Reproducible Research and Correct Conclusions

Naomi S. Altman[*]

Abstract

The demonstration that even highly influential research may not be reproducible and the reasons for irreproducibility have recently become topics of hot debate. Ioannidis (2005a) used a search of highly cited papers in highly respected medical journals to bring attention to the fact that often the results were contradicted or weakened in subsequent studies. In a provocative follow-up commentary, Ioannidis (2005b) used probabilistic arguments to demonstrate that *most* published research findings are likely to be false. Since then, the issue has been taken up in both the research literature and the popular press shedding much heat and some light. This article continues the discussion, defining positive and negative reproducibility and negative predictive power.

Key Words: p-value; false discovery rate; false non-discovery rate; PPV; positive predictive value; negative predictive value.

1. Introduction

Periodically since at least 1962, studies of the medical and psychometric literature (e.g. Cohen, 1962; Moher et al, 1994; Chung et al, 2002; Breau et al 2006) have suggested that many studies have little power to detect effects of interest due to poor study design and/or low sample size. In modern parlance, these studies suggested an unacceptably high false negative rate (FNR), representing lost opportunities for discoveries and wasted research resources. In 2005, Ioannidis (2005a) challenged the medical research community on another front: studying highly cited articles from high impact journals, he found that many of the results were not upheld in follow-up studies of comparable or higher quality, indicating an unacceptably large false discovery (positive) rate (FDR). His results are summarized in Figure 1. By 2010, Ioannidis' findings had made it into the popular press raising alarm bells about how science is conducted (e.g. Freedman, (2010); *The Economist*, 2013).

A number of reasons for Ioannidis' findings have been proposed. On the purely statistical side, assuming that the prior probability of a non-null result is small, most statistically significant results should be false positives. Even among the true positive results, standard probability arguments imply that when the effect size is small, detected (statistically significant) results are likely to have estimated effects that are larger than the truth, particularly when power of the study is low. Study design also has a large effect on replicability. Ioannidis found that observational studies are less replicable than randomized trials, which may be due to selection bias or uncontrolled confounding variables. Finally, the "winner take all" aspects of scientific endeavor play a role: negative results are seldom published unless they contradict a previously reported high profile positive result; research funds go to the innovators, with few rewards for attempts to replicate previous findings; publication, funding and fame go to first claimant, thus rewarding premature publication of results without adequate follow-up testing. Ioannidis (2005b) makes some of this rigorous by defining the positive predictive value - the probability that a significant result comes from a study with a non-zero effect size - as a function of the prior probability that the effect is non-zero, the level of the test, the power of the study and researcher or block effects.

Several writers have pointed to lack of stringency in statistical testing as a cause for lack of replicability, implicating the "$p < 0.05$" rule-of-thumb as too lax (e.g. Hayden, 2013) or noting that p-values are statistics with inherent variability. Button et al (2013) argue that lack of power is a major cause of false discovery, and that low

[*] Department of Statistics, The Pennsylvania State University, State College, PA, USA

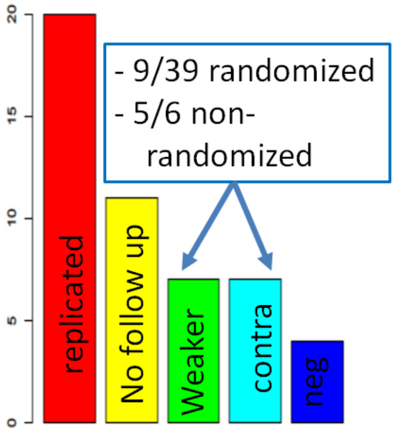

Figure 1: Summary of results of follow-up studies of highly cited papers from Ioannidis ()

powered studies with true effects tend to over-estimate those effects. Note however that the emphasis in the "lack of reproducibility" literature is diametrically opposite to the earlier work on lack of power, focusing on false discovery rather than false nondiscovery.

Whether false nondiscovery or discovery is more important, of course, depends the relative costs of lost opportunities versus following up null effects. The relative frequency of false nondiscoveries and discoveries depends on what percentage of studies have a non-null result. While some of the literature seems to argue that real effects should be rare and this may well be the case for small pilot studies or secondary effects gleaned from larger studies, investigators planning large expensive studies carefully select their primary hypotheses after preliminary studies with the explicit objective of improving the percentage of "real" effects under study. On the other hand, secondary hypotheses are often tested on the same data and these analyses, undoubtedly driven by the data availability and possibly post hoc observation of patterns in the data, likely examine a smaller percentage of "real" effects. As well, driven by publication pressures, dramatic results from small pilot studies sometimes make their way to publication.

