Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007
Advance Access publication on September 25, 2013

R

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

LEAH R. JAGER

Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA

JEFFREY T. LEEK*

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

jleek@jhsph.edu

Summary

The accuracy of published medical research is critical for scientists, physicians and patients who rely on these results. However, the fundamental belief in the medical literature was called into serious question by a paper suggesting that most published medical research is false. Here we adapt estimation methods from the genomics community to the problem of estimating the rate of false discoveries in the medical literature using reported P-values as the data. We then collect P-values from the abstracts of all 77 430 papers published in *The Lancet*, *The Journal of the American Medical Association*, *The New England Journal of Medicine*, *The British Medical Journal*, and *The American Journal of Epidemiology* between 2000 and 2010. Among these papers, we found 5322 reported P-values. We estimate that the overall rate of false discoveries among reported results is 14% (s.d. 1%), contrary to previous claims. We also found that there is no a significant increase in the estimated rate of reported false discovery results over time (0.5% more false positives (FP) per year, $P = 0.18$) or with respect to journal submissions (0.5% more FP per 100 submissions, $P = 0.12$). Statistical analysis must allow for false discoveries in order to make claims on the basis of noisy data. But our analysis suggests that the medical literature remains a reliable record of scientific progress.

Keywords: False discovery rate; Genomics; Meta-analysis; Multiple testing; Science-wise false discovery rate; Two-group model.

1. Introduction

Scientific progress depends on the slow, steady accumulation of data and facts about the way the world works. The scientific process is also hierarchical, with each new result predicated on the results that came before. When developing new experiments and theories, scientists rely on the accuracy of previous discoveries, as laid out in the published literature. The accuracy of published research is even more critical

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

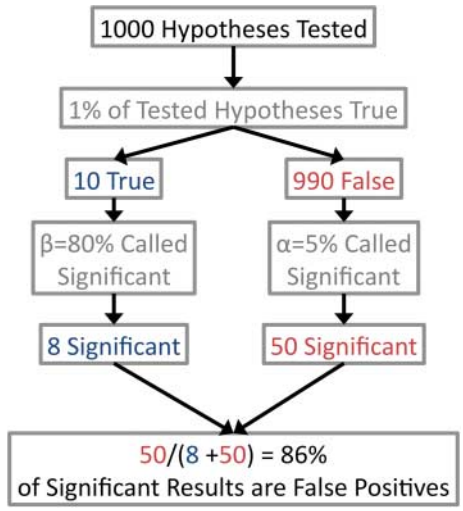

Fig. 1. The theoretical argument suggests that most published research is false. If the probability a research hypothesis is true is low, then most tested hypotheses will be false. The definition of P-values and customary significance cutoffs mean that $(\alpha \cdot 100)\%$ of false-positive hypotheses and $(\beta \cdot 100)\%$ of true-positive hypotheses will be called significant. If only 1% of tested hypotheses are true and the customary values of $\alpha = 0.05$ and $\beta = 0.8$ are used, then 86% of reported significant results will be false positives. This final percent of published results corresponding to false positives is the quantity that we estimate. A version of this figure appeared on the blog Marginal Revolution and is reproduced with permission (Tabarrok, 1989).

in medicine – physicians and patients may make treatment decisions on the basis of the latest medical research.

Ioannidis (2005) suggested that most published medical research is actually false, calling into serious question the fundamental belief in the medical literature. The claim is based on the assumption that most hypotheses considered by researchers have a low prestudy probability of being successful. The suggested reasons for this low pre-study probability are small sample sizes, bias in hypothesis choice due to financial considerations, or bias due to over testing of hypotheses in "hot" fields. On the basis of this assumption, many more false hypotheses would be tested than true hypotheses.

The rest of the argument is based on standard properties of hypothesis tests. If all P-values less than some value α are called significant than on average $(\alpha \cdot 100)\%$ of false hypotheses will be reported as significant. The usual choice for α in most medical journals is 0.05, resulting in 5% of false hypotheses being reported as significant. Meanwhile, true hypotheses will be called significant at a much higher rate, β. Many studies are designed so that β is a large value like 0.8; so that 80% of true hypotheses will be called significant.

However, if many more false hypotheses are tested than true hypotheses, the fraction of significant results corresponding to true hypotheses will still be low. The reason is that 5% of a very large number of hypotheses is still larger than 80% of a very small number of hypotheses (Figure 1, reproduced with permission from Tabarrok, 1989). The argument is plausible, since hypothesis testing is subject to error (van Belle *and others*, 2004), the pressure to publish positive results is strong (Easterbrook *and others*, 1991; Bhopal *and others*, 1997), and the number of submitted papers is steadily growing (Figure 2).

