Summary and discussion of: "Why Most Published Research Findings Are False"

Statistics Journal Club, 36-825

Dallas Card and Shashank Srivastava

December 10, 2014

1 Introduction

Recently, there has been a great deal of attention given to a perceived "replication crisis" in science, both in the popular press (Freedman 2010, "Trouble at the lab" 2013, Chambers 2014), and in top-tier scientific journals (Ioannidis 2005a, Ioannidis 2005b, Schooler 2014).

To a certain extent, science is understood to be an accumulative, iterative, self-correcting endeavour, where mistakes are a normal short-term side-effect of a long-term process of accumulating knowledge. However, concern over a number of high profile retractions (Wakefield et al. 1998, Potti et al. 2006, Sebastiani et al. 2010) has had people questioning the standards and ethics of both scientists and science journals.

Most discussions of this topic begin with a reference to a 2005 paper called by John Ioannidis, where he argues that under reasonable assumptions, more than half of claimed discoveries in scientific publications are probably wrong, particularly in fields such as psychology and medicine. Ioannidis has published a large number of papers on this theme, both empirical investigations of the literature, and the more theoretical 2005 paper. This document will summarize the important points in this body of work, as well as a dissenting view (Jager and Leek 2014).

2 Why most published research findings are false

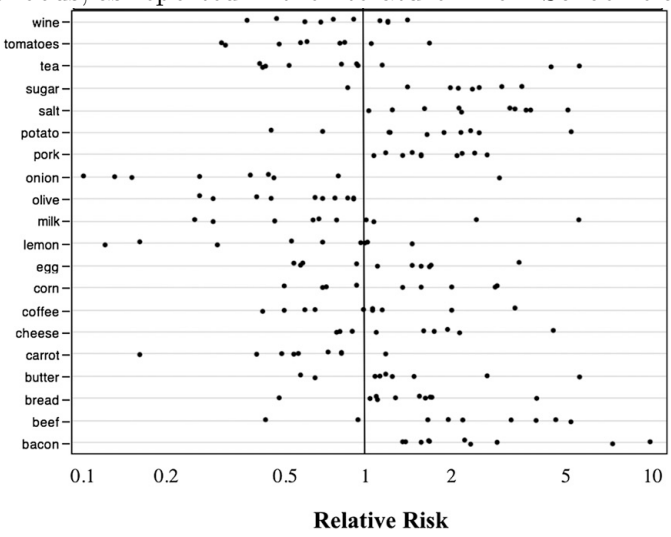
2.1 Empirical evidence

There is now a compelling amount of empirical evidence that initial claims in medical research, particularly those based on observational studies, are likely to be exaggerated, and prone to subsequent refutation or correction by larger or better designed experimental studies.

In an illustrative study, Schoenfeld and Ioannidis (2013) choose the first 50 ingredients from randomly chosen recipes from the Boston Cooking-School Cook Book, and searched the medical literature for studies linking these ingredients to various forms of cancer. For 40 of these, they found at least one study, and for 20 they found at least 10 studies. Examining the reported associations, the authors found a pattern of strong effects reported

with relatively weak statistical support, with larger effect sizes reported in individual studies than in meta-analyses. In addition, they found a dramatically wide range of reported relative risks associated with an additional serving per day with each food item, well beyond what seems credible for more well-established effects (see Figure 1). In addition, the authors also observed a bimodal distribution of normalized (z) scores associated with P-values, a finding which conforms to the idea of a publication bias, in which journals favour the publication of significant results over null findings.

Figure 1: Estimated relative risk of various types of cancer associated with an extra serving per day of various foods, as reported in the literature. From Schoenfeld and Ioannidis 2013.



A more ambitious study (Ioannidis 2005a) the most highly-cited interventional studies in the most influential medical journals published between 1990 and 2003. Of the 49 studies with more than 1000 citations examined by Ioannidis, four were null results which contradicted previous findings, seven were later contradicted by subsequent research, seven were subsequently found to have weaker effects than claimed in the highly-cited papers, 20 were replicated, and 11 had been largely unchallenged. In all cases, a judgement of replication or refutation was based on subsequent studies that included a larger study population, or a more rigorous protocol, such as a randomized control trial as opposed to an observational study. Observational studies and those with small study populations were also found to be more likely to be subsequently contradicted.

Citation count is not the same as influence or importance. Moreover, there is no guarantee that a failed replication attempt invalidates an earlier study. The point here, however, is that the majority of published findings in medical research are not based on large, randomized-control trials, but small observational studies. As the author concludes, because of the propensity for subsequent contradictory evidence to even some of the most-highly cited studies, "evidence from recent trials, no mater how impressive, should be interpreted with caution, when only one trial is available." (ibid.)

