Difficulties in making inferences about scientific truth from distributions of published p-values

Andrew Gelman and Keith O'Rourke*

1 Mar 2013

1. Background

There has been much unease in recent years about our current default system of evaluating and reporting experiments and observational studies. The profusion of dubious and unreplicated claims in subfields ranging from social psychology to brain imaging to medicine has led many observers including ourselves to feel that the scientific publication system is failing. Four flashpoints of this ongoing conversation have been:

- John Ioannidis's 2005 article on most published findings being false, which resonated both with researchers within applied fields and with outsiders who felt whipsawed by various medical research claims that hit the newspapers but rarely seemed to pan out;
- Daryl Bem's much-publicized paper about ESP, which was published in a leading psychology journal despite serious methodological flaws and a highly implausible scientific framework (Bem, 2011, Yarkoni, 2011, Francis, 2012). This case became even more notorious when a failed replication of Bem's study was denied publication in that same journal on the grounds that the journal "does not publish replications" (French, 2012).
- Recent scandals about faked research by leading psychologists Mark Hauser and Diederik Stapel, along with reporting such as that on the blog *Retraction Watch* of low quality control and defensive, non-scientific attitudes among editors and publishers of scientific journals (Marcus and Oransky, 2010–2013).
- The 2009 Duke clinical trial scandal where a critical data processing error (a row slip) was identified by an outside group at MD Anderson and when the Duke clinical trialists were informed of it, they did their best to dismiss, deny and then obscure it rather than correct it. Following up, the MD Anderson group realized similar errors were all too prevalent in their own research institute and have since taken measures to help prevent and detect them (Baggerly and Coombes, 2009, 2012).

In the article under discussion, Jager and Leek present a bold idea to measure the "science-wise false discovery rate" by taking the distribution of published p-values and then applying an estimation procedure used in genetic linkage studies to estimate the proportion of p-values coming from null and alternative distributions. We admire the authors' energy and creativity, but we find their claims unbelievable (as well as implausibly precise, as for example their estimate with a margin of error of 1%) and based on unreasonable assumptions.

Jager and Leek may well be correct in their larger point, that the medical literature is broadly correct. To answer such a question would require additional care in defining what it means for a study to be correct. The research paradigm of effects being either zero or nonzero is not, we believe, particularly helpful, for two reasons. First, we almost always care about the direction of an effect,

*Discussion of the paper, "Empirical estimates suggest most published research is true," by Leah Jager and Jeffrey Leek, for *Biostatistics*. We thank Anastasios Tsiatis and Geert Molenberghs for arranging this discussion and the Institute of Education Sciences for partial support of this work.

not merely its existence. Second, the magnitude of any effects or comparisons are also important and, in fact, connect directly to concerns about replicability of scientific phenomena.

Medical researchers are mostly studying real effects (setting aside certain wacky examples and desperate clinical research areas involving high mortalities). But there's a lot of variation. A new treatment will help in some cases and hurt in others. Also, studies are not perfect, there are various sorts of measurement error and selection bias that creep in, hence even the occasionally truly zero effect will not be zero in statistical expectation (i.e., with a large enough study, effects will be found). For these reasons, we are not so concerned with Type 1 or Type 2 errors in the classical sense (and as considered by Jager and Leek), rather we worry about Type S errors (someone says an effect is positive when it's actually negative) and Type M errors (someone says an effect is large when it's actually small, or vice versa).

For example, the notorious study of beauty and sex ratios described by Gelman and Weakliem (2009) had a Type M error: the claim was an 8 percentage point difference in the probability of a girl (comparing the children of beautiful and non-beautiful parents), but we are sure any actual difference is 0.3 percentage points or less, it could go in either direction, and there is no reason to suppose it persists over time. The point in that example is not that the true effect is or is not zero (thus making the original claim "false" or "true") but rather that the study is noninformative. If it got the sign right it's by luck, and in any case it's overestimating the magnitude of any difference by more than an order of magnitude.

Our point about Type 1 errors is not primarily "semantics" or "philosophy." The framework of the Jager and Leek paper under discussion is admirably clear—our problem is that we don't think it applies well to reality. We have a problem with the identification of *scientific* hypotheses as *statistical* "hypotheses" of the "$\theta = 0$" variety. We understand that the authors chose to follow the model used by the much-cited Ioannidis (2005), but that does not recuse them from dealing with the logical difficulties involved with that model.