The assumptions that lead to the original result have been called into question, but counter arguments have focused on the logic behind the approach, specifically considering how evidence was potentially

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

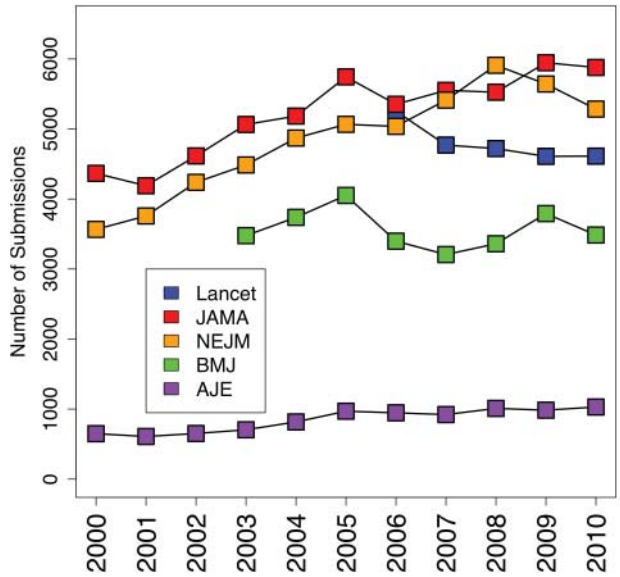

Fig. 2. Major medical journal submissions are increasing over time. A plot of the number of submissions to the major medical journals *The Lancet*, *The Journal of the American Medical Association* (*JAMA*), *The New England Journal of Medicine* (*NEJM*), *The British Medical Journal* (*BMJ*) and the flagship epidemiological journal *The American Journal of Epidemiology* (*AJE*) between the years 2000 and 2010. Submission data are available only for the years 2006–2010 for *The Lancet* and the years 2003–2010 for *The BMJ*.

misquantified (Goodman and Greenland, 2007). On the other hand, evidence-based medicine focuses on determining whether specific medical hypotheses are true or false by aggregating data from multiple studies and performing meta-analyses (von Elm and Egger, 2004; Altman, 2002). But, to date, there has been no empirical approach for evaluating the rate of false discoveries across an entire journal or across multiple journals.

To fill this gap, here we develop a statistical algorithm to directly estimate the proportion of false discoveries in the medical literature based only on reported, significant P-values. Our approach is derived from similar statistical methods for estimating the rate of false discoveries in genomic (Kendziorski *and others*, 2003; Newton *and others*, 2001; Efron and Tibshirani, 2002; Storey and Tibshirani, 2003) or brain-imaging studies (Genovese *and others*, 2002, where many hypotheses are tested simultaneously. There are serious problems with interpreting individual P-values as evidence for the truth of the null hypothesis (Goodman, 1999). It is also well established that reporting measures of scientific uncertainty such as confidence intervals are critical for appropriate interpretation of published results (Altman, 2005). However, there are well established and statistically sound methods for estimating the rate of false discoveries among an aggregated set of tested hypotheses using P-values (Kendziorski *and others*, 2003; Newton *and others*, 2001; Efron and Tibshirani, 2002). Since P-values are still one of the most commonly reported measures of statistical significance, it is possible to collect these P-values and use them as data to estimate the rate of false discoveries in the medical literature.

We collected all 5322 P-values reported in the abstracts of the 77 430 papers published in *The Lancet*, *JAMA*, *NEJM*, *BMJ*, and *AJE* between the years 2000 and 2010 (Section 2.2). Based on these data we are able to calculate an empirical estimate of the rate of false discoveries in the medical literature and trends in false discovery rates over time. We show that despite the theoretical arguments to the contrary, the medical literature remains a reliable record of scientific progress.

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

2. METHODS

2.1 *Estimating the science-wise false discovery rate*

Our goal is to estimate the rate that published research results are false discoveries. False discoveries, for our analysis, are cases where the null hypothesis is true in a hypothesis testing framework, but the results are reported as significant. We will call this quantity π_0. This is the same rate that was considered in the original paper claiming most published results are false (Ioannidis, 2005). Similar efforts have suggested that almost no null hypotheses are true in the economic literature and most tests that fail to reject the null do so because of a lack of power (De Long and Lang, 1989). When the goal is to estimate the false discovery rate among published results in the medical literature, we do not know all of the hypotheses tested by all researchers, so we cannot estimate the probability that a priori research hypotheses are false. But we can observe the reported P-values in the medical literature and use them to estimate the science-wise false discovery rate.