This article extends some of the ideas in Ioannidis (2005b) in explaining the probabilistic reasons for lack of reproducibility (assuming honest experiments and reporting). While Ioannidis primarily focused on the reproducibility of statistically significant results, this article discusses both significant and nonsignificant outcomes and their relationship to zero and non-zero effect sizes.

2. Reproducibility

Ioannidis (2005a) discussed two reproducibility issues in follow-up studies: a change from statistical significance to non-significance (or vice versa) and a change of effect size (usually from larger to smaller). The papers selected for analysis were "high impact" so not surprisingly only 4 of the base papers reported negative results (and these were themselves follow-ups to studies which reported significant results). The focus in this article is the first issue: reproducibility and correctness of statistical significance or nonsignificance. One-sided tests are assumed, because a sign change of the effect is an even more severe lack of reproducibility than a change in significance level and no reasonable researcher would consider two studies with the same 2-sided p-value but opposite effect directions to be "reproduced".

We start by considering 2 identical but independent studies. We define a study as *positive reproducible* if the effect was statistically significant and remained significant in the follow-up study. We define a study as *negative reproducible* if the effect was statistically nonsignificant and remained nonsignificant in the follow-up study.

Lets consider the standard case in which we reject at level α and have power $\beta(\alpha, \delta)$ for some effect size δ. If there is really no effect, then we have positive reproducibility with probability α^2 and if the effect size is exactly $\delta > 0$ then we have positive reproducibility with probability $\beta^2(\alpha, \delta)$. Using the often used values of $\alpha = 0.05$ and $\beta(\alpha, \delta) = 0.8$ we see that when the effect is actually 0, we will have positive reproducibility with probability only 0.0025, which is reassuring - we are unlikely to make the same mistake twice. However, when the effect size is actually δ positive reproducibility is only 0.64, which means that we will miss the effect in one or both studies an alarming 36% of the time. Conversely, if there is really no effect, then we have negative reproducibility with probability 0.9 and if the effect size is exactly δ then the negative reproducibility is 0.04.

One of the suggested "fixes" for the lack of reproducibility problem is to use more stringent p-values (e.g. Hayden, 2013). However, in any given study, if a smaller level is used, the power of the test is also reduced. For example, for a t-test, if the original test had $\beta(0.05, 1.308) = 0.8$ with $\alpha = 0.05$ with sample size 8, then reducing to $\alpha = 0.01$ changes the power to $\beta(0.01, 1.308) = 0.47$. This means that our probability of positive reproducibility with true effect size $\delta = 1.308$ is now only 22% while negative reproducibility is now increased to 26% – there is a non-negligible chance that this effect will not be declared statistically significant in either the original study or the follow-up. On the other hand, if we increase the sample size to 14 (the minimum required to achieve power at least .8 with $\alpha = 0.01$) the actual power is 0.83. Then with $\delta = 1.308$, the positive reproducibility is 69% and the negative reproducibility is only 3%. If $\delta = 0$, the positive reproducibility is .01% and the negative reproducibility is 98%. Increased stringency is an excellent idea, but only if it accompanied by an increase of sample size and hence power.

Of course, the analysis above does not address observed reproducibility of the original study, because we do not know the true effect size. To handle this situation, Ioannidis introduced the prior probability that there is a non-zero effect, π_0. In studies in which many hypotheses are tested, such as differential expression studies in biological "omics" π_0 has a frequentist interpretation as the percentage of features with non-null effects in the study and is readily estimated from the distribution of p-values once the study has been completed. In other contexts the Bayesian interpretation of π_0 as the prior probability that the effect currently under study is null may be more reasonable and can be placed in a frequentist context as the percentage of studies with similar prior evidence of a non-zero effect for which the true effect size is actually zero.

Some of the literature on reproducibility suggests that π_0 should be very close to 100%. However, this depends very much on the type of study. For preliminary "blue sky" "let's see what's out there" studies, this is probably a reasonable assumption. However, most federally funded research there is evidence from preliminary studies that guides the research and should lower π_0. The situation is similar to testing for rare diseases in which we know that screening to enrich the percentage of true positives that are tested can be highly effective in reducing the high false positive rate. It is not clear just how high we can expect π_0 to be, but in medical trials of new drugs or devices, it seems reasonable to assume at most equipoise - i.e. $\pi_0 \leq 50\%$. On the other, when secondary analysis are performed on the data, especially those proposed due to "findings" from the same study or due to lack of interesting results from the primary hypothesis, we can expect π_0 to be much higher. This is recognised implicitly in classical multiple testing for analysis of variance where methods such as Scheffé's (1969) procedure "adjust" the p-value for the effects of looking at all possible comparisons of treatments.

