Biostatistics (2013), *pp.* 1–6
doi:10.1093/biostatistics/kxt038

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

LEAH R. JAGER, JEFFREY T. LEEK*

Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health,
Baltimore, MD 21205, USA

jleek@jhsph.edu

1. Introduction

We would like to thank the discussants for their insight and efforts in producing this entertaining and thoughtful discussion. We believe that they have raised interesting points regarding our analysis and have made valuable contributions that suggest ways to build on our original idea. Our response focuses on four main points: (1) incorporating data into the discussion, (2) the importance of reproducible research, (3) a new statistical formulation for the science-wise false discovery rate (SWFDR), and (4) are most published medical results "true"?

2. Incorporating data into the discussion

One of our primary goals in this manuscript was to identify, collect, and publish a new source of data for evaluating the rate of false discoveries in the medical or scientific literature. Given the stakes—both economic and professional—involved in high-profile medical and scientific research, it is not surprising that there are strong opinions about the actual rate of false discoveries among published results. The interest in this question has led to significant discussion in both the scientific literature (Ioannidis, 2005; Goodman and Greenland, 2007; Pfeiffer and Hoffmann, 2009) and the popular press (Freedman, 2010; Staff, 2007; Hotz, 2007). Unfortunately, it is difficult to collect data to directly estimate the rate of false discoveries for at least two reasons.

First, *replicating scientific results is time consuming, expensive, and often impossible*. Scientific studies, particularly in the medical literature, are expensive due to recruitment, treatment, and measurement costs. Limited resources and the high cost ensure that only the most scientifically and medically significant results will be subjected to a true replication—two or more studies with identical population, parameter of interest, and study design. Even replication studies may use different study populations or technologies, making them difficult to compare. As Cox (2013) points out in his discussion, there have been relatively few published true replications before whole genome association studies, likely due to the cost/benefit tradeoff of performing clinical studies (Fletcher, 1989).

Second, *replication is only a surrogate for truth*. Gelman and O'Rourke (2013) and Ioannidis (2013) point out the failed replication of individual or multiple studies as evidence for false positives in the

*To whom correspondence should be addressed.

scientific literature. In addition to the economic challenges, replication poses statistical challenges that we think deserve further investigation. For example, a result may be either a true or false discovery in the original study being performed. In the replication study, the result may replicate or not. If it replicates, we do not know if both results are true discoveries or false discoveries. If it does not replicate, we are not absolutely certain whether the first result was a true discovery and the second a false non-discovery or if the original result was a false discovery. Successful replication is thus only a surrogate endpoint for the underlying veracity of hypotheses being considered.

In light of the aforementioned difficulties, it is not surprising that early efforts to discuss the SWFDR focused on theoretical arguments (Ioannidis, 2005; Goodman and Greenland, 2007) or anecdotal evidence (Ioannidis *and others*, 2001; Gelman and O'Rourke, 2013). Important efforts to calculate more comprehensive estimates of the SWFDR in specific cases have, by necessity, relied on noisy high-throughput measurements to stand in for a "gold standard" of truth (Pfeiffer and Hoffmann, 2009). We therefore sought to collect a set of data that could be used to address the question in a more comprehensive and empirical fashion. Our approach was based on a relatively crude text mining algorithm for extracting p-values from the abstracts of published papers in the medical literature. As the discussants have pointed out, our initial sample could be improved by:

(1) Making it more comprehensive and considering a larger variety of journals (Ioannidis, 2013; Goodman, 2013; Benjamini and Hechtlinger, 2013).
(2) Focusing on specific subtypes of endpoints (e.g. primary versus secondary) (Ioannidis, 2013; Gelman and O'Rourke, 2013; Goodman, 2013; Benjamini and Hechtlinger, 2013).
(3) Focusing on specific study types, such as randomized trials or epidemiological studies (Ioannidis, 2013; Gelman and O'Rourke, 2013; Goodman, 2013).
(4) Handling rounding more carefully for all possible rounded values (Benjamini and Hechtlinger, 2013).
(5) Collecting p-values from the full text rather than abstracts (Ioannidis, 2013; Benjamini and Hechtlinger, 2013; Gelman and O'Rourke, 2013).
(6) Making different decisions about the threshold on the p-values (Cox, 2013; Benjamini and Hechtlinger, 2013).

Ioannidis (2013), Benjamini and Hechtlinger (2013), and Schuemie *and others* (2013) point out data-backed sensitivities of our results to a variety of choices we made during the data analysis process. While we attempted to evaluate the potential impact of certain types of p-value manipulation in our paper, we agree that addressing the selection into the abstract effect and different choices of threshold for our original collection of p-values will ultimately lead to different estimates of the SWFDR.