By definition, the P-values for tests under the null distribution are uniformly distributed between 0 and 1 when correctly calculated (Casella and Berger, 2002). The reported P-values in the literature represent a subset of all P-values computed. Specifically, they are most frequently the statistically significant P-values. The usual threshold for significance is $P < 0.05$, so we focus on developing a method that can be applied to only the reported P-values less than 0.05, since they can be observed. Another statistical property means that if we consider only the P-values that are less than 0.05, the null P-values must be distributed uniformly between 0 and 0.05, denoted Uniform(0, 0.05). Each true alternative P-value may be drawn from a different distribution, but statistical research has shown that when considered as one group, the distribution of true alternative P-values can be modeled as a Beta distribution, Beta(a, b) or mixtures of Beta distributions (Allison *and others*, 2002, 2006; Pounds and Morris, 2003; Leek and Storey, 2011). For a fixed $a < 1$, the parameter b quantifies the right skew of the Beta distribution. In other words, for a fixed $a < 1$, larger values of b correspond to distributions with more significant P-values. If we only consider the P-values less than 0.05, the distribution for alternative P-values can be modeled by a Beta distribution truncated at 0.05, denoted by t Beta$(a, b; 0.05)$. This conditional distribution models the behavior of reported P-values when scientists report all P-values less than 0.05 as significant, including the case where they test many hypotheses and only report the P-values less than 0.05 (see supplementary material available at *Biostatistics* online).

For any specific P-value, we do not know whether it comes from a truly null or alternative hypothesis. So, with probability π_0, it corresponds to the null distribution and with probability $(1 - \pi_0)$ it corresponds to the alternative. So, the P-value distribution can be modeled as a mixture of these two distributions (Newton *and others*, 2001; Efron and Tibshirani, 2002):

$$f(p|a, b, \pi_0) = \pi_0 \,\text{Uniform}(0, 0.05) + (1 - \pi_0) t \,\text{Beta}(a, b; 0.05) \tag{2.1}$$

A similar approach is taken in genomics or brain-imaging studies when many hypotheses are tested simultaneously, although in that case all P-values are observed, not just those <0.05. The parameters in equation (2.1) can be directly estimated using the expectation-maximization (EM) algorithm (Dempster *and others*, 1977).

One additional complication to equation (2.1) is that some P-values are left-censored. For example, a P-value of 0.000134343 may be reported as $P < 0.01$. This is a similar problem encountered in the analysis of survival data, when patients are lost to follow-up (Kaplan and Meier, 1958). If we treat the P-values that are reported as censored (those reported with $<$ or $\leqslant$, rather than $=$, in the abstract), then we can apply standard parametric survival analysis methods to extend model (2.1) to handle the censored observations (Kleinbaum and Klein, 2005). The key assumption here, as in standard survival analysis, is that censoring is independent of the P-value distributions. This assumption may be reasonable because

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

it is unlikely that among all scientists the decision to round is strongly correlated with the behavior of statistical models across scientists, labs, and journals.

A second complication is that some P-values are rounded, so that a P-value of 0.013 is reported as 0.01. Since P-values are continuous random variables, the probability of being exactly a round value like 0.01 is zero. So, the observation that many P-values are reported as 0.01, 0.02, 0.03, 0.04, or 0.05 strongly suggests that these P-values are rounded. We model these P-values as multinomial random variables, taking on one of the five values: 0, 0.01, 0.02, 0.03, 0.04, and 0.05. The probability of each rounded value is equal to the total probability assigned to all the P-values that round to that value. We can calculate these probabilities by integrating the distributions over the intervals rounding to 0, 0.01, 0.02, 0.03, 0.04, and 0.05: [0, 0.005), [0.005, 0.015), [0.015, 0.025), [0.025, 0.035), [0.035, 0.045), and [0.045, 0.05], respectively (MacDonald and Pitcher, 1979; Wengrzik and Timm, 2011). Again, the assumption is that rounding is independent of P-value, which is again likely because it would only happen if there is a correlation between the choice to round and the choice of statistical methods and analyses across scientists, labs, and journals.

With these modifications in place, we can now use the EM algorithm to estimate the parameters in the model. Specifically, we can estimate π_0, the rate of false discoveries. We can apply our algorithm to the P-values from all journals and all years to estimate an overall rate of false discoveries, or we can apply our algorithm individually to specific journals or years to estimate journal and year specific false discovery rates. Full mathematical details of our approach and R code for implementing our models are available at https://github.com/jtleek/swfdr.