Finally, one of Ioannidis's earlier studies examined replication in genetic association studies (Ioannidis et al. 2001). The authors analyzed 26 meta-analyses of 370 studies

focused on 36 genetic associations with a variety of diseases, as well as the first individual study for each of these associations. For eight of the 36 associations, the results of the first study differed significantly from the eventual findings in subsequent meta-analyses (see Figure 2a). For eight other studies, the initial paper did not claim a statistically significant association, but such as association was demonstrated by subsequent meta-analyses (see Figure 2b). Of the remaining 20, 12 found no significant associations in the initial study or at the end of the meta-analysis, and eight found a significant association in the initial study with no disagreement beyond chance in subsequent research. In general, there was modest correlation between the initially-reported strength of association and the findings of meta-analyses. In at least 25 of the 36 cases, the strength of the effect claimed in the initial study was stronger than what was found in subsequent research, again suggesting that we should be skeptical of the conclusions reached by the first published study on any topic.

2.2 Theoretical argument

In his 2005 PLoS Medicine paper, Ioannidis presents a straightforward argument as to why we should not be surprised by poor agreement of subsequent research with initial findings, particularly in fields such as medicine and psychology. In essence, testing a large number of hypotheses which are in fact false will lead to many false discoveries. Moreover this situation will be aggravated by multiple independent investigations, and by various types of bias.

Assume that a study is investigating c null hypotheses, of which some number are in fact true, and the remainder are false. Let the ratio of false null hypotheses to true null hypotheses be R. The total number of false null hypotheses is therefore given by $cR/(1+R)$ and the number of true null hypotheses by $c/(1+R)$. We expect that some (but not all) of the false null hypotheses will in fact be rejected, and claimed as "discoveries", and this number is given by $(1-\beta)cR/(1+R)$, where $(1-\beta)$ is the power. Similarly, some number of the true null hypotheses will also be rejected (so called "false positives"), as a function of the type I error rate, α: $\alpha c/(1+R)$. These results are summarized in Table 1:

Table 1: Expected number of research findings and true relationships

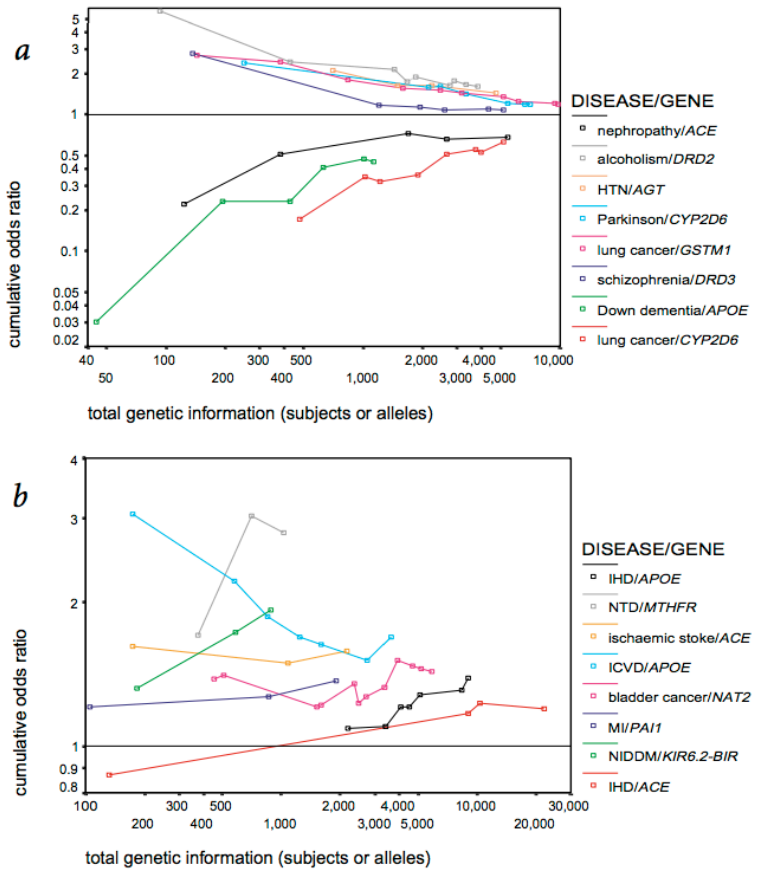
	H_0 false	H_0 true	Total
All	$\frac{cR}{1+R}$	$\frac{c}{1+R}$	c
Reject H_0	$\frac{(1-\beta)cR}{1+R}$	$\frac{\alpha c}{1+R}$	$\frac{(1-\beta)cR+\alpha c}{1+R}$

Dividing the number of discoveries by the total number of rejected null hypotheses gives us the positive predictive value (PPV), which is equal to one minus the false discover rate (FDR):

$$\text{PPV} = 1 - \text{FDR} = \frac{(1-\beta)R}{(1-\beta)R+\alpha}$$

Thus, as long as $(1-\beta)R \geq \alpha$ (which is completely reasonable for an small but adequately powered experiment), we would expect the FDR to be less than 50%. We can, however, see that as the Type I error rate (α) increases, or power (1-β) decreases, or the ratio of false to

Figure 2: Cumulative odds ratio as a function of total genetic information as determined by successive meta-analyses. a) Eight cases in which the final analysis differed by more than chance from the claims of the initial study. b) Eight cases in which a significant association was found at the end of the meta-analysis, but was not claimed by the initial study. From Ioannidis et al. 2001.



true null hypotheses (R) decreases, the FDR will increase.

