*Biostatistics* (2013), *pp.* 1–4
doi:10.1093/biostatistics/kxt032

# Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek

YOAV BENJAMINI*

*Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel*
*Sagol School for Neuroscience, Tel Aviv University, Tel Aviv 6997801, Israel*
ybenja@tau.ac.il

YOTAM HECHTLINGER

*Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 6997801, Israel*



The original paper by Ioannidis (2005) has done much to raise the awareness to the possible sources of biases in current scientific work. However, Ioannidis described what *may* be happening in the scientific literature if scientists use blindly the marginal 0.05 significance level rule throughout their work. From this possibility, he reached the alarming conclusion that most research findings are false. In the current work, Prof. Jager and Prof. Leek took it upon themselves to find what actually is happening in current medical research if we assume that the practice is to publish only the findings significant at the 0.05 level. Since Ioannidis was actually arguing about the false discovery rate (FDR) in research (without using this terminology), they utilized FDR methods that are being used in a single research project to study the phenomenon across studies. They further developed modeling tools to address the peculiarities of the data they encountered, such as rounding to specific digits, and they developed relevant text extraction algorithms, allowing them to collect data on the entire population of important medical papers. Finally, they produced the analysis in a reproducible way, which allows everyone to repeat, check, and modify their analysis (as we did). All in all, we congratulate them for shouldering the challenge in such a way.

We raise three questions. (1) Does the $\sim 14 \pm 1\%$ science-wise FDR reflect the situation in medical research? (2) How can we improve the estimation? (3) Whatever the actual number is, if it is well above the perceived 5% FDR, how can it be brought under control?

(1) Is it $14 \pm 1\%$?

Unfortunately, this is not clear. There are problems with some of the decisions made as to the statements collected for the analysis: (i) only "$p < 0.05$" and typographical variations to it, but not "$p \leqslant 0.05$", (ii) only statements regarding significance expressed as "$p =$" and typographical variations to it, but not confidence intervals, and (iii) only statements from the Abstracts were analyzed. Each one of these decisions affects the estimate.

*To whom correspondence should be addressed.

With regard to decision (i): repeating their analysis (using their available and accessible tools) while moving the threshold slightly to "$p \leqslant 0.05$" increased their science-wise FDR estimate to 20.5%, much beyond the reported $\sim 14 \pm 1\%$.

With regard to decision (ii): selection of the significant findings at the 0.05 level is a very common practice even when reporting confidence intervals (CIs). It takes the form of reporting in the abstract the marginal 95% CIs that do not cover 0 (or do not cover 1 for odd ratios and relative risks). Although we did not quantify this phenomenon, many studies report only CIs and hence were entirely excluded from the analysis, with unknown effect on the estimate. More importantly, in some studies that report CIs and $p$-values, reported $p$-values tend to escort only the more decisive CIs.

With regard to decision (iii): examining in depth a sample of 25 papers from those used in the original analysis, 5 from each journal, revealed that the number of significance statements in the body of the papers varied much, from 1 to 68 with a median of 24, while the median of abstract-only data in the same papers was 4, close to the value 2 reported for the entire database. The implication is that there is a strong selection-into-the-abstract effect. As further evidence, we observed that in 19 of the 25 papers that we studied in depth the smallest $p$-value in the paper was further selected into the abstract.

One result of such a selection is that the distribution of the $p$-values of false discoveries in the abstracts that are smaller than 0.05 is stochastically smaller than uniform. The authors did study the implication of this problem on their estimate. From their Figure 5(b), which describes the effect of selecting the smallest of 20 uniformly distributed $p$-values into the abstract, one can approximate that the "true science-wise" FDR is about 30% when the estimate is 20%. While Jager and Leek discuss it as an extreme deviation from the assumptions, in view of the above we see it as a realistic or even optimistic scenario.

So far we discussed the data being used. There are also some problems with the analysis. The authors treated appropriately the values of 0.04, 0.03, 0.02, and 0.01 as being the result of rounding. But then they avoided using the same approach to address the practice of "rounding on a log-base-10 scale", even though only 7.8% of the $p$-values are at 0.01, while 36.9% are at 0.001, and 8.8% at 0.0001. More importantly, any $p$-value reported as "$p \leqslant x$" was treated as left censored at $x$. This means that if $p \leqslant 0.04$ is reported, it is treated as $0.00 < p \leqslant 0.04$, and not as $0.03 < p \leqslant 0.04$. Indeed 88% of the $p$-values at 0.001 and 90% of those at 0.0001 were censored. These two issues seem minor but may have a noticeable effect on the estimator. Different combinations of these assumptions can raise the estimated FDR level to above 30%.

As we see, these problems all affect the estimator in the liberal direction, giving rise to the assertion that the true value is much higher than 14%. A different point of view on the science-wise FDR in medical research is offered in a report by Kaplan (2008), written in collaboration with the American Food and Drug Administration, which revealed in those years an increase in the Phase III failure rate from 30% to 50%. Since Phase III studies are conducted only for results that were supported by research at earlier phases, the 50% figure can be considered as another estimate of the science-wise FDR in medical research, possibly an upper one. In summary, 20% and 50% should bracket the true science-wise FDR in medical research.

(2) Improving the estimation

It is important to realize that there are different reporting methodologies in different research areas, and even within subareas. In medical research involving human subjects, a study is usually summarized by the result for a primary endpoint, as noted by the authors. In regulatory research, a study is considered a failure if the primary endpoint fails to show the effect of treatment. A primary endpoint is often single and so is not affected by selection in the published paper. However, the primary endpoints may suffer from drawer and publication biases, where results that are not significant are not submitted for publication and if submitted are less likely to get published. An analysis of published data cannot help, and here we may need to rely on databases that include results regarding all clinical trials performed.