2.2 *Collecting P-values from medical publications*

We wrote a computer program in the R statistical programming language (http://www.R-project.org/) to collect the abstracts of all papers published in *The Lancet*, *JAMA*, *NEJM*, *BMJ*, and *AJE* between 2000 and 2010. Our program then parsed the text of these abstracts to identify all instances of the phrases "P =", "P <", "P =", allowing for a space or no space between "P" and the comparison symbols. Our program then extracted both the comparison symbol and the numeric symbol following the comparison symbol. We scraped all reported P-values in abstracts, independent of the study type. The P-values were scraped from http://www.ncbi.nlm.nih.gov/pubmed/ on January 24, 2012. A few manual changes were performed to correct errors in the reported P-values due to variations in the reporting of scientific notation as detailed in the R code. To validate our procedure, we selected a random sample (using the random number generator in R) of abstracts and compared our collected P-values to the observed P-values manually. The exact R code used for scraping and sampling and the validated abstracts are available at https://github.com/jtleek/swfdr.

2.3 *Obtaining journal submission data*

We also directly contacted the editors of these journals, and they supplied data on the number of submitted manuscripts to their respective journals. We were only able to obtain publication data for the years 2006–2010 for *The Lancet* and for the years 2003-2010 for *BMJ*.

2.4 *False discovery rates over time*

We used a linear model to estimate the rate of increase or decrease in false discovery rates over time (McCulloch and Searle, 2001). The dependent variable was the estimated false discovery rate for a journal in each year. The independent variable was the year. We also fit a linear model relating the false

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

discovery rate to the number of submissions for each journal. Each model included an adjustment for journal.

2.5 *Nonuniform null P-values*

Theoretically, P-values should be distributed Uniform$(0, 1)$ under the null hypothesis. However, there are a variety of reasons that this may not be true, including common, unmeasured confounders (Leek and Storey, 2008) or P-value hacking, where individual steps in the statistical processing pipeline are changed to get more significant results (Simmons *and others*, 2011). Under these conditions, the null P-values may not be uniformly distributed, which would violate the assumptions of our algorithm. While the extent and frequency of these problems in the medical literature has not been estimated or documented, it could lead to bias in our estimates. So, we performed two simulation studies, one where the P-values followed our stated assumptions and one where each researcher tested 20 hypotheses and reported only in the minimum P-value. This represents a clear violation of our assumptions. Next we performed a simulation where all P-values were floored rather than rounded to two significant digits. Finally, we performed an additional simulation where the alternative distribution was held fixed and the null distribution was allowed to become progressively more anti-conservative to evaluate the sensitivity of our estimates to changes only to the null distribution.

2.6 *Reproducible research*

A key component of computational research is that the results be reproducible. We have adhered to the standards of reproducible research by both outlining all of our statistical methods in the supplemental material and making the exact R code used to produce our results available in supplementary material available at https://github.com/jtleek/swfdr. By distributing our methods, code, and software our approach adheres to the highest standards of reproducible computational research (Peng, 2012). We hope that by providing all codes used to perform our calculations and analysis that other users will be able to evaluate and improve on our science-wise false discovery rate estimation procedure.

3. RESULTS

Our computer program collected the abstracts from all 77 430 papers published in *The Lancet*, *JAMA*, *NEJM*, *BMJ*, and *AJE* between the years 2000 and 2010. The abstracts were mined for P-values, as described in Section 2.2. Of the mined abstracts, 5322 reported at least one P-value according to our definition. The relatively low rate of P-values in abstracts is due to (i) many articles not reporting P-values in the abstracts since they are essays, review articles, or letters and (ii) a trend toward decreased reporting of P-values and increased reporting of other statistical measures such as estimates and confidence intervals. Most reported P-values were less than the usual significance threshold of 0.05: 80% for *The Lancet*, 76% for *JAMA*, 79% for *NEJM*, 76% for *BMJ*, and 75% for *AJE*. Among the papers that reported P-values, a median of the two P-values were reported (median absolute deviation 1.5). Our validation analysis of 10 randomly selected abstracts showed that we correctly collected 20/21 P-values among these abstracts, with no falsely identified P-values (see supplementary material available at *Biostatistics* online).

