*Biostatistics* (2013), *pp.* 1–9
doi:10.1093/biostatistics/kxt036

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

JOHN P. A. IOANNIDIS

*Departments of Medicine, Health Research and Policy, and Statistics, Stanford University, 1265 Welch Road, MSOB Room X306, Stanford, CA 94305, USA*

jioannid@stanford.edu

SUMMARY

Jager and Leek have tried to estimate a false-discovery rate (FDR) in abstracts of articles published in five medical journals during 2000–2010. Their approach is flawed in sampling, calculations, and conclusions. It uses a tiny portion of select papers in highly select journals. Randomized controlled trials and systematic reviews (designs with the lowest anticipated false-positive rates) are 52% of the analyzed papers, while these designs account for only 4% in PubMed in the same period. The FDR calculations consider the entire published literature as equivalent to a single genomic experiment where all performed analyses are reported without selection or distortion. However, the data used are the $P$-values reported in the abstracts of published papers; these $P$-values are a highly distorted, highly select sample. Besides selective reporting biases, all other biases, in particular confounding in observational studies, are also ignored, while these are often the main drivers for high false-positive rates in the biomedical literature. A reproducibility check of the raw data shows that much of the data Jager and Leek used are either wrong or make no sense: most of the usable data were missed by their script, 94% of the abstracts that reported $\geqslant 2$ $P$-values had high correlation/overlap between reported outcomes, and only a minority of $P$-values corresponded to relevant primary outcomes. The Jager and Leek paper exemplifies the dreadful combination of using automated scripts with wrong methods and unreliable data. Sadly, this combination is common in the medical literature.

*Keywords*: Bias; False discovery rate; Science; Selection bias; $P$-value.

## 1. INTRODUCTION

Jager and Leek apply a standard false-discovery rate (FDR) approach to assess abstracts of papers published in the New England Journal of Medicine (NEJM), Journal of the American Medical Association (JAMA), Lancet, British Medical Journal (BMJ), and American Journal of Epidemiology (AJE). They claim to provide an "empirical estimate" of false positives, but no empirical estimates are provided: there is no replication effort to validate published results. Instead, inferences depend on modeling with implausible assumptions. As discussed below, the work is fundamentally flawed in its data, sampling, calculations, and conclusions.

## 2. OVERVIEW OF MAIN FLAWS

### 2.1 *Failure to recognize that bias is often the major cause of high false-positives*

Instead of challenging my real arguments about false research findings (Ioannidis, 2005a), Jager and Leek attacked a strawman thereof, using a diagram from a blog and claiming that this represents "The theoretical argument suggesting most published research is false". That diagram refers to a research world without bias. Indeed, if we reach that perfection one day, FDR ∼14% may become possible. Fourteen percentage is still much higher than the nominal errors suggested by the reported $P$-values (largely $<1\%$ in the analyzed sample of papers), but desirable. Conversely, my conclusion that "for most study designs and settings, it is more likely for a research claim to be false than true" was drawn from simulations that accounted for potential bias. The high false-positive rates inferred by these simulations have now been repeatedly validated empirically by numerous replication efforts in very diverse fields of biomedical research and beyond (see Section 2.7).

### 2.2 *Conglomeration of proposed discoveries with validated results*

In my PLoS Medicine paper (Ioannidis, 2005a) I separated first proposed discoveries ("research findings") versus subsequent evidence. I have repeatedly made a plea for replication efforts and larger-scale evidence from additional studies and careful meta-analyses, to increase the credibility of scientific results. Conversely, Jager and Leek conglomerated initial discovered effects, subsequent replication efforts, and even meta-analyses in their samples.

### 2.3 *Ignoring that false-positive rates vary enormously across fields and study designs*

A central point of the PLoS Medicine paper (Ioannidis, 2005a) was how false-positive rates are expected to vary greatly depending on the field and type of study design from ∼0% to ∼100%, pointing that some study designs have high credibility, e.g. results from well-powered randomized controlled trials (RCTs) with little bias were estimated to be 85% likely to be true even with $P \sim 0.05$ ($>95\%$ with $P < 0.0001$). The same was estimated to apply for well-conducted confirmatory systematic reviews/meta-analyses. Putting in the same FDR analysis $P$-values from RCTs, meta-analyses, genetics, prognostics, diagnostics (and more) makes no sense.