In any case, π_0 only tells us the expected percentage of null effects, it does not give us the distribution of effect sizes when the effect is not zero. A fully Bayesian analysis would propose a distribution of effect sizes, which would then lead to a distribution of power for the test (and this would be very suitable for use with "omics" data as well). However, despite the use of the π_0, the intent of this paper is the analysis of frequentist tests. Therefore, we will consider the case when there is some value δ and the effect size is either 0 with probability π_0 or $\delta > 0$ with probability $1 - \pi_0$.

Thus the probability of positive reproducibility $\mathcal{P}_{pos}$ when the identical experiment is done twice is

$$\mathcal{P}_{pos} = \pi_0 \alpha^2 + (1 - \pi_0)\beta^2(\alpha, \delta) \tag{1}$$

and the negative reproducibility $\mathcal{P}_{neg}$ is

$$\mathcal{P}_{neg} = \pi_0(1 - \alpha)^2 + (1 - \pi_0)\left(1 - \beta^2(\alpha, \delta)\right). \tag{2}$$

Of course, it is quite unusual for an experiment to be repeated when the initial result was negative. Hence we might interested instead in the probability of a second rejection following a second identical but independent experiment given that the null hypothesis was rejected in the first experiment. Letting $\mathcal{R}_i$ denote the event that we reject the null on the i^{th} experiment, H_0 be the event that the null is true, and H_A be the event that the null is false. Since δ is fixed, this is this is:

$$\text{Prob}\left(\mathcal{R}_2 | \mathcal{R}_1\right) \quad = \tfrac{\pi_0 \alpha^2 + (1-\pi_0) * \beta^2(\alpha,\delta)}{\pi_0 \alpha + (1-\pi_0) * \beta(\alpha,\delta)} \tag{3}$$

Returning to our previous example in which we know that the effect size is either 0 or a fixed δ giving $\beta(.05, \delta) = 0.8$ and $\pi_0 = 95\%$, then $\mathcal{P}_{pos} = 3.4\%$ while $\mathcal{P}_{neg} = 86\%$, while the probability of a second significant experiment given that the first one was significant is only 39%. This is because over half of the significant tests in the first experiment are false positives. However, if due to preliminary results, $\pi_0 = 50\%$ (equipoise) then positive reproducibility increases to 32% but negative reproducibility is reduced to 47%, while the probability of a second significant test after the first one was significant is 76%.

Enrichment of the truly non-zero effects clearly improves reproducibility, particularly positive reproducibility. Of course a major motivation for preliminary studies is to reduce the waste of effort in chasing zero effects. Other motivations include estimating effect sizes and variability of the responses to design experiments with sufficient power.

For the same effect size, reducing the significance threshold to $P \leq 0.01$ reduces the power to $\beta(0.01, \delta) = 0.47$, which with $\pi_0 = 95\%$ yields $\mathcal{P}_{pos} = 1.3\%$ while $\mathcal{P}_{neg} = 94\%$, while the probability of a second significant experiment given that the first one was significant is reduced to 33.8%. Now only about a quarter of the tests significant in the first experiment are false positives, and both our ability to reject the true positives in the second experiment and to reproduce the discovery is much reduced.

3. Correctness

Of course reproducibility is not the gold standard of good research – it is correctness. Question we should be asking are how likely it is that a significant result actually has effect size $\delta > 0$ and how likely it is that a nonsignificant result has $\delta = 0$? This is the classic "true positive" and "true negative" problem that is usually introduced in elementary probability classes using Bayes' rule to invert probabilities. Letting F_i be the event that we fail to reject the null hypothesis for the i^{th} experiment, it is readily seen that:

$$\text{Prob}\left(\delta > 0 | \mathcal{R}_1\right) = \frac{(1 - \pi_0)\beta(\alpha, \delta)}{\pi_0 \alpha + (1 - \pi_0)\beta(\alpha, \delta)} \tag{4}$$

while

$$\text{Prob}\left(\delta = 0 | F_1\right) = \frac{\pi_0(1 - \alpha)}{\pi_0(1 - \alpha) + (1 - \pi_0)\left(1 - \beta(\alpha, \delta)\right)}. \tag{5}$$

Ioannidis (2005b) calls $\mathrm{Prob}\,(\delta > 0|\mathcal{R}_1)$ the positive predictive value (PPV) of the experiment. (Note that his β is our $1 - \beta(\alpha, \delta)$.) By analogy, we call $Prob(\delta = 0|F_1)$ the negative predictive value (NPV). When $\pi_0 = 0.95$, $\alpha = 0.05$ and $\beta(\alpha, \delta) = 0.8$, $PPV = 0.457$ and $PNV = 0.989$. However, if $\pi_0 = 0.5$ then $PPV = .941$ and $PNV = 0.826$. Clearly the conclusions of significance testing are much more likely to be correct when based on well-supported alternative hypotheses.