The distributions of collected P-values showed similar behavior across journals and years (Figure 3). Most P-values were substantially <0.05, with spikes at the round values 0.01, 0.02, 0.03, 0.04, and 0.05. There were some variations across journals, probably due to variations in editorial policies and types of manuscripts across the journals. We applied our algorithm to all the P-values from all journals and

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

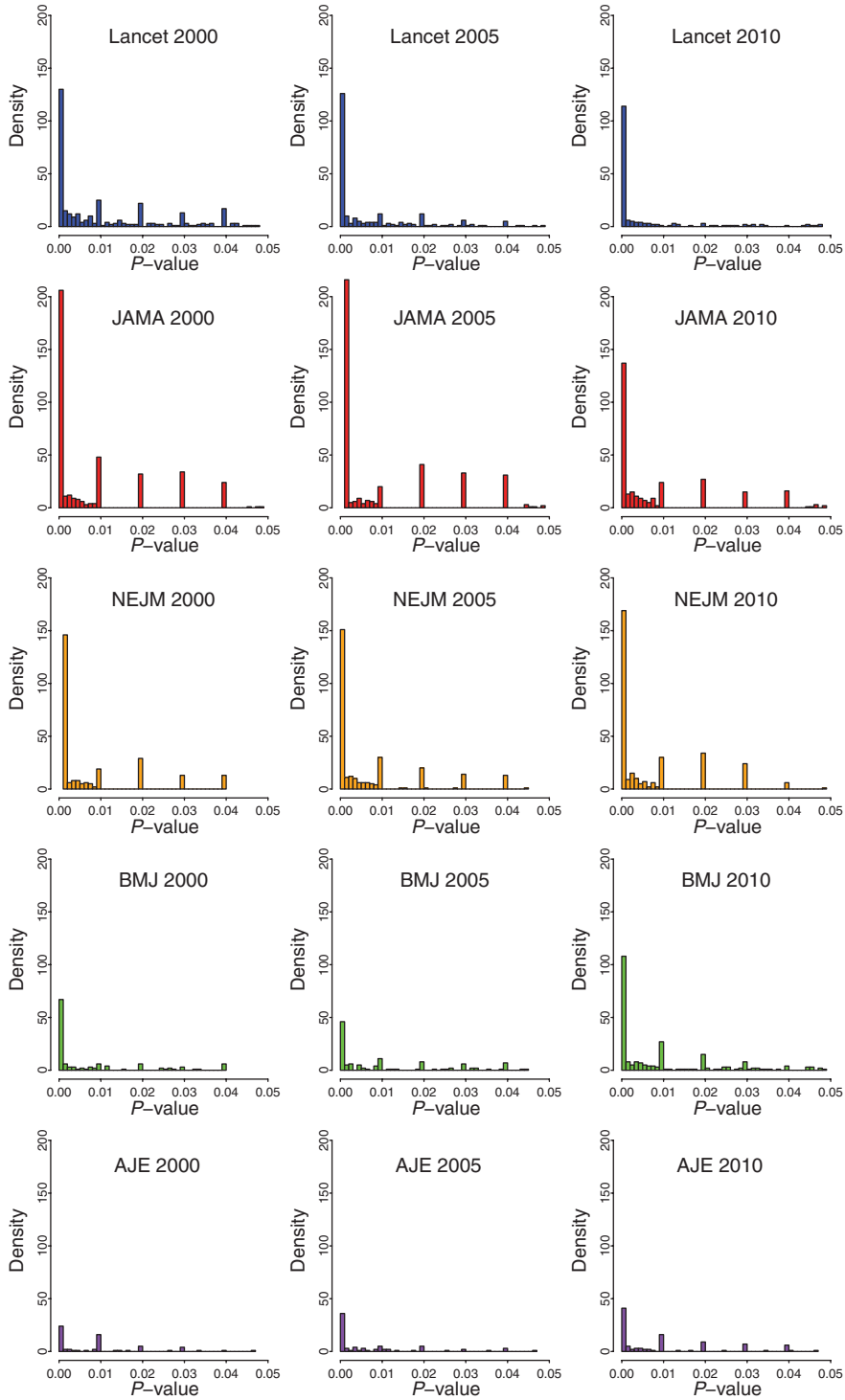

Fig. 3. Histogram of P-values < 0.05. The observed P-value distributions for all $P < 0.05$ scraped from PubMed for *AJE*, *JAMA*, *NEJM*, *BMJ*, and *The Lancet* in the years 2000, 2005, and 2010.