Moreover, the effect of bias can alter this dramatically. For simplicity, Ioannidis models all sources of bias as a single factor u, which is the proportion of null hypotheses that would not have been claimed as discoveries in the absence of bias, but which ended up as such because of bias. There are many sources of bias which will be discussed in greater detail in section 2.3.

The effect of this bias is to modify the equation for PPV to be:

$$\text{PPV} = 1 - \text{FDR} = \frac{(1-\beta)R + u\beta R}{(1-\beta)R + u\beta R + \alpha + u(1-\alpha)}$$

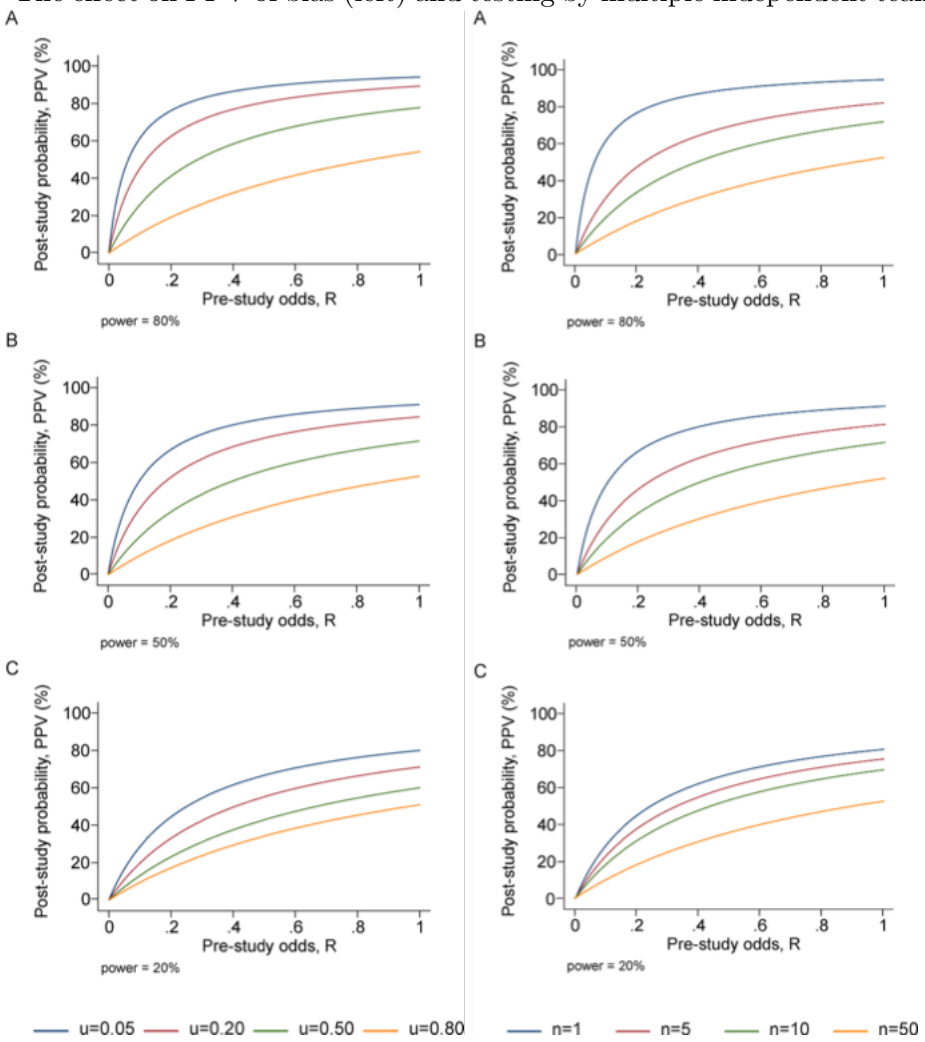
Thus, as bias increases, PPV will go down, and FDR will go up. The effect of this can be seen for various levels of power and bias in Figure 3 (left).

Similarly, multiple groups independently replicating the same experiment will also increase the FDR. In particular, if n is the number of teams investigating the same hypotheses, the equation for PPV (in the absence of bias) becomes:

$$\text{PPV} = 1 - \text{FDR} = \frac{(1-\beta^n)R}{(1-\beta^n)R + 1 - (1-\alpha)^n}$$

The effects of this are illustrated in Figure 3 (right).

Figure 3: The effect on PPV of bias (left) and testing by multiple independent teams (right)



In the paper, Ioannidis gives some rough estimates of values for these various parameters in different settings, and the resulting PPV. For example, in a large, randomized control trail, the associated expense means that the hypothesis being tested is likely to be true, that the study will be adequately powered, and that it will hopefully be done in a relatively

unbiased manner. In addition, pre-registration helps to ensure that the number of teams investigating the question is known. With values of $\alpha = 0.05$, $(1 - \beta) = 0.8$, $R = 1 : 1$, $u = 0.1$, and $n = 1$, we achieve an expected FDR of 0.15. By contrast, with a high Type I error rate, low power, high bias, investigation by many independent teams, or many false null hypotheses being tested, this analysis suggests an FDR well above 0.5 can easily be attained.