### 2.4 *Sampled papers were mostly those with lowest false-positive rates*

The sample of papers analyzed includes predominantly designs with the lowest possible false-positive rates: 52% were either RCTs or systematic reviews, while these two designs account for only 7% of all papers (30% of those with abstracts) in the same journals (Table 1). Sampled RCTs and systematic reviews from these premier journals are among the largest and best in the literature. The problem with statistically significant well-conducted RCTs and meta-analyses is not so much that they find false positives, but inflated effect size estimates due to the "winner's curse". See Ioannidis (2008) for references on this issue; Siontis *and others* (2011) on inflated effects in major journals; and Pereira *and others* (2011) on inflation of large effects in $>85\,000$ meta-analyses. Jager and Leek did not address effect sizes.

### 2.5 *The assumed distributions of P-values are implausible*

The main assumption that "by definition, the $P$-values for false-positive findings are uniformly distributed between 0 and 1" and that true positives are represented by a peculiarly censored beta distribution is entirely implausible. This standard FDR method works in single datasets using the same platform (e.g. assessing

Table 1. *Number of articles and types of designs in the sample analyzed by Jager and Leek versus all the articles published in the five journals and in the entire PubMed database for 2000–2010*

| | | Proportion (%) with different designs | | | |
|---|---|---|---|---|---|
| | Number of articles | RCT | SR | RCT or SR | RCT or SR or CT |
| Jager and Leek sample | | | | | |
| AJE | 499 | 3 | 5 | 8 | 15 |
| BMJ | 667 | 44 | 21 | 59 | 73 |
| JAMA | 1425 | 38 | 15 | 49 | 56 |
| Lancet | 1317 | 50 | 6 | 56 | 62 |
| NEJM | 1414 | 63 | 1 | 63 | 70 |
| All five journals | 5322 | 45 | 9 | 52 | 59 |
| Same five journals: all articles | 76 862 | 4* | 2* | 7* | 8* |
| Same five: articles with abstracts | 15 658 | 21* | 11 | 30* | 38* |
| All PubMed: all articles | 7 199 704 | 2* | 2* | 4* | 6* |
| All PubMed: articles with abstracts | 5 865 305 | 3* | 2* | 4* | 7* |

RCT, randomized controlled trial; SR, systematic review; CT, clinical trial.
The type of study is as assigned by PubMed. PubMed may assign more than one type to the same article and the type assigned is subject to some error, but it is overall very reliable for papers published in 2000–2010. The Jager and Leek sample represents the articles they analyzed in their paper (PMIDs derived from their pvalueData R file).
* $<0.001$ for comparison of the Jager and Leek sample versus other articles.

thousands of gene tags or brain signals) and where all analyses are ideally reported without any selection and distortion. Conversely, the medical literature includes thousands of different studies with very different designs and very different types of measurements and analyses, and suffers selective reporting and other biases. Reported $P$-value distributions are primarily determined by reporting biases and field-specific conventions.

### 2.6 *P-values in abstracts are extremely selected*

An inferential model that assumes that the $P$-values reported in the abstracts are representative of all the $P$-values generated in all primary statistical analyses is totally implausible. As Gotzsche (2006) has shown empirically, one cannot believe $P$-values in abstracts. The average abstract of a research paper reports only a couple of $P$-values, while hundreds, thousands, or even millions of $P$-values are generated in analyzing data that end up being summarized in a published paper. $P$-values reported in the full text of the paper are a distorted sample of those obtained in the universe of all analyses, typically favoring more significant results. $P$-values reported in the abstract are an even more highly distorted sample of the already distorted sample of $P$-values reported in the main paper. Not surprisingly, most investigators report some of the best-looking $P$-values in their abstracts and Jager and Leek indeed show that 81% of the $P$-values in abstracts were $<0.05$. Empirical data have demonstrated that non-significant results are disappearing from the published literature and that least credible scientific fields publish the highest proportions of significant results (Fanelli, 2010a, 2010b, 2012). Significance chasing and poor replication practices are documented in diverse fields of science, including psychological sciences, economics, marketing, etc. (see selected references in Appendix References 1 of supplementary material available at *Biostatistics* online). The Jager and Leek paper fits in this expanding literature, but for the exact opposite reason than what they infer. Journals with more prominently significant $P$-values may not have a lower FDR (as Jager and Leek would infer with their method), but simply worse selective reporting bias. As an extreme scenario, let us

suppose a hypothetical "null scientific field" where all the tested associations are null; if scientists in that field report in abstracts only whatever $P$-values reach $<0.001$, Jager and Leek would estimate FDR $\sim$0% in that field, instead of the true FDR $\sim$100%.