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

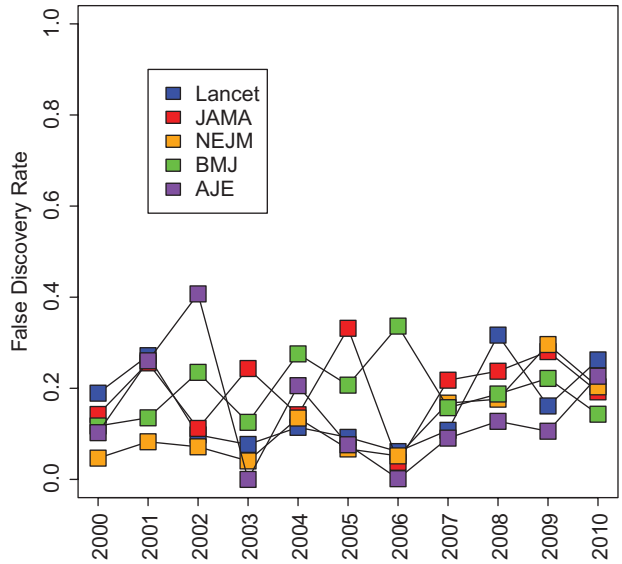

Fig. 4. Estimated false discovery rates for the years 2000–2010 by journal. The estimated false discovery rates for *AJE*, *JAMA*, *NEJM*, *BMJ*, and *The Lancet* in the years 2000, 2005, and 2010. There is no significant trend in false discovery rates over time or with increasing numbers of submissions.

all years to estimate the overall rate of false discoveries in the medical literature. The estimated rate of false discoveries among published results was 14% (s.d. 1%). We did find variation among journals with false discovery rates of 19% (s.d. 3%) for *The Lancet*, 17% (s.d. 2%) for *JAMA*, 11% (s.d. 2%) for *NEJM*, and 17% (4%) for *BMJ*. To compare with the four major medical journals, we calculated an estimate of the rate of false discoveries for *AJE*. The estimated rate was 11% (s.d. 4%), similar to that of the *NEJM*. The *AJE* is an epidemiological journal and publishes substantially different types of studies than the medical journals. Specifically, there is a substantially greater focus on observational studies. This suggests that the false discovery rate is somewhat consistent across different journal and study types.

Next we considered the rate of false discoveries over time. We applied our algorithm to estimate the rate of false discoveries separately for each journal and each year (Figure 4). We fit a linear model adjusting for a journal as detailed in the methods, and there was no significant trend in false discovery rates for any journal over time (0.5% more FP per year, $P = 0.18$). Similarly, the positive trend in the submission rates for these journals (Figure 2) was not associated with an increase in false discovery rates over the period 2000–2010. (0.5% more FP per 100 submissions, $P = 0.12$).

To evaluate the robustness of our approach to the assumption of uniformity of null P-values, we performed four simulation studies. In the first case we simulated 100 journals where researchers correctly calculated P-values and reported all P-values less than 0.05. These P-values exactly satisfy our model assumptions. In this case, we show that our estimates of the false discovery rate are very close to the true value (Figure 5a). In the second case, we simulated widespread P-value hacking, where in each study performed, only the minimum P-value was reported (Figure 5b). In this case, we underestimate the science-wise false discovery rate, but less so for small values of the false discovery rate like we have observed. Next we simulated a scenario where all P-values were always rounded down, rather than to the nearest value -