While these numbers are very approximate, and actual values for R or u will never be known in practice, this paper provides a convincing argument that the actual rate of FDR in the literature is much larger than the nominal rate of 5%, and arguably higher than 50%, especially when one considers that effect of various types of bias, to which we turn next.

2.3 Bias

In the article, Ioannidis describes some guidelines as to what to be critical of in assessing the probability that a claimed discovery is in fact true. Some of these follow directly from the model, namely low power studies (for example, because of a small sample size or small effect size), and those that test many hypotheses, are less likely to have true findings. Others have more to do with bias. Important factors to consider include:

Fraud / conflict of interest Large financial interests are an important part of the research community in medicine, and financial conflicts of interest should be taken into consideration. This includes not only outright fraud (e.g. Wakefield et al. 1998, Potti et al. 2006), but also bias with respect to which hypotheses to test, selective reporting, and lack of objectivity.

"Hot" areas of science The more excitement surrounding a particular field, the more teams will there be investigating specific questions, with strong competition to find impressive "positive" results. Similarly, journals eager to publish discoveries in these fields may end up with lower standards or lack of critical judgement.

Flexibility in analysis The more flexibility an experiment offers in terms of designs, outcomes, and analysis; the more potential there is to transform 'negative' results into 'positive' ones. In this regards, this suggests that standardization of methodology and analysis is likely to reduce the risk of false positives. This subsumes several different types of biases that can creep into analysis (e.g., sampling bias, exclusion bias, systemic errors, etc).

Pre-selection If a greater number of hypothesis are tested, the more likely it is to find false positives. Hence, hypothesis pre-selection is important for high fidelity research findings. This also implies that research findings are much more likely to be true in confirmatory designs (such as RCTs or meta-analyses), rather than high throughput preliminary hypothesis generating experiments such as using microarrays.

3 Counter-view (Jager and Leek): Why most research findings are true

The body of work by Ioannidis and others raised recognition of the kinds of biases that may lead to spurious research findings, and also called into question research findings which could not be corroborated on subsequent enquiry. However, recent work by Jager and Leek 2014 argues that Ioannidis overstates this case, and that false findings in research literature are not as common as Ioannidis suggests.

The major arguments against Ioannidis's analysis question his assumption that most tested hypothesis have low pre-study probabilities of being true. Additionally, Ioannidis's analysis is purely theoretical (apart from very informal estimates of R values), and does not directly draw any support from any empirical data. From such a perspective, it could be claimed that the Ioannidis 2005b model only explains what *might* happen if scientists blindly use a fixed significance level threshold for all analysis. However, this estimate would be too aggressive if researchers usually study hypotheses that have high pre-study likelihood of being true.

Jager and Leek 2014 instead suggest using a data-driven model to estimate the FDR across a large number of studies. For the purpose of their analysis, they use the empirical distribution of P-values found among abstracts of five leading bio-medical journals. Their rationale for the use of P-values for estimating the science-wise false discovery rate (SWFDR) is that P-values are ubiquitous as data-sources in research literature. Also the use of P-values for hypothesis testing remains the most widely-used approach in statistically screening results in several areas, such as medicine and epidemiology. Their procedure hinges on collecting the empirical distribution of a large number of P-values, and then borrowing an estimation procedure used in genetic studies to estimate the proportion of P-values coming from the null and alternative distributions.

3.1 Two groups model and the false discovery rate

Efron and Tibshirani 2002 first presented the 'two-groups model' for modeling P-values from multiple hypothesis tests. Under this model, observed P-values are assumed to come from a mixture distribution.

$$p \sim \pi_0 f_0 + (1 - \pi_0) f_1$$

Here, π_0 is the proportion of P-values corresponding to tests where the null hypothesis is true. The density f_0 represents the density of the observed P-values under the null hypothesis, whereas f_1 denotes the density of P-values under the alternative hypothesis. Since in general each test may have a different alternative distribution, the density f_1 may be seen as a mixture distribution in itself.

Under a correct model, P-values are distributed as $p \sim \mathcal{U}(0, 1)$ under the null hypothesis. At the same time, the distribution of P-values under the alternative is often parametrized using a Beta distribution (Allison et al. 2002).

However, reported P-values in publications usually do not represent the full range of P-values. In particular, most publications only report P-values that are smaller than a significance threshold level, most commonly 0.05 in medical literature. Hence, Jager and Leek 2014 modify the 'two groups' model by conditioning the distribution on the event $p \leq 0.05$. Under these adjustments, the null distribution f_0 still corresponds to a uniform distribution $\mathcal{U}(0, 0.05)$, whereas the conditional alternate distribution is parametrized as a truncated Beta distribution, i.e. a renormalized Beta distribution truncated at 0.05 $p|\{p \leq 0.05\} \sim \frac{1}{F_{a,b}(0.05))}\beta(a, b) = t\beta(a, b; 0.05)$, where a and b are the shape parameters of the Beta distribution. The conditional mixture distribution then models the behaviour of reported P-values when researchers report all P-values less than 0.05 as significant.

$$p \sim \pi_0 \mathcal{U}(0, 0.05) + (1 - \pi_0) t\beta(a, b; 0.05)$$

Now, the fraction π_0 represents the fraction of *reported* P-values that actually come from the null distribution, and hence corresponds to the FDR.