Medical research publications report more than primary endpoints: secondary endpoints, safety endpoints, and subset analysis (or "heterogeneous treatment effect"). One would preferably have a different estimate of the FDR for each of these families of inferences.

Thus, we think it to be unavoidable to get deeper into the meaning of the statements made in each paper, find what types of findings are reported, define the relevant families, and collect their associated $p$-values before proceeding to the science-wise estimation stage. Also, research in epidemiology and psychology is endowed with a complex structure, offering different types of conclusions in a single study, and the difficulty is confounded by the extensive use of CIs that are harder to mine than $p$-values.

In the above areas of research, mining the relevant data directly from the database may be impossible. We found the alternative of studying in depth a sample of the publications to be feasible and useful for these purposes.

(3) Estimate science-wise FDR, or try to control it?

In spite of the possible impression from the discussion so far, we do not think that much effort should be invested in conducting a thorough investigation, overcoming the limitations of the current study, and offering better and more refined estimates. We think that the study of Jager and Leek is enough to point at the serious problem we face: even though most findings may be true, whether the science-wise FDR is at the more realistic 30% or higher, or even at the optimistic 20%, it is certainly too high.

It may be clear by now that we attribute the high level of science-wise FDR first and foremost to the problem of unadjusted selective inference. It has been demonstrated in the work of Jager and Leek that selecting the significant findings in the abstract at a 0.05 level results in an FDR that is too high. This was the point made originally by Soriç (1989), a physician by training, who warned about the naïve practice of 0.05 testing in hypotheses rich medical research that may result in most discoveries being false, a point reiterated forcefully by Ioannidis (2005). The work of Soriç was followed by the introduction of the FDR concept and testing procedures that transformed the warning of Soric into an operational goal. The goal is simple: requiring that the type I error, averaged over the selected discoveries, will be controlled in expectation; in more general terms, requiring that the original property of the inference will hold "on the average over the selected". This is the essence of selective inference, and should be a minimal requirement when selecting the promising leads from the many considered.

Taking the original Fisherian point of view, significance testing is an effort to address the selection of an interesting finding regarding a single parameter from the background noise. Modern science faces the problem of selection of promising findings from the noisy estimates of many. Assuming that the case of truth-seeking honest researchers is the rule, these researchers face the real problem of inference on the selected. Theoretical research shows that the selective inference problem does not disappear even if you assume that the null hypothesis can never be true (see Abramovich *and others*, 2006), or even if you replace hypothesis testing with CIs (see Benjamini and Yekutieli, 2005).

In genomics, proteomics, imaging, and other areas where "many" parameters mean tens or hundreds of thousands, the dangers have become obvious, and therefore are often addressed. In medical research for drug registering it is also addressed. Unfortunately, this is not the case in general medical research (nor is it in experimental Psychology or Epidemiology), where adjustment for the effect of selection is rarely done (see Fletcher and Colditz, 2002). For example, the study by Rami Cohen of a sample of 60 papers in The New England Journal of Medicine from 2001 to 2006 (reported in Benjamini, 2010) reveals that all had multiple endpoints, but 47 of the papers did not address multiplicity at all, and even in those that addressed it did so partially. Attending to selective inference issues via FDR controlling tests and False Coverage-statement Rate controlling CIs is an appropriate approach that places the horse before the cart: if you care about the FDR in science, take the needed means to control it in your own study. Then, let meta-analysis take care of the selection effect across studies. This combined strategy is crucial for assuring the replicability of scientific discoveries.

There is still much work ahead. To address the actual needs of researchers in particular areas of science, appropriate methodologies should be developed, as not all methods are relevant for all situations. Some of the situations that lack methodologies are selective inference for primary and secondary endpoints (Cohen's thesis, 2013), safety analysis (Benjamini, 2010), and heterogeneous treatment effect analysis,

along with many other specific problems. We hope that the huge interest that the paper by Jager and Leek has already aroused in the statistical community following its web publication will be channeled toward developing remedies for the too high science-wise FDR. They should therefore be thanked for their influential effort.

## ACKNOWLEDGMENTS

*Conflict of Interest:* None declared.

## FUNDING

The research reported in this paper was supported by a European Research Center grant (PSARPS).

## REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. AND JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics* **34**, 584–653.

BENJAMINI, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* **52**(6), 708–721, doi:10.1002/bimj.200900299.

BENJAMINI, Y. AND YEKUTIELI, Y. (2005). False discovery rate controlling confidence intervals for selected parameters. *Journal of the American Statistical Association* **100**, 71–80.

COHEN, R. (2013). Hierarchical weighted methods that control the false discovery rate and their use in medical research. Tel Aviv University, [PhD. Thesis].

FLETCHER, S. W. AND COLDITZ, G. A. (2002). Failure of estrogen plus progestin therapy for prevention. *Journal of the American Medical Association*, **288**, 366–369.

KAPLAN (2008). *Current Dilemma in Drug Development Increasing Failure Rate of investigational Drugs in Phase 3*. http://powershow.com/view1/183eb3-YjZmY/Current_Dilemma_in_Drug_Development_Increasing_Failure_Rate_of_Investigational_Drugs_in_Phase_3_Clin_powerpoint_ppt_presentatio.

SORIÇ, B. (1989). Statistical "discoveries" and effect size estimation. *Journal of the American Statistical Association* **84**, 608–610.