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

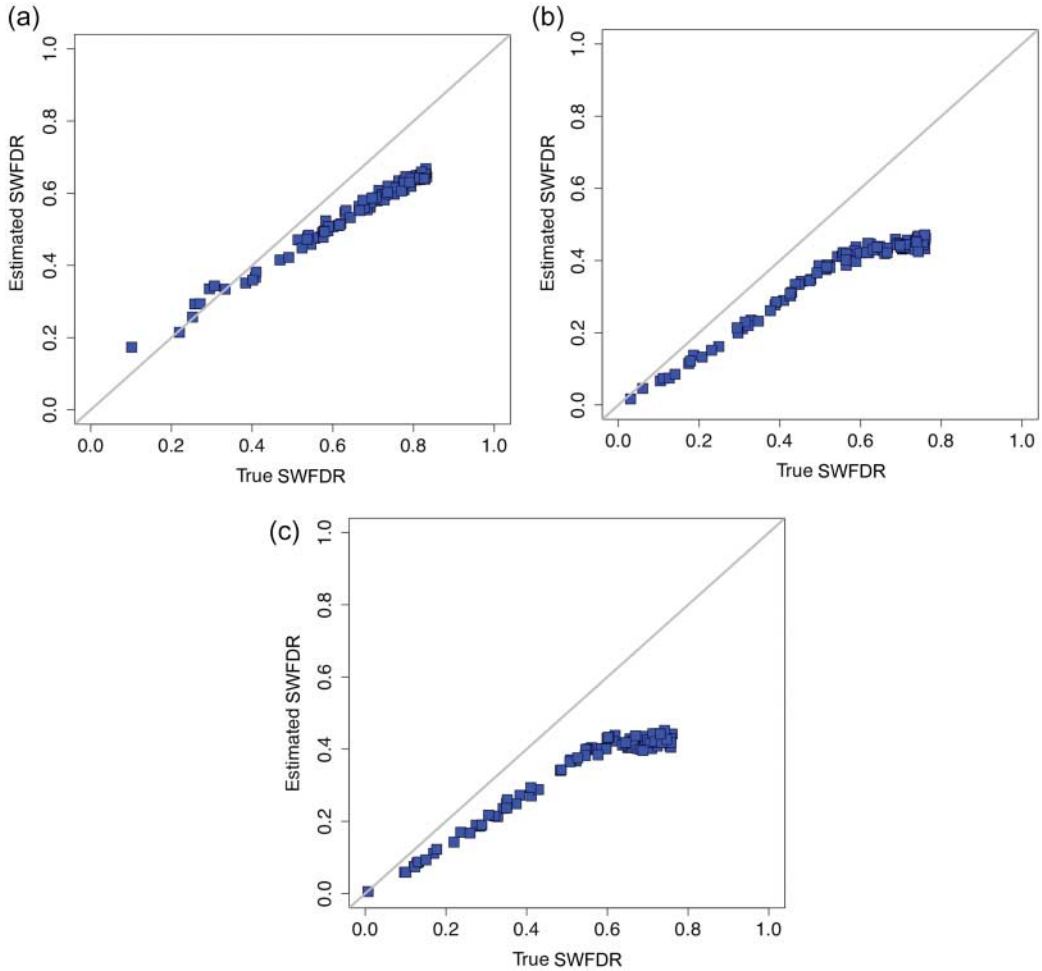

Fig. 5. Estimated versus true science-wise false discovery rate. (a) One hundred simulated journals were created where the P-values reported were all the P-values <0.05. In this case, our estimates match the true rate of false discoveries closely. (b) One hundred simulated journals were created where the P-values reported were only the minimum P-values from 20 hypothesis tests. This represents and extreme case of P-value hacking. In this case, as expected, our estimates are anti-conservatively biased. (c) One hundred simulated journal where all rounded P-values were floored to the next lowest hundredth rather than rounded to the nearest hundredth. Scenarios (a) and (b) represent consistent and widespread altering of P-values in an anti-conservative way.

a major violation of our rounding assumption. In this extreme case, we again underestimate the science-wise false discovery rate (Figure 5c). The latter two cases represent extreme scenarios where researchers are consistently and intentionally misrepresenting the analyses they have performed. To get a better idea of the sensitivity of our estimates to the null distribution of P-values we also performed a simulation where the alternative distribution was held fixed and the null distribution was allowed to become more and more anti-conservatively biased. As expected, when the null distribution and the alternative distribution are the same our estimates will be anti-conservative (Figure S1, see supplementary material available at *Biostatistics* online).

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

4. DISCUSSION

Here we proposed a new statistical method for estimating the rate of false discoveries in the medical literature directly based on reported P-values. We then collected data from the major medical journals and showed that most published medical research results in these journals are not false. A similar result held for the epidemiological journal *AJE*, even though this journal publishes a substantially different type of result in general. Our results suggest that while there is an inflation of false discovery results above the nominal 5% level (van Belle *and others*, 2004), but the relatively minor inflation in error rates does not merit the claim that most published research is false. Although our results disprove the original claim that most published research is false, they do not invalidate the criticisms of standalone hypothesis tests for statistical significance that were raised. Specifically, it is still important to report estimates and confidence intervals in addition to or instead of P-values when possible so that both statistical and scientific significance can be judged by readers (Altman, 2005; Altman and Gardner, 1986). In some cases, a truly alternative hypothesis may still have a very small effect size and still lead to a small P-value if the sample size is large enough, although these cases may not be scientifically interesting.

Here we have adapted methods from the genomic literature for false discovery rate estimation. But there are important distinctions between multiple testing in genomics studies and across unrelated studies. In most genomic studies, there are a large number of variables measured on the same set of samples. Generally, the hypothesis test performed on each of the measured variables will have the same form. In genomic studies, many of the null hypotheses are expected to be true and the primary variations in power are due to differences in effect sizes. In the medical literature, it is a substantially more complicated situation. Each study may have a different sample size, different statistical techniques may be used, and it is not clear that most null hypotheses are expected to be true. As long as P-values are correctly calculated under the null hypothesis and our stated reporting conditions hold, our estimates of the science-wise false discovery rate will be sound. However, it is clear that when considering hypothesis tests across multiple studies there will be violations of these assumptions to a greater extent than in genomic studies. An interesting avenue for future research would be to estimate the robustness of science-wise false discovery rate estimators to variations in the extent and the type of P-value hacking (Simmons *and others*, 2011) beyond the basic simulations that we have performed here.