Needless to say, independent replication of the experiment provides important information about correctness of the result. Here we only consider the case in which the same result was achieved in all of the n experiments. Then,

$$\mathrm{Prob}\,(\delta > 0|\mathcal{R}_1 \cdots \mathcal{R}_n) = \frac{(1-\pi_0)\beta(\alpha,\delta)^n}{\pi_0\alpha^n + (1-\pi_0)\beta^n(\alpha,\delta)} \tag{6}$$

and

$$\mathrm{Prob}\,(\delta = 0|F_1 \cdots F_n) = \frac{\pi_0(1-\alpha)^n}{\pi_0(1-\alpha)^n + (1-\pi_0)\left(1-\beta^n(\alpha,\delta)\right)}. \tag{7}$$

Going back to our example, if $\pi_0 = .95$ and we reject when $P \leq 0.05$ and we replicate the experiment just twice, the probability that there is actually an effect is 93% when we rejected twice, compared to only 46% for the first experiment only. This comes at a small improvement as well to the true negatives, which are detected 98.8% correctly from one experiment and 99.8% correctly from two. 99.6% of positives are correct for two experiments compared to 94% for one experiment when $\pi_0 = .5$. In this case 83% of negatives from one experiment are correct, but 96% of results that were negative twice are correct.

4. Better Experiments

Better experiments improve the power of the tests. Improvements can be made by better design, which reduces variability and bias, and by increase in sample size.

Figure 2 shows the positive replicability of a study for different sample sizes for $\pi_0 = 0.95$ and 0.5 and for $\alpha = 0.05$ and 0.01. Using a more stringent p-value scarcely changes positive replicability when π_0 is large. Preliminary studies which reduce π_0 to 0.5 effectively improve positive replicability, especially for larger sample sizes and more stringent tests.

Figure 3 shows the probability that the effect is truly non-zero following a test with $P < \alpha$ for $\pi_0 = 0.95$ and 0.5 and for $\alpha = 0.05$ and 0.01. The efficacy of increasing the sample size is very sharp for small low-power studies, but asymptotes as the power approaches 100%.

Suppose instead of doing 2 independent experiments we double our sample size. Note that when doing 2 experiments, we have to reject twice to declare statistical significance, so our power is $\beta^2(\alpha, \delta)$. On the other hand, to have a false discovery, we need to reject twice when the null hypothesis is true, so that we expect false discoveries only α^2 of the time rather than α. The power of a single test rejecting at $P \leq 0.0025$ and sample size 16 is $\beta(.0025, 1.308) = .74$ compared to the power of $0.8^2 = 0.64$ from rejecting when both experiments with sample size 8 have $P < 0.05$. In theory, at least, it is better to increase the sample size than to do independent replicates of a small experiment.

In practice, when an independent replicate of an experiment is performed (by a different investigator in a different lab) there is almost always some type of "block effect" which is the intra-class correlation of observations within the replicate. What this means is that relative to observations from different replicates of the experiment, the observations within a replicate *are not* independent. The observations have replicable (but unknown) latent factors due to instrumentation, laboratory protocols and the skills of the experimenter which induce biases that reduce the observed estimate of variability. The smaller the intraclass correlation, the closer we are to the ideal situation discussed in the first paragraph. The larger the intraclass correlation, smaller the effective sample size achieved by

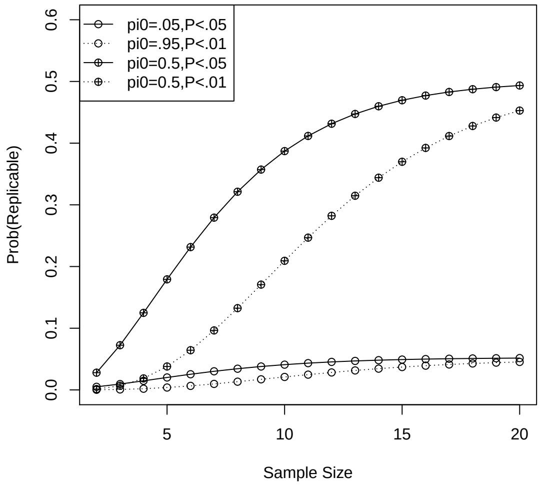

Figure 2: Positive replicability of a two-sample t-test for effect size 1.308 at different values of π_0 and rejection level α.