3.2 Left-censoring and rounding

Jager and Leek also propose modeling tools to address certain types of P-value alterations that are common in literature. Specifically, they address the problems of *left-censoring* i.e, small P-values such as 0.0032 may sometimes be reported as $p < 0.01$; as well as P-value *rounding*, i.e. a P-value of 0.012 is often reported as 0.01, leading to spikes of the kind seen in Figure 4

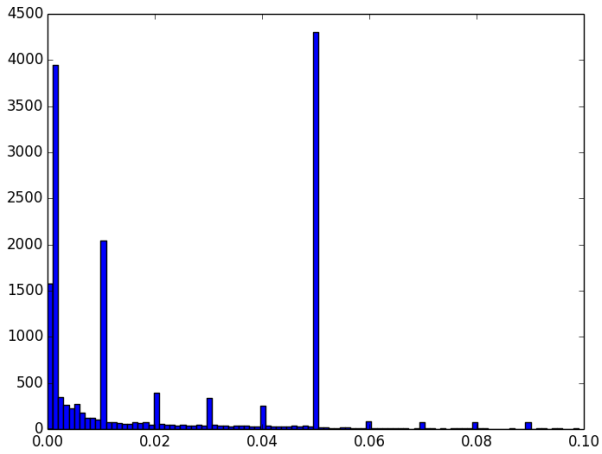


Figure 4: Histogram showing distribution of P-values scraped from a collection of about 3000 PLoS-One articles on cardiology.

Both of these phenomenon are identified using some basic heuristics. P-values reported with $<$ or $\leq$ signs, rather than $=$ in the abstracts are taken to be left-censored; and observations reported as one of the values among 0.01, 0.02, 0.03, 0.04 or 0.05 are taken to be rounded. The problem of left-censoring is treated using standard methods from standard

parametric survival analysis, by treating an observation of the kind $p < p_1$ as a uniform average over all possible P-values values in the range $(0, p_1)$. For the case of rounded observations, the P-values are modeled as multinomial random values, where the probability of each rounded value bin is the total probability of all P-values that would be assigned to that bin by rounding. For example, an observation like $p = 0.02$ is represented in the model as an integration over the assumed distribution of P-values over the range $(0.015, 0.025)$.

3.3 Estimating SWFDR

With the modified 'two groups' model that accounts for certain types of P-value hacking, a simple expectation-maximization (EM) formulation based on latent variables can be used to estimate the parameters of the model. Specifically, the E-step consists of estimating the probability of an observation coming from the null-distribution conditional on the observed data. The M-step consists of maximizing the log-likelihood with respect to the mixture proportion π_0 (which has a simple closed form), and the shape parameters of the truncated Beta distribution a and b (which is done numerically). Since left-censoring and rounding are treated as fully observed, the formulation is not significantly changed by these modifications. As would be expected with such a model, initialization can have a significant effect on convergence of the procedure. The authors also provide bootstrap standard errors for the estimates of $\hat{\pi_0}$, by sampling with replacement sets of observations (each observation is a tuple of a P-value, and its rounding and left-censoring indicators).

The algorithm is then applied to specific observations of P-values from several journals for different years for estimates of journal and year specific false discovery rates; as well as across journals and different years for an estimate of the overall rate of false discoveries in research publications (see Figure 5).

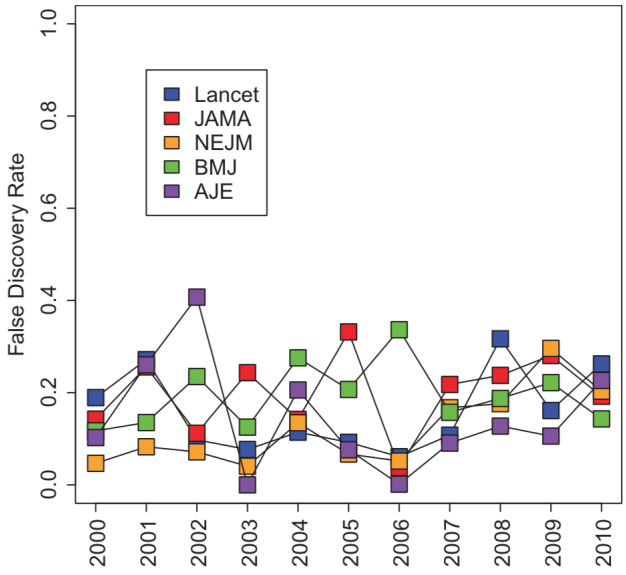


Figure 5: Estimated false discovery rates for the years 2000-2010 by journal.

Of the abstracts from five journals mined for P-values, 5322 reported occurrences of P-values. The overall estimated false discovery rate among all the published results was found to be 14% (with a bootstrap standard deviation of 1%). According to the authors, the rate was also consistent between the various journals; even though the different journals focus on different types of studies (Randomized Controlled Trials, Case controls, Observational studies, etc). Jager and Leek 2014 also performed basic regression analysis to show that the rates of false discoveries are fairly consistent and not significantly correlated with time, or increasing number of submissions in a journal.