We are encouraged, however, that several of the discussants collected additional data to evaluate the impact of the above decisions on the SWFDR estimates. The discussion illustrates the powerful way that data collection can be used to move the theoretical and philosophical discussion on to a more concrete, scientific footing—discussing the specific strengths and weaknesses of a particular empirical approach. Moreover, the interesting additional data collected by the discussants on study types, journals, and endpoints demonstrate that data beget data and lead to a stronger and more directed conversation.

3. THE IMPORTANCE OF REPRODUCIBLE RESEARCH

Reproducible research is a burgeoning field (Peng, 2009, 2011) that has grown in importance with the discovery of statistical coding errors in recent high-profile medical (Baggerly and Coombes, 2009) and economic (Herndon *and others*, 2013) studies. Access to a study's raw data is one component of reproducible research, but another equally critical component is access to the code and software used to perform

analyses (Peng, 2011). As part of our original submission, we included all of the R code we used to collect the *p*-value data and perform the statistical modeling. We also included a version of the data collected using our code.

Our code and data were used by Ioannidis (2013), Benjamini and Hechtlinger (2013), and Schuemie *and others* (2013) to evaluate the impact of different modeling choices on our estimates of the SWFDR. The benefit of reproducible research is that the discussion of our results could begin where our analysis ended. The discussions that focused on additional quantitative analysis using our code raised particularly useful points about the threshold we chose for selecting *p*-values, the types of studies performed, and the bias in the selection to the abstract effect. Without our code and data, it would not have been possible for the discussants to make important contributions to our analysis so quickly; it might have been months or years before this happened. The details of our code also ensured that the discussants did not need to retread the analyses we had performed since the exact procedure used was available to them immediately.

Our code was distributed through the Github (https://github.com/jtleek/swfdr) platform that provides a web interface for the git version control system. By using web-based version control, we were able to correct a minor typo identified by a discussant and push the changes to the repository quickly (Ioannidis, 2013). This type of responsive and rapid error correction would not have been possible with static code directories published through journal interfaces and suggests the power of public code/data repositories linked to powerful version control tools for empirical data sciences. Equally critical is making the code/data available in a format that is easily accessible and remixable to create new analyses.

We believe this process also illustrates a key point in empirical research—if we expect authors to produce code and data to support their conclusions we should expect that minor corrections be necessary and embrace them. Otherwise, we may succumb to the temptation to throw out very good research in search of perfection.

4. Statistical framework for the SWFDR

In addition to making a new data resource available, we also proposed a meta-analytic framework for estimating the SWFDR. We focused on *p*-values and hypothesis testing for two reasons: (1) hypothesis testing remains the most widely used approach for statistically screening results in the medical and epidemiological literature and (2) the most widely read analysis of the SWFDR focused on the hypothesis testing framework (Ioannidis, 2005). As has often been pointed out by the discussants (Cox, 2013; Gelman and O'Rourke, 2013) and in the statistical literature (Berger and Sellke, 1987), hypothesis testing has known limitations as a method for hypothesis screening. While that discussion is beyond the scope of this work, we point out that our choice to use *p*-values was not based on an explicit endorsement of hypothesis testing, but rather the ubiquity of *p*-values as a data source that could be used to estimate the SWFDR.

Our estimation approach is based on a similar framework for estimating the FDR in genomic experiments (Efron and Tibshirani, 2002; Storey and Tibshirani, 2003), with some adaptations to the sampling problems peculiar to meta-analysis. In some ways, this analogy is apt: there are a large number of hypotheses being tested and the quantity of interest is the rate at which discoveries are false—where the null hypothesis is true but the result is declared significant.

In other ways, the analogy is less direct. The primary point raised by multiple discussants is the structural difference in the way *p*-values are calculated in genomic experiments versus in meta-analysis of scientific results. Gelman and O'Rourke (2013), Benjamini and Hechtlinger (2013), Ioannidis (2013), and Schuemie *and others* (2013) point out that while the theoretical definition requires that *p*-values be uniform under the null hypothesis, this may not be the case for *p*-values collected from abstracts. The reasons for this deviation may be the selection into the abstract effect (Benjamini and Hechtlinger, 2013; Ioannidis, 2013; Goodman, 2013), unmodeled confounders in the regression analysis (Schuemie *and others*, 2013; Ioannidis, 2013), publication bias (Goodman, 2013; Ioannidis, 2013), researcher degrees of freedom

(Gelman and O'Rourke, 2013; Ioannidis, 2013), p-value hacking (Gelman and O'Rourke, 2013), and fraud (Goodman, 2013). In some cases, the discussants have included empirical data supporting these deviations and in others they have posited theoretical ways in which the null distribution may deviate from the uniform distribution.