Jager and Leek have added an interesting extension in their Figure 5 trying to model the potential impact of some types of biases. However, they have not modeled anything close to the strong biases likely to prevail in the literature. For example, they claim that their simulation, where the $P$-values reported were only the minimum $P$-values from 20 hypothesis tests, represents "an extreme case of $P$-value hacking". Selection forces for $P$-values in the abstracts of biomedical papers are often far more intense. For example, in many omics studies, the $P$-values reported in abstracts are less than one millionth of those obtained in analyses performed, and even in traditional epidemiology often $\ll$1% of the $P$-values obtained in analyses are reported in the abstract. Jager and Leek ignore some more appropriate exploratory methods for bias modeling, such as selection models—see, for example, Pfeiffer *and others* (2011), Hedges and Vevea (1996, 2005), and Vevea and Hedges (1995). Bias modeling is an interesting exploratory avenue for future research, but it remains as speculative as any post hoc exercise.

### 2.7 *A rich literature on empirical replication rates is ignored*

Jager and Leek claim that "to date there has been no empirical approach for evaluating the rate of false positives across an entire journal or across multiple journals". This is misleading. "Empirical approach" means independent replication in subsequent studies. There are thousands of replication studies and numerous empirical evaluations, each including multiple studies, where researchers have tried to replicate previously proposed findings. These evaluations confirm ubiquitously the high rates of false positives and the high rates of inflated effects in original discoveries. For example, several empirical evaluation studies of attempted large-scale replications of over a thousand proposed candidate gene genetic associations show replication rates ranging from 0% to 6%, with an average of 1.2%, i.e. FDR $\sim$98.8% (see Ioannidis *and others*, 2011 for an overview). Empirical evaluations by scientists at Amgen and Bayer of preclinical studies done by leading academic investigators also show that the large majority of claimed discoveries were false positives that could not be reproduced (Prinz *and others*, 2011; Begley and Ellis, 2012). The literature on non-replication includes already many thousands of references.

### 2.8 *The sample of journals is highly selected*

Lancet, JAMA, BMJ, NEJM, and AJE represent a tiny and highly select sample that is not representative of the medical literature. They try to cherry pick papers with very low acceptance rates. These five journals published only 0.27% of the 5 865 305 PubMed articles with abstracts in 2000–2010 (Table 1). They published many RCTs and systematic reviews (designs with expected low FDR), while these two designs accounted for only 4% of PubMed articles with abstracts (Table 1).

### 2.9 *Simpson's paradox invalidates time-trend analyses*

Jager and Leek find a constant FDR over time. Time-trend analysis without stratification by design type is misleading. For example, even if the FDR increased for all types of designs over time, the average FDR for the five journals would remain the same (or decrease), if these journals selected more stringently over time for papers with high-credibility designs and left the others to lesser journals. For example, the 5 journals published 158 systematic reviews in 2000, but 218 in 2010, despite diminishing their total published articles with abstracts from 1787 in 2000 to 1285 to 2010. Moreover, some current designs did not even exist in 2000 (e.g. genome-wide association studies).

### 2.10 *The sample of studies with P-values in the abstract is highly selected*

Jager and Leek found 5322 papers with $P$-values among 77 430 papers screened, i.e. inferences are made about only 7% of the papers in the five journals. This is a highly selected sample, adding another cause of selective distortion. Moreover, studies that give numbers in their abstract are often of better quality and more likely to represent designs with lower false-positive rates, e.g. RCTs.