3.3.1 Robustness to selection bias and P-value hacking

This papers also includes some synthetic simulations where alternate P-values come from a $\beta(1, 100)$ distribution truncated at 0.05, and null P-values come from a Uniform distribution (hence exactly satisfying the model assumptions). In this case, the estimate FDRs are estimated reasonably accurately. Next, the P-value distribution for the null hypothesis is warped from uniform to the case where only the minimum of 20 uniformly generate P-values is reported. Finally, to simulate a case of systemic P-value hacking; the P-values are always rounded down, rather than the nearest value. In both these cases, the FDR is consistently under-estimated, as see in Figure 6. However, the authors argue that this is less so for small values of FDR; and that both these scenarios probably represent extreme scenarios that are unrealistic. [1]

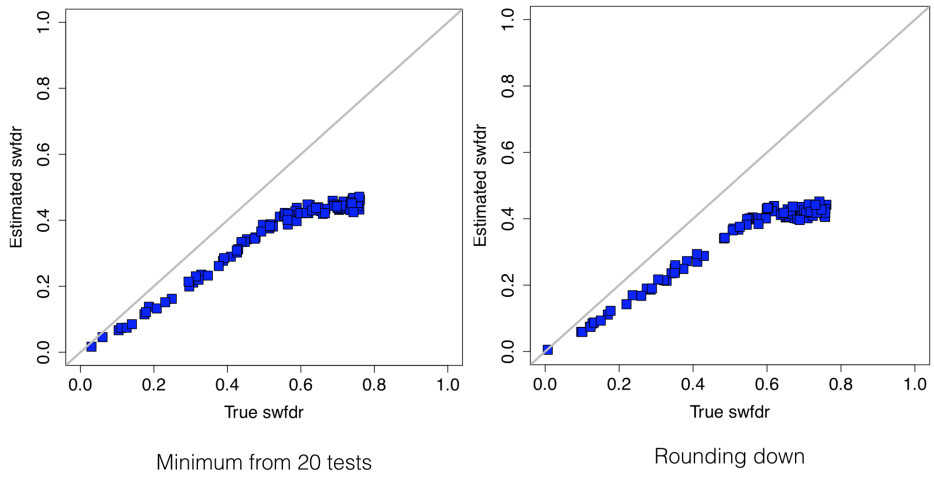


Figure 6: Effects of selection bias and P-value hacking on estimated FDR

Hence, the authors conclude that although FDR rates in published research literature are inflated above the nominal 0.05 level used for significance testing, the inflations are relatively minor (14%). Even the introduction of systemic biases and P-value hacking only increases the estimated FDR only nominally (up to 20%), and hence that empirical evidence suggests that most research findings are true.

[1] Code and data for calculating SWFDR and these simulations are made available by the authors at https://github.com/jtleek/swfdr

3.3.2 Experiments with NIPS data

For exploration, we attempted to get FDR-estimates for a non-medical domain. However, scraping through publications in the NIPS conference over a span of 13 years (2000 to 2012), we could collect only about 100 P-values in the entire texts of papers.[2] This would be comparable to the data for a single year in any of the five journals considered in the original paper(but which considers abstracts only). On this collection, the procedure gives a FDR of about 38%.

3.4 Criticism

The analysis by Jager and Leek is the first to use empirical data to estimate FDR. However, the work has also received some criticism for inappropriate choices of methods and data, ignoring sources of error such as biases in the analysis, model mis-specifications, and implausible model assumptions.

Firstly, the choice of using only P-value data mined from abstracts indicates a potential for **selection bias** since the abstracts typically report some of the most significant (lowest) P-values among possibly numerous experiments performed in the research endeavour. In a small analysis on a sample of the analyzed publications, it was found that the typical paper had about 24 P-values in its entire text, whereas the abstract had only 2. In about 76% of the papers, the best P-value is reported in the abstract (Benjamini and Hechtlinger 2013). The selective incorporation of the best P-values would deflate the empirical estimates of the SWFDR. Also, from the reported numbers, it is possible that the scenario of choosing the best among 20 P-values as discussed in the paper is not so unrealistic, and might be in fact optimistic for several domains.

Secondly, even if the reported P-values are not cherry-picked, and FDR analysis is performed on whole texts of publications rather than just abstracts; the accumulated P-value data might still be biased due to the **file-drawer effect** (scientists are less likely to submit results that do not show strong statistical significance numbers) and **publication bias** (even if submitted, papers with less significant P-values are unlikely to be accepted for publication at premier journals). Such phenomenon would distort the conditional null distribution of P-values from the uniform, and diminish FDR estimates.

Thirdly, it can be argued that the data used in the analysis has significant **sampling bias**. In particular, most of the studies (52%) considered in the analysis are either RCTs or systematic reviews; which are preferred study designs known to have among the lowest false-positive rates. However, the proportion of these research methodologies in actual published literature is much lower. The vast majority of published studies are observational; which are known to be very susceptible to selection biases and confounding variables. Even within a paper abstract, multiple low reported P-values are often highly correlated, and are often minor variations of the same experiment (Ioannidis 2014). Treating these as independent observations would again deflate FDR estimates.

