reports on randomized controlled trial abstracts from four high-impact general medical journals. *Trials*, **13**, 77.

Goodman, S. N. (1992). A comment on replication, *P*-values and evidence. *Statistics in Medicine* **11**, 875–879.

Goodman, S. N. (1999). Towards evidence-based medical statistics, II: the Bayes factor. *Annals of Internal Medicine* **130**, 1005–1013.

Goodman, S. N., Altman, D. G. and George, S. L. (1998). Statistical reviewing policies of medical journals: caveat lector? *Journal of General Internal Medicine* **13**(11), 753–756.

Goodman, S. N. and Greenland, S. (2007). *Assessing the Unreliability of the Medical Literature: A Response To "Why Most Published Research Findings Are False"*, Berkeley Electronic Press, Johns Hopkins University Biostatistics Working Paper #135, www.bepress.com/jhubiostat/paper135.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine* **2**(8), e124.

Ioannidis, J. P., Haidich, A. B., Pappa, M., Pantazis, N., Kokori, S. I., Tektonidou, M. G., Contopoulos-Ioannidis, D. G. and Lau, J. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Journal of the American Medical Association* **286**(7), 821–830.

Ioannidis, J. P. and Tzoulaki, I. (2012). Minimal and null predictive effects for the most popular blood biomarkers of cardiovascular disease. *Circulation Research* **110**(5), 658–662.

Jager, L. and Leek, J. (2013). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* **15**(1), 1–13.

Kyzas, P. A., Loizou, K. T. and Ioannidis, J. P. (2005). Selective reporting biases in cancer prognostic factor studies. *Journal of the National Cancer Institute* **97**(14), 1043–1055.

Panagiotou, O. A. and Ioannidis, J. P. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology* **41**(1), 273–286.

Pereira, T. V., Horwitz, R. I. and Ioannidis, J. P. (2012). Empirical evaluation of very large treatment effects of medical interventions. *Journal of the American Medical Association* **308**(16), 1676–1684.

Pitkin, R. M., Branagan, M. A. and Burmeister, L. F. (1999). Accuracy of data in abstracts of published research articles. *Journal of the American Medical Association* **281**(12), 1110–1111.