Biostatistics (2013), *pp.* 1–4
doi:10.1093/biostatistics/kxt037

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

MARTIJN J. SCHUEMIE*, PATRICK B. RYAN

Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD 20814, USA
Janssen Research and Development LLC, Titusville, NJ 08560, USA
mschuemi@its.jnj.com

MARC A. SUCHARD

Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD, USA
Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA

ZACH SHAHN, DAVID MADIGAN

Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD, USA
Department of Statistics, Columbia University, New York 10027, NY, USA

The paper "An estimate of the science-wise false discovery rate (SWFDR) and application to the top medical literature" by Jager and Leek provides an interesting perspective on FDRs in published medical research studies. As the authors point out, the distribution of p-values under the null hypothesis in these studies may well depart from uniform, and for this reason the authors explore the consequences of "p-hacking" and rounding of p-values on the FDR. Even after accounting for p-hacking, the authors conclude that the FDR is not alarmingly high. However, the