Distributions of p-values of null effects in different application domains

Jager and Leek's key assumption is that p-values of null effects will follow a uniform distribution. We argue that this will be the case under only very limited settings and thus they are mistaken to use their analysis to make claims about a "science-wise false discovery rate" or even more modest claims regarding the medical literature, given the presence of selection (both of what analyses to perform and what to publish) and all the mishaps and measurement error of all kinds that can occur (see Greenland, 2005).

Jager and Leek write that their model is commonly used to study hypotheses in genetics and imaging. We could see how this model could make sense in those fields: First, at least in genetics we could imagine a very sharp division between a small number of nonzero effects and a large number of effects that are essentially null. Second, in these fields, a researcher is analyzing a big data dump and gets to see all the estimates and all the p-values at once, so at that initial stage there is no selection. But we don't see this model applying to published medical research, for two reasons. First, as noted above, we expect to see a continuous range of effects rather than a sharp division between null and non-null effects; and, second, there is just too much selection going on for us to believe that the conditional distributions of the p-values would be anything like the theoretical distributions suggested by Neyman-Pearson theory.

Figure 1 of the paper under discussion illustrates the central problem, which is the assumption that the 990 false hypotheses yield only 50 statistically significant p-values. We believe that scientists are ingenious enough, and do experiments in the presence of enough measurement error (a term which here we are using to include all non-sampling error including problems of reliability

and validity, noncompliance, missing data, and so on), that they would have *no problem* getting statistical significance at much more than a 5% rate even in the presence of null effects. We are not talking about cheating, just about good scientific practice, looking for interesting patterns and when using standard classical statistical techniques in which prior information is not used to regularize-toward-zero the estimates of unanticipated new findings.

Technically, even if the the null hypothesis is exactly true, a uniform distribution of p-values does not in general follow. The uniform distribution is an ideal target that is achieved only for continuous outcomes under assumptions of exponential and location-scale models for certain parameterizations of treatment effects, and perhaps occasionally by using higher order asymptotic approximations or just closely with permutation/randomization test methods. The uniform distribution will *not* be achieved for discrete outcomes (without the addition of subsequent random noise), or for instance when a t.test is performed using the default in the R software with small sample sizes (unequal variances). These deficiencies are easily identified by simple simulations—we would recommend the two-group binary outcome randomized clinical trial as a dramatic case study. Though one might argue or hope that the distributions are close enough to uniform for this not to matter, the onus is on authors to establish this before recommending the methods and making empirical claims.

Even aside from selection, mishaps and measurement error, we see problems with the assumed model when is applied outside very controlled settings. Even for continuous outcomes with convenient distributional assumption, the uniform distribution does not happen in empirical research simply by making assumptions. Researchers have to make it happen and the only widely accepted method for doing that is randomization. A considerable amount of work is required when there was not randomization to even get close and the assumptions are largely untestable (Rosenbaum, 2010). Perhaps more importantly, these advanced methods are not commonly used (yet) in most clinical research. Additionally, even with randomization, inevitable clinical trial limitations such as unblinding, missing data, non-compliance, withdrawal, etc. can considerably affect the distribution of p-values in ways that cannot be remedied.

We recognize that the above criticisms apply to much of our applied research as well; that our own confidence intervals, p-values, and posterior distributions are only approximate and can depend highly on key assumptions which must be strongly questioned. It is because of these concerns, which we see even in our own work (considering that we rarely if ever analyze clinical trials with continuous measurements, known distributions, and prechosen outcomes) that we are skeptical of leveraging an assumed uniform distribution of p-values, or anything close to it, to make broader claims about the validity of scientific reports.

More pragmatically, the authors' model applies to a single well-controlled study, with a single protocol managed by one group with control over the data and all of its analysis. But the population of p-values the authors need to represent is generated by a myriad of studies, with differing protocols over many groups with many interests and intentions. With enough unmeasured confounding in a non-randomized clinical trial, things may be far worse than in the authors' sensitivity analyses.

Also of great concern, the paper under discussion uses a sample of p-values from a variety of sources. Some sense of the variety can be gained from the random set of ten abstracts they present in their appendix, in which 3 studies are randomized and 7 studies are not. The sampled p-values vary from being for primary comparisons, secondary comparisons, for covariate imbalance, etc. In the full sample we would expect p-values for meta-analysis, checks of patient blinding, checks of data entry error rates, and almost anything else someone might have thought to test and report in the abstract.