### 2.11 *Additional problems exist with epidemiological studies and their journals*

Jager and Leek note that AJE publishes mostly observational studies and interpret the data saying that "this suggests that the FDR is somewhat consistent across different journal and study types". This is highly misleading. Simply AJE (and the observational studies that it tends to publish) select and report very nice-looking $P$-values and effects in the abstracts (Kavvoura *and others*, 2007). The possibility that epidemiological studies have the same (or even better) FDR as the typically large RCTs published in the other four major journals goes against (epidemiological) theory and extensive empirical evidence (Ioannidis, 2005b; Young and Karr, 2011). For example, consider hormone therapy for preventing coronary heart disease, low-fat diet, vitamin E and coronary heart disease, beta-carotene and cancer chemoprophylaxis, estrogen and dementia, and so forth (see Appendix References 2 of supplementary material available at *Biostatistics* online). Young and Karr (2011) have shown that of 52 major claims made by observational studies, none was validated when tested in RCTs.

Reported $P$-values in observational studies do not suffer just from selective reporting but also from confounding bias. In a typical observational dataset, almost all variables may seem associated with all the others (as nicely demonstrated by Smith *and others*, 2007) due to the dense correlation patterns (Ioannidis *and others*, 2009). This does not mean that correlation is causation or that these associations are true and replicable in other populations where confounder patterns are different; let alone that one can improve important outcomes by modifying these variables. If unaccounted confounding could be accounted, most of these associations might evaporate.

### 2.12 *Selecting only the* $P < 0.05$ *range deflates FDR estimates*

The attraction of passing the 0.05 mark is not absolute (Ioannidis and Trikalinos, 2007): $P$-value clustering extends in the range of $P = 0.05$ and possibly even higher (Kavvoura *and others*, 2008). Even in RCTs, investigators claim significance for many non-significant results by applying "spin" (Boutron *and others*, 2010). For example, Jager and Leek exclude a huge peak of 470 $P$-values at 0.05 from FDR calculations, even though most are truncated (i.e. $P < 0.05$ anyhow) and almost all are considered "significant" by their authors. This seemingly subtle exclusion markedly affects the overall FDR estimates, and totally invalidates FDR estimates comparing journals and years (see Appendix text 1 of supplementary material available at *Biostatistics* online).

### 2.13 *Much of the data is either wrong or makes little sense*

I probed further some raw data. First, I checked whether indeed only 5322 papers had usable data. For studies that report only effect sizes and confidence intervals (CIs) without $P$-values, one can easily obtain equivalent $P$-values: $z = \theta/[(\text{upper 95\% CI} - \text{lower 95\% CI})/3.92]$. These are important to include, since effect sizes and CIs are more commonly used when 95% CIs approach the null, i.e. $P$-values are modest. A search of papers with abstracts in the 5 journals AND ($P$ [tw] OR CI [tw] OR confidence [tw]) for 2000–2010 yielded 10 805 articles: 19/20 (95%) randomly sampled were eligible. Moreover, many of the 5322 papers retrieved also had additional usable effect sizes and CIs. In all, Jager and Leek missed most of the eligible data, and selected those with systematically lower $P$-values, thus deflating FDR estimates.