While we concur with the discussants that there are multiform ways in which the null distribution may be non-uniform, we are happy to see the discussion centered on a specific statistical framework for modeling false positives quantitatively. Proposition 1 in supplementary material available at *Biostatistics* online of our paper addresses some of these concerns by demonstrating that the null behavior of the p-values is uniform if the decision to report the p-value is based only on whether it meets the threshold of $p < 0.05$ (or more generally α). As pointed out by Goodman (2013), this thresholding was actually the basis for the original discussion of the rate of false discoveries in the medical literature, although, as Cox (2013) points out, a helpful avenue for expansion would be a complete decision-theoretic framework for estimating the SWFDR that does not rely on a specific threshold. While we concur that such an analysis would potentially be interesting, the difficulties raised by selection bias are compounded when the full distribution of p-values is considered.

We also made efforts to evaluate the robustness of our proposed SWFDR estimates when the null distribution was incorrect but the alternative distribution was held fixed. In the case of p-value hacking, we show that unless widespread hacking is being performed on the null hypotheses while the alternative distribution is unchanged, our estimates will be robust to minor deviations from the expected null distribution. Some discussants have suggested that this type of hacking is widespread either for fraudulent (Gelman and O'Rourke, 2013; Ioannidis, 2013; Goodman, 2013) or purely accidental (Schuemie *and others*, 2013) reasons. We admire the efforts of the discussants to quantify the potential magnitude of the bias due to non-uniformity of p-values. However, while the potential for bias is clearly demonstrated, the extent of the bias has yet to be quantified in real studies. We are hopeful that our framework could be used to perform sensitivity analyses when considering policies on screening results for biomedical or scientific journals.

Of course, as the discussants point out, there are some forms of bias that are not covered by Proposition 1 or our sensitivity analysis. Some cases, such as the rounding of p-values at finer scales (Benjamini and Hechtlinger, 2013), could be directly addressed by minor modification of the SWFDR estimator we have proposed. Other cases, such as the selection-into-the-abstract bias (Benjamini and Hechtlinger, 2013; Goodman, 2013; Ioannidis, 2013; Gelman and O'Rourke, 2013), could also be addressed with straightforward collection of more data and minor modifications of our framework to correct potential biases.

Issues such as missing multiplicity adjustments (Benjamini and Hechtlinger, 2013; Ioannidis, 2013) or comparisons of primary versus secondary endpoints (Ioannidis, 2013; Gelman and O'Rourke, 2013; Goodman, 2013) could be ameliorated by a more complete collection of data from the papers in question. A complete extraction of all potential confounders, types of analyses, and numbers of comparisons performed could be used to calculate a conditional estimate of the SWFDR given specific covariates. While this data collection was beyond the scope of our original manuscript, the availability of the raw data in Pubmed, and the power of natural language processing tools do not rule out the possibility of more thorough data collection.

Other cases, such as outright fraud (Goodman, 2013; Gelman and O'Rourke, 2013), subtle latent confounding in the calculation of statistical significance (Ioannidis, 2013; Schuemie *and others*, 2013), or unspecified bias in the scientific process (Ioannidis, 2013) are less straightforward to address by data collection and conditional estimation. Even in these challenging scenarios, we hope that our statistical modeling framework could be used to perform sensitivity analyses to measure the potential range of values for the SWFDR. Schuemie *and others* (2013) have shown one example of this strategy by sampling from biased empirical p-values for the null hypothesis.

5. Is most published research false?

Our efforts have focused on collecting quantitative data, establishing a statistical framework, and moving the discussion about the SWFDR to more empirical ground. We were happy to see that the discussion of our work has largely moved in this direction. But ultimately the question remains: what is the rate that reported scientific discoveries are false? We reported an estimate of 14% in our original manuscript. As the discussants have pointed out through the use of our data and additional data they have collected, this number may be optimistic for several reasons. One discussant suggests that bounds on the SWFDR may range between 20% and 50% (Benjamini and Hechtlinger, 2013), while others suggest more extreme values (Ioannidis, 2013) or are not comfortable making an estimate (Gelman and O'Rourke, 2013).