What can be construed from such a mixture? Recall that there are different issues and motivations for each sort of test. For instance, secondary comparisons are often exploratory and multiplicity not accounted for and moderate false positive rates reasonable or in some sense op-

timal in the long run. With checking for covariate balance, in a randomized trial it is expected (hoped) that there is balance. Here, if investigators have successfully randomized, the null should almost always be true and a high false positive rate good! In observational studies, covariates are widely suspected of being unbalanced and the alternative should usually be true. Here a low false positive rate would at most be reassuring and just of that suspicion of imbalance.

One of us worked for many years across multiple research groups at a large medical school and was occasionally asked to evaluate the work of other research groups, with access to all data and staff. It was alarming how often mistakes occurred and went undetected. One notable arose when staff at a teaching hospital evaluated the predictive value of some new measure to predict hospital mortality and it was dramatically successful. This made some members of senior management very worried so they asked us to verify the findings. After a few difficult meetings the researchers agreed to do double data entry, and it was discovered that a row slip had occurred and it was the next admitted patient's measure that predicted the last admitted's mortality. This has to be rare, but what isn't is some error that leads to a publishable finding that all the research team pounces upon and spends all its energies trying to get into the highest prestige journal possible, as soon as possible. Any suggestion for careful checking is unlikely to be welcome. The 2009 Duke clinical trail scandal sounds similar to this and has been thoroughly and publicly documented. The glimmer of hope from that scandal is that the group from MD Anderson that identified the error realized similar errors were happening at their own research institute (row slips, coding errors, etc.) and have taken steps to help rectify it. Lets hope many more research institutes will take similar steps, but that won't help immediately with past published research.

There is also empirical research of how often statistical errors occur in publications of clinical research, though without providing a sense of exactly how important. Garcia and Alcaraz (2004). searched through given volumes of Nature and NEJM and noted 10% of reported p-values did not correspond with the statistics they were purportedly for. That would have a big impact on the distribution of p-values (in addition to all the problems listed above which would occur even in the absence of this sort of mistake). Strasak et al. (2007) reviewed statistical errors in NEJM and presented a detailed table. Readers may disagree on whether all should be considered errors and how much of impact these would have on the distribution of p-values, but there are many that surely would.

Summary

Jager and Leek describe their work as an "empirical analysis of the rate of false discoveries in the major medical journals." To us, an empirical estimate would involve looking at some number of papers with p-values and then follow up and see if the claims were replicated. But this is not what was done. We are concerned that if this paper is taken at face value, that clinical research managers might take this claim as providing evidence that resources currently directed or planned for quality control and increased care in clinical research are not needed and can be spent better elsewhere.

We think what Jager and Leek are trying to do is hopeless, at least if applied outside a very narrow range of randomized clinical trials with pre-chosen endpoints. It's not a matter of a tweak to the model here or there; we just don't think it's possible to analyze a collection of published p-values and, from that alone, infer anything interesting about the distribution of true effects. One might say that the same objection would hold for any meta-analysis, but this case seems more problematic to us here. The approach is just too driven by assumptions that are not even close to plausible and a catch-all sample of $p$values that would not be representative of any conceivable population of interest. The authors claim that "we are able to empirically estimate the rate of false positives in the medical literature and trends in false positive rates over time," but this is done by

assuming the model that is being questioned, in which null p-values have uniform distributions and taking a strange sample as estimating a ill defined rate. Given the huge problems with this model, we don't see the resulting estimates just can't be "empirical" in any real sense of the word.

This technique might be effective for single studies such as in genetic linkage or perhaps carefully designed randomized clinical trials but not for collections of general published studies.

That said, we admire the authors' boldness in advancing their approach. Various scholars have been performing analysis of published p-values as a way to study the scientific process (Simmons, Nelson, and Simonsohn, 2011, Francis, 2013). Pulling in methods from related areas such as statistical genetics, following ideas of Efron and Tibshirani (2002), seems like a promising idea. Given clarification and further development we can anticipate that the ideas being explored by Jager and Leek will have many useful applications if applied carefully.

References

Baggerly, K. A., and Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics* **3**, 1309–1334.

Baggerly, K. A., and Coombes, K. R. (2012). A saga starter set. `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/StarterSet/index.html`

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.

Francis, G. (2013). Replication, statistical consistency, and publication bias (with discussion). *Journal of Mathematical Psychology*, to appear.

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

Gelman, A. (2013). P-values and statistical practice. *Epidemiology* **24**, 69–72.

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

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`

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

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

Greenland, S., and Poole, C. (2013). Living with P-values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* **24**, 62–68.

Rosenbaum, P. R. (2010). *Observational Studies*, second 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 Jan. 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/