The analysis also seems to exclude a large peak of values exactly at the 0.05 level from its analysis of P-values. Many of these values are preceded by $<$ and $\leq$ operators, and

[2] On NIPS data, the expressions $p <$, $p \leq$ and $p =$, suggested by Jager and Leek, are not very helpful in identifying P-value occurrences. p has an obviously different primary connotation in NIPS (probability), compared with medical literature

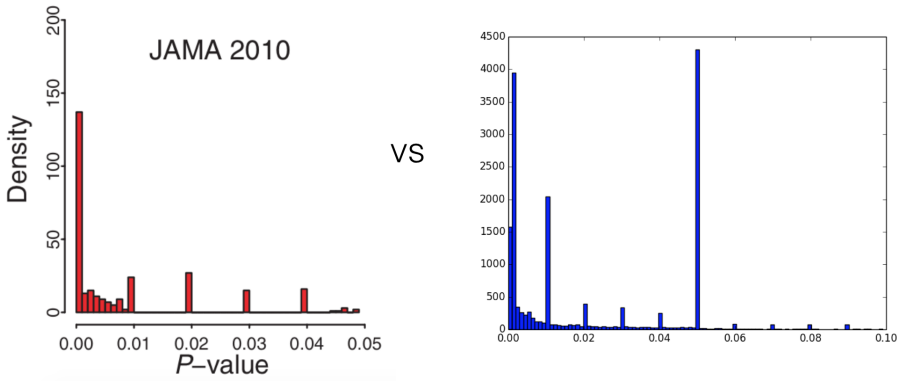


Figure 7: Lager and Leek (2014) exclude peak at the 0.05 level in their P-value analysis

almost all are considered significant by the authors. A histogram of P-value distributions from the paper is compared against an empirical distribution we scraped from a selction of PLoS-One articles in Figure 7. This omission of data seems to show a reasonably clear **exclusion bias**. If the values at the 0.05 level are incorporated in the analyses, the overall FDR as reported by the author jumps to about 21% from 14%.

Next, the assumption of uniform distribution under the null is often unrealistic due to biases, incorrect model specifications and measurement errors. Schuemie et al. 2014 exhibit this by a study analyzing hypothesis that were known to be negative. For three different types of observational studies, Figure 8 plots the distribution of P-values under the null hypothesis. For these distribution, they show that Jager and Leek's FDR estimation procedure drastically under-estimates the true FDR.

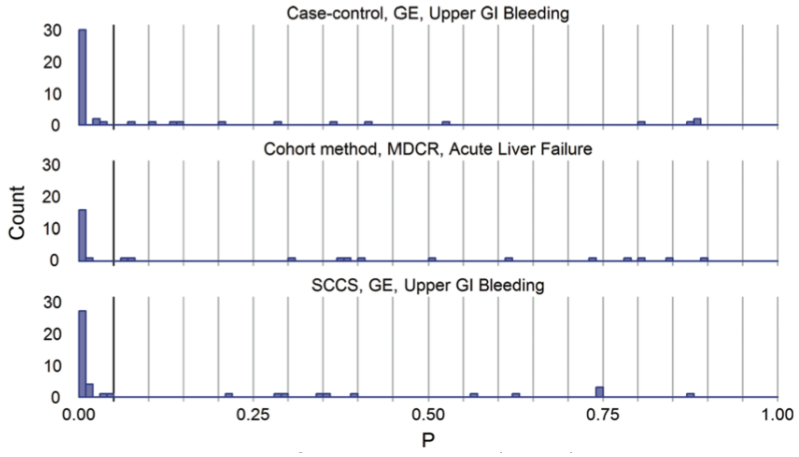


Figure 8: P-value distribution under the null hypothesis for three observational study types in an epidemiological experiment

A major criticism of Jager and Leek's analysis questions the conflation of data from different kinds of study types, as well as different kinds of hypothesis. In a large number of studies, the abstracts don't show any statistical significance for the primary hypothesis

under test, but may quote P-values for ancillary observations and hypothesis, which are often exploratory and might not be pre-registered. Especially from a epidemiological standpoint, the connotations of a FDR value would be very different for primary, secondary and tertiary end-points in an observational study (Goodman 2013). From this perspective, an analyses of FDRs which delineates results according to the hypothesis; and distinguishes between initial studies, replications and meta-analyses, would be more meaningful.

Finally, it is unclear if the choice of estimating FDR as a mixture proportion is appropriate. The iterative EM procedure in this case does not give any guarantees, and convergence would depend on initialization. Many commentators have questioned the choice of importing a procedure from genomics to a large scale study over diverse types of experiments and methodologies. In genomic studies looking for relations, power usually comes from large effect sizes, and a reasonably sharp division between null and non-zero effects (Gelman and O'Rourke 2014). On the other hand, reported P-values in research literature are expected to show a more continuous range of effects, with relatively smaller effect sizes. Also, in genomic studies, the analysis is done on a large data dump, and hence there is no bias due to pre-selection. Conversely, for research literature, selection bias may be a major factor, as discussed earlier.