Table 2. *In-depth analysis of 20 randomly selected abstracts with P-values*

| PMID | Design (*n*) | *P*-values | Primary (sign) | Correlation/overlap of outcomes with reported *P*-values |
|---|---|---|---|---|
| 18068514 | RCT (50 + 22) | 0.634, 0.561* | 0 (0) | Yes (diastolic and systolic blood pressure) |
| 16525139 | RCT (323) | <0.001, 0.01, 0.90, 0.001* | 1 (1) | Yes (complete response, 5-year event-free survival, 5-year overall survival, median survival after relapse) |
| 18950853 | RCT (203) | <0.0001, 0.0001* | 1 (1) | No |
| 15173147 | RCT (347) | 0.01, 0.03, 0.40* | 1 (0) | Yes (1-year hemoglobin, 1-year and 2-year age-adjusted hemoglobin <11) |
| 15928285 | RCT (70) | <0.001, <0.001, 0.01* | 2 (2) | Yes (recurrence of atrial fibrillation, hospitalization, atrial fibrillation episodes) |
| 16267322 | RCT (905) | 0.05, 0.12, <0.001, 0.003* | 2 (1) | Yes (response, remission, sustained response, remission) |
| 15257998 | PO (326 events) | 0.99, 0.12* | 0 (0) | Yes (cancer risk with folate intake, cancer risk with energy-adjusted folate intake) |
| 12867108 | RCT (1701) | 0.03, 0.01, 0.0004* | 0 (0) | Yes (ipsilateral invasive disease, ipsilateral ductal cancer, ipsilateral ductal cancer) |
| 19502645 | RCT (2368) | 0.97, 0.89, 0,70, 0.13, 0.01, 0.002, 0.003* | 4 (0) | Yes (death, death or major event, death, death or major event, death or major event in stratum, interaction for stratum by study group, severe hypoglycemia) |
| 11812558 | PO (180) | <0.0001, 0.002* | 1 (0) | Yes (disease-free survival between three tumor groups, disease-free survival for Duke A versus B in one of the three groups) |
| 15950715 | CC (334) | 0.07* | 0 (0) | Only one *P*-value reported |
| 19181729 | RCT (18858) | 0.08, 0.02, <0.001* | 1 (0) | Yes (breastfeeding, age, and income associations) |
| 11705561 | CC (604) | 0.03, 0.05, 0.017* | 0 (0) | Yes (wheeze risk with hookworm infection, wheeze risk with Der p1 level) |
| 11907289 | RCT (213) | <0.001* | 1 (1) | Only one *P*-value reported |
| 19741227 | PO (19) | 0.02, 0.02* | ? (?) | Yes (JC virus in urine, JC virus in blood cells) |
| 12215131 | RCT (176) | <0.001, 0.02* | 1 (1) | Yes (success at 3 months, success at 18 months) |
| 12241661 | CC (no data) | <0.0001 | 0 (0) | Only one *P*-value reported |
| 12181103 | CC (264) | 0.0001* | 0 (0) | Only one *P*-value reported |
| 11943693 | CS (1927) | 0.02, <0.0001* | 0 (0) | Yes (spinal BMD, femoral BMD) |
| 14693873 | RC (513) | <0.001, <0.02, 0.01, 0.03, 0.001 | 1 (1) | Yes (unadjusted, adjusted, propensity matched, matched and adjusted, subgroup adjusted effect of valve surgery on death) |

*n*, sample size; RCT, randomized controlled trial; PO, prospective observational; CC, case–control; CS, cross-sectional; RC, retrospective cohort; primary (sign), number of *P*-values in abstract that refer to characterized primary endpoints (in parenthesis number with <0.05); BMD, bone mineral density.
12558 [PubMed - indexed for MEDLINE].
*Additional study outcomes are stated in the results of the abstract, but without *P*-value given. ?, unclear.

Then I read 20 randomly selected abstracts with $P$-values (the 10 sampled by Jager and Leek and another 10 that I sampled) (Table 2, Appendix References 3 of supplementary material available at *Biostatistics* online). Of the 20, 11 papers (55%) represent RCTs, and 10/11 have larger sample size than the average RCT in the literature which is $n = 80$ (Chan and Altman, 2005). Yet, only 11 of the 20 papers give $P$-values for what are defined as pre-specified primary outcomes. Even then it is unclear whether the exact analysis was indeed pre-specified: empirical evidence from RCTs (Chan *and others*, 2008) shows that deviations between protocols and publications are very common for primary outcome analyses (25/42), and almost ubiquitous for subgroup analyses (25/25), and adjusted analyses (23/28).

The raw data that Jager and Leek extracted and analyzed make little sense. Even though 17/20 abstracts (85%) give some significant $P$-value(s), only 7/20 (35%) give significant $P$-values for (claimed) primary endpoints. Out of the 20, 17 (85%) abstracts report some results qualitatively without giving their $P$-values—apparently these are mostly non-impressive (14/17 report qualitatively "negative" results). Whenever there is $\geqslant 2$ $P$-values reported in the abstract, with one exception (15/16, 94%) the $P$-values refer to analyses that are not independent but highly correlated and/or overlapping. Overlapping, correlated $P$-values largely drive the calculations. For example, the automated script extracts five "independent" $P$-values for PMID = 14693873; but simply reading the abstract one realizes that it is the very same outcome presented in unadjusted, adjusted, propensity-matched, matched-plus-adjusted, and subgroup-adjusted analyses. The paper is possibly best represented by the $P = 0.03$ for the model that is propensity-matched plus adjusted; but the other four $P$-values with the more spectacular 0.02, 0.01, 0.001, $<0.001$ are also counted, deflating the estimated FDR. Finally, many of the non-primary significant $P$-values represent largely irrelevant analyses. For example, PMID = 19181729 is an RCT that assesses whether setting new breastfeeding groups can affect breastfeeding and it shows a "negative" $P = 0.08$ for the primary outcome (breastfeeding). However, the abstract incidentally reports also two "positive" $P$-values (0.02 and $<0.001$) on the association of age and income with characteristics that were clearly neither primary nor secondary outcomes; this is information of little or no relevance on factors whose modification is impossible (age) or notoriously difficult unfortunately (income). The abstracts of the crème de la crème of the biomedical literature are a mess. No fancy informatics script can sort out that mess. One still needs to read the papers.