An important consideration is that we have focused our analysis on the major medical journals. Another interesting avenue for future research would be to consider less selective or more specialized medical journals to determine journal characteristics that associate with variations in the false discovery rate. A limitation of our study is that we consider only P-values in abstracts of the papers. We chose to focus on the abstracts of papers as the P-values in abstracts primarily correspond to main effects. A potentially informative analysis could focus on breaking down the rate of false discoveries by whether the P-values corresponded to a primary or secondary analysis or other characteristics of the analyses performed—such as the type of hypothesis test performed. However, such a calculation would require more extensive development of text mining methods for automatically detecting differences in the characteristics of tested hypotheses. Despite these limitations, we have performed the first global empirical analysis of the rate of false discoveries in the major medical journals and we have shown that theoretical claims to the contrary, the data show that the medical and scientific literature remain a reliable record of scientific progress.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org and supplementary code is available from https://github.com/jtleek/swfdr.

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

Acknowledgments

Conflict of Interest: None declared.

Funding

This work was supported by a Johns Hopkins School of Public Health Faculty Innovation Fund Award to J.T.L.

References

Allison, D. B., Cui, X., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Prolla, T. A. and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20.

Altman, D. G. (2002). Poor-quality medical research: what can journals do? *Journal of the American Medical Association* **287**, 2765–2767.

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery* **29**, 554–556.

Altman, D. G. and Gardner, M. J. (1986). Confidence intervals rather than P-values: estimation rather than hypothesis testing. *British Medical Journal* **292**, 746–750.

Bhopal, R., Rankin, J., McColl, E., Thomas, L., Kaner, E., Stacy, R., Pearson, P., Vernon, B. and Rodgers, H. (1997). The vexed question of authorship: views of researchers in a British medical faculty. *British Medical Journal* **314**, 1009–1012.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edition. Pacific Grove, CA: Duxbury.

De Long, J. B. and Lang, K. (1989). Are all economic hypotheses false? *Journal of Political Economy* **100**, 1257–1272.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Easterbrook, P. J., Berlin, J. A., Gopalan, R. and Matthews, D. R. (1991). Publication bias in clinical research. *Lancet* **337**, 867–872.

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

Genovese, C. R., Lazar, N. A., Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15**, 870–878.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: the P value fallacy. *Annals of Internal Medicine* **130**, 995–1004.

Goodman, S. and Greenland, S. (2007). Why most published research findings are false: problems in the analysis. *PLoS Medicine* **4**, e168.

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014

KENDZIORSKI, C. M., NEWTON, M. A., LAN, H. AND GOULD, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

KLEINBAUM, D. G. AND KLEIN, M. (2005). *Survival Analysis: A self learning text*, 2nd edition. New York, NY: Springer.

LEEK, J. T. AND STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 18718–18723.

LEEK, J. T. AND STOREY, J. D. (2011). The joint null criterion for multiple hypothesis tests. *Statistical Applications in Genetics and Molecular Biology* **10**, 28.

MACDONALD, P. D. M. AND PITCHER, T. J. (1979). Age-groups from size-frequency data: a versatile and efficient method of analysing distribution mixtures. *Journal of Fisheries Research Board of Canada* **36**, 987–1001.

MCCULLOCH, C. E. AND SEARLE, S. R. (2001). *Generalized, Linear, and Mixed Models*, 1st edition. New York, NY: John Wiley and Sons.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

PENG, R. D. (2012). Reproducible research in computational science. *Proceedings of the National Academy of Sciences of the United States of America* **334**, 1226–1227.

POUNDS, S., MORRIS, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of P-values. *Bioinformatics* **19**, 1236–1242.

SIMMONS, J. P., NELSON, L. D. AND SIMONSOHN, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 1359–1366.

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

TABARROK, A. (2005). Why most published research findings are false. *Marginal Revolution*. http://marginalrevolution.com/marginalrevolution/2005/09/why_most_publis.html.

VAN BELLE G., FISHER, L. D., HEAGERTY, P. J. AND LUMLEY, T. (2004). *Biostatistics: A Methodology for the Health Sciences*, 2nd edition. Hoboken, New Jersey: John Wiley and Sons.

VON ELM, E. AND EGGER, M. (2004). The scandal of poor epidemiological research. *British Medical Journal* **329**, 868–869.

WENGRZIK, J. W. AND TIMM, J. (2011). Comparing several methods to fit finite mixture models to grouped data by the EM algorithm. *Proceedings of the World Congress on Engineering* **1**, 1.

[Received January 25, 2013; revised February 4, 2013; accepted for publication February 7, 201]

Downloaded from http://biostatistics.oxfordjournals.org/ by guest on January 7, 2014