In light of the additional data and analyses collected by the discussants, we agree with (Benjamini and Hechtlinger, 2013) that there is room for additional variability in our estimates of the SWFDR and that the estimated standard deviation of 1% is likely optimistic. Yet, we also believe that the estimates of the SWFDR—at least for well-designed and executed studies—is likely not $>50\%$. An interesting avenue for extension of our research is to consider conditional analyses including additional covariates for study type, journal, endpoint, and characteristics of the analysis performed.

The outcome of this future research aside, it is clear that the importance of the SWFDR inspires keen interest both among scientists and the general public. We were surprised by the rapid dissemination of an earlier version of our paper after quietly posting it to the ArXiv, but were encouraged by the thorough and thoughtful comments that our paper inspired (Neuroskeptic, 2013; Gelman, 2013; Leek, 2013)—leading to improvements in the paper even before it was submitted. We would therefore like to thank the discussants at Biostatistics (Cox, 2013; Gelman and O'Rourke, 2013; Schuemie *and others*, 2013; Benjamini and Hechtlinger, 2013; Goodman, 2013; Ioannidis, 2013) and worldwide who have also brought rational, empirical reasoning to bear on this important topic, and provided outstanding ideas for further development of our discussion of the SWFDR.

Supplementary material

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

Acknowledgments

Conflict of Interest: None declared.

References

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

Benjamini, Y. and Hechtlinger, Y. (2013). Discussion of: an estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, to appear.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association* **82**(397), 112–122.

Cox, D. R. (2013). Discussion of: an estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, to appear.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic epidemiology* **23**(1), 70–86.

FLETCHER, R. H. (1989). The costs of clinical trials. *Journal of the American Medical Association* **262**(13), 1842–1842.

FREEDMAN, D. H. (2010). *Lies, Damned Lies, and Medical Science*. www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269.

GELMAN, A. (2013). *I Don't Believe the Paper, "empirical estimates suggest most published medical research is true." That is, Most Published Medical Research May Well Be True, But I'm Not at All Convinced by the Analysis Being Used to Support this Claim*. andrewgelman.com/2013/01/24/i-dont-believe-the-paper-empirical-estimates-suggest-most-published-medical-research-is-true-that-is-the-claim-may-very-well-be-true-but-im-not-at-all-convinced-by-the-analysis-being-used.

GELMAN, A. AND O'ROURKE, K. (2013). Discussion of: an estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, to appear.

GOODMAN, S. (2013). Discussion of: an estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, to appear.

GOODMAN, S. AND GREENLAND, S. (2007). Why most published research findings are false: problems in the analysis. *PLoS Medicine* **4**(4), e168.

HERNDON, T., ASH, M. AND POLLIN, R. (2013). *Does High Public Debt Consistently Stifle Economic Growth?: A Critique of Reinhart and Rogoff*. Political Economy Research Institute. MA: Gordon Hall of Amherst.

HOTZ, R. L. (2007). *Most Science Studies Appear to be Tainted by Sloppy Analysis*. online.wsj.com/article/SB118972683557627104.html.

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

IOANNIDIS, J. P. (2013). Discussion of: an estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, to appear.

IOANNIDIS, J. P., NTZANI, E. E., TRIKALINOS, T. A. AND CONTOPOULOS-IOANNIDIS, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics* **29**(3), 306–309.

LEEK, J. (2013). *Why I Disagree with Andrew Gelman's Critique of My Paper About the Rate of False Discoveries in the Medical Literature*. simplystatistics.org/2013/01/24/why-i-disagree-with-andrew-gelmans-crit ique-of-my-paper-about-the-rate-of-false-discoveries-in-the-medical-literature.

NEUROSKEPTIC (2013). *Is Medical Science Really 86% True?* neuroskeptic.blogspot.com/2013/01/is-medical-science-really-86-true.html.

PENG, R. D. (2009). Reproducible research and biostatistics. *Biostatistics* **10**(3), 405–408.

PENG, R. D. (2011). Reproducible research in computational science. *Science (New York, Ny)* **334**(6060), 1226–1227.

PFEIFFER, T. AND HOFFMANN, R. (2009). Large-scale assessment of the effect of popularity on the reliability of research. *PLoS One* **4**(6), e5996.

SCHUEMIE, M. J., RYAN, P. B., SUCHARD, M. A., SHAHN, Z. AND MADIGAN, D. (2013). Discussion of: an estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, to appear.

STAFF, D. (2007). *Is Most Published Research Really False?* www.sciencedaily.com/releases/2007/02/070227105745.html.

STOREY, J. D. AND TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**(16), 9440–9445.