### 2.14 *Final comment*

I congratulate Jager and Leek on their adoption of reproducible research practices. This has allowed probing their work and discovering more promptly the major errors made, thus speedily putting their claims to rest. Instead of proving the reliability of the medical literature, their paper exemplifies how badly things can go when automated scripts are combined with wrong methods and unreliable data. Sadly, this dreadful triplet remains common in published papers, even those by the best authors and in the best journals.

## Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

## Acknowledgments

*Conflict of Interest:* None declared.

## References

Begley, C. G. and Ellis, L. M. (2012). Drug development: raise standards for preclinical cancer research. *Nature* **483**, 531–533.

Boutron, I., Dutton, S., Ravaud, P. and Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal of the American Medical Association* **303**, 2058–2064.

Chan, A. W. and Altman, D. G. (2005). Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* **365**, 1159–1162.

Chan, A. W., Hrobjartsson, A., Jorgensen, K. J., Gotzsche, P. C. and Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *British Medical Journal* **337**, a2299.

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One* **5**, e10271.

Fanelli, D. (2010b). "Positive" results increase down the Hierarchy of the Sciences. *PLoS One* **5**, e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904.

Gotzsche, P. C. (2006). Believability of relative risks and odds ratios in abstracts: cross sectional study. *British Medical Journal* **333**, 231–234.

Hedges, L. V. and Vevea, J. L. (1996). Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics* **21**, 299–332.

Hedges, L. V. and Vevea, J. L. (2005). Selection method approaches. In: Hannah R. Rothstein, Alexander J. Sutton, Michael Borenstein (editors), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. New York: Wiley.

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

Ioannidis, J. P. (2005b). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* **294**, 218–228.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648.

Ioannidis, J. P., Loy, E. Y., Poulton, R. and Chia, K. S. (2009). Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Science Translational Medicine* **1**, 7ps8.

Ioannidis, J. P., Tarone, R. and McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* **22**, 450–456.

Ioannidis, J. P. and Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials* **4**, 245–253.

Kavvoura, F. K., Liberopoulos, G. and Ioannidis, J. P. (2007). Selection in reported epidemiological risks: an empirical assessment. *PLoS Medicine* **4**, e79.

Kavvoura, F. K., McQueen, M. B., Khoury, M. J., Tanzi, R. E., Bertram, L. and Ioannidis, J. P. (2008). Evaluation of the potential excess of statistically significant findings in published genetic association studies: application to Alzheimer's disease. *American Journal of Epidemiology* **168**, 855–865.

Pereira, T. V., Horwitz R. I. and Ioannidis, J. P. (2011). Empirical evaluation of very large treatment effects of medical interventions. *Journal of the American Medical Association* **308**, 1676–1684.

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.

Prinz, F., Schlange, T. and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* **10**, 712.

Siontis, K. C., Evangelou, E. and Ioannidis, J. P. (2011). Magnitude of effects in clinical trials published in high-impact general medical journals. *International Journal of Epidemiology* **40**, 1280–1291.

Smith, G. D., Lawlor, D. A., Harbord, R., Timpson, N., Day, I. and Ebrahim, S. (2007). Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine* **4**, e352.

Vevea, J. L. and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60**, 419–435.

Young, S. S. and Karr, A. (2011). Deming, data, and observational studies: a process out of control and needing fixing. *Significance* **8**, 116–120.