4 Discussion

In our presentation, we discussed the issue of spurious findings in research, and briefly looked at some of the the issues and factors that are at play in affecting research findings through hypothesis testing. The two main papers presented different claims regarding the extent of false positives in research, but both concur in that the actual occurrence of false positives is much higher than the nominal 5%. While the importance of good methodology (avoiding systemic biases in experiment design, randomization, propensity matching) and analysis (checking code, data, and statistical analysis) cannot be over-stated, a bigger focus for discussion was how to enforce and popularize these practices, since they might sometimes seem to restrict rather than aid research productivity, and provide no tangible benefits.

Replicability of research, and the issue of standardizing good scientific practice require motivating and incentivizing better scientific methodology among the broader research community. One of the ideas suggested in class discussed the possibility of public platforms (such as 'PLoS-Zero') for publishing negative results, and also rewarding and highlighting research papers than conform to high standards of scientific methodology. Such venues can lead to a better awareness of good methodologies in a domain, as well as provide normalizing ideals to conform to. At the same time, publicizing negative results would reduce the frequency of initial false positives. It would also assist research groups in keeping track of people working on similar problems, and would encourage replicability of experiments. Making results in publications truly reproducible can be difficult, but where possible, sharing code and data allows others to check your work and perform various secondary analyses.

Another point of discussion revolved on the role of peers and reviewers in evaluating a claimed scientific discovery. It can be argued that with an adequately stringent peer review, some false initial claims could be more quickly put to rest, hence improving publication quality. In this context, it was mentioned that a significant proportions of very high quality journals provide P-value details that do not in fact match with the corresponding test

statistics. Issues such as these and replicability of experiments (in some fields) could be dealt with effectively if reviewers made in-depth evaluations of research findings. One of the ideas mentioned was that this process could be made more effective if reviewers had financial incentive to critically evaluate each claim. Another suggestion was making reviews of any publication visible to the public. This would be additional incentive to a reviewer to evaluate a submission well.

Finally, part of the discussion also centered on who has the power and responsibility to improve the situation. Obviously education and awareness are part of it, but funding agencies and journals arguably have the most power to influence actual practice. For example, requiring pre-registration of studies for clinical trials could be required by either group, leading to a much better understanding of the full range of clinical outcomes.

References

Allison, David B. et al. (2002). "A mixture model approach for the analysis of microarray gene expression data". In: *Computational Statistics & Data Analysis* 39.1, pp. 1–20.

Benjamini, Yoav and Yotam Hechtlinger (2013). "Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek". In: *Biostatistics*, kxt032.

Chambers, Chris (2014). "Physics envy: Do 'hard' sciences hold the solution to the replication crisis in psychology?" In: *The Guardian*. URL: `http://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology` (visited on 12/12/2014).

Efron, Bradley and Robert Tibshirani (2002). "Empirical Bayes methods and false discovery rates for microarrays". In: *Genetic epidemiology* 23.1, pp. 70–86.

Freedman, David H. (2010). "Lies, Damned Lies, and Medical Science". In: *The Atlantic*. URL: `http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/?single_page=true` (visited on 12/12/2014).

Gelman, Andrew and Keith O'Rourke (2014). "Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values". In: *Biostatistics* 15.1, pp. 18–23.

Goodman, Steven N. (2013). "Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature". In: *Biostatistics*, kxt035.

Ioannidis, John P. A. (2005a). "Contradicted and initially stronger effects in highly cited clinical research". In: *The Journal of the American Medical Association* 294.2, pp. 218–228.

— (2005b). "Why most published research findings are false". In: *PLoS Medicine* 2.8, e124.

— (2014). "Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false". In: *Biostatistics* 15.1, pp. 28–36.

Ioannidis, John P. A. et al. (2001). "Replication validity of genetic association studies". In: *Nature genetics* 29.3, pp. 306–309.

Jager, Leah R. and Jeffery T. Leek (2014). "An estimate of the science-wise false discovery rate and application to the top medical literature". In: *Biostatistics* 15 (1), pp. 1–12.

Potti, Anil et al. (2006). "Genomic signatures to guide the use of chemotherapeutics". In: *Nature Medicine* 12.11, pp. 1294–1300.

Schoenfeld, Jonathan D. and John P. A. Ioannidis (2013). "Is everything we eat associated with cancer? A systematic cookbook review". In: *The American journal of clinical nutrition* 97.1, pp. 127–134.

Schooler, Jonathan W. (2014). "Metascience could rescue the 'replication crisis'". In: *Nature* 515 (7525).

Schuemie, Martijn J. et al. (2014). "Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature". In: *Biostatistics* 15.1, pp. 36–39.

Sebastiani, Paola et al. (2010). "Genetic Signatures of Exceptional Longevity in Humans". In: *Science*.

"Trouble at the lab" (2013). In: *The Economist*. URL: `http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble` (visited on 12/12/2014).

Wakefield, AJ et al. (1998). "RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children". In: *The Lancet* 351.9103, pp. 637–641.