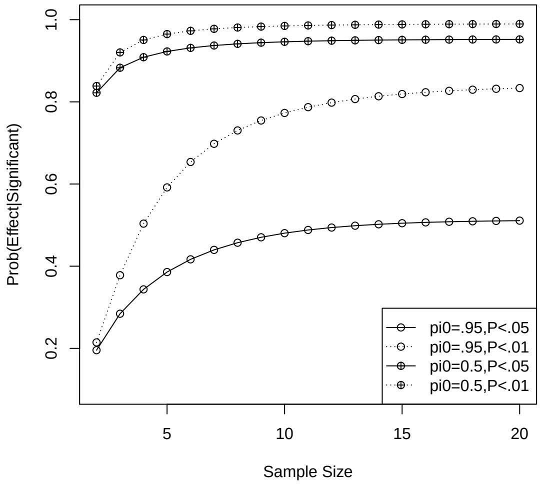

Figure 3: Correctness of a two-sample t-test for effect size 1.308 at different values of π_0 and rejection level α.

increasing the sample size, and hence the more modest the gain in power. What this means in practice is that while each investigator improves his/her probability of producing reproducible and correct effects by increasing sample size, any estimate of reproducibility and correctness is almost certainly optimistic. Ioannidis (2005b) refers to this as the study bias.

With fixed resources for performing experiments, the researcher seeking to do replicable research leading to correct results is therefore faced with some hard choices. On one hand, the researcher may increase the sample size to allow for high power even when rejecting at a very stringent p-value. On the other hand, the researcher can seek collaboration with an independent lab which can attempt to replicate interesting findings. Finally, the researcher may seek to verify the results using follow-up experiments which are expected to have different biases and which will succeed only if the initial results are correct. Of these options, the first option leads to the most optimistic estimate of reproducibility and correctness, while the last is the most risky.

5. Discussion

The recent focus on reproducible research has high-lighted some of the failings of the modern scientific process, especially in some of the high-stakes, high variance research disciplines such as the life and medical sciences, biology and psychology. A better understanding of the roles of sample size, independent replication, and preliminary studies should help design studies that waste fewer resources while improving our ability to reach true conclusions.

The use of formal meta-analysis to combine information across studies that are independent but not identical has proven valuable both in estimating effect sizes when multiple studies are available and in shedding light on the biases induced by the search for significance (e.g. Francis, 2013).

Other initiatives include registering research hypotheses in advance, to make it clearer when published results come from primary rather than secondary analyses, repositories for primary data (which are now standard in "omics" studies funded by U.S. federal agencies) and publication of data analysis pipelines as supplemental documents to papers.

While "reproducibility" of the research has become the catchword of the day, researchers and statisticians understand that correctness of the conclusions is the gold standard to which we aspire. Attention to good experimental design, good statistical analysis, accurate records of *all* the analyses that were done and *all* the hypotheses that were tested should help us towards this goal.

Acknowledgements

Naomi Altman acknowledges partial support from NSF DMS 1007801, NSF IOS 0820729 and NIH UL1RR033184. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

REFERENCES

Breau, R., Carnat, T., and Gaboury, I. (2006). Inadequate statistical power of negative clinical trials in urological literature. *The Journal of Urology*, **176**, 263–266.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafó, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**, 365–376.

Cohen, J. (1962). The statistical power of abnormal-social psychological-research - a review. *Journal of Abnormal Psychology*, **65**, 145–153

Chung, K., Kalliainen, L., Spilson, S., Walters, M., and Kim, H. (2002). The prevalence of negative studies with inadequate statistical power: an analysis of the plastic surgery literature. *Plastic Reconstructive Surgery*, **109**, 1–6.

Economist (2013). Trouble at the lab. *The Economist*, page 199.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, **57**, 153–169.

Freedman, D. (2010). Lies, damned lies, and medical science. *Atlantic Magazine*.

Hayden, E. (2013). Weak statistical standards implicated in scientific irreproducibility. *Nature*.

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

Ioannidis, J. A. (2005b). Why most published research findings are false. *PLoS Med*, **2**, e124.

Moher, D., Dulberg, C., and G.A., W. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, **272**, 122–124.

Scheffé, H. (1953); (1969). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104; Corrigenda. *Biometrika*, **56**, 229.

Wacholder S, Chanock S, Garcia-Closas M, Elghormli L, Rothman N (2004) Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst*, **96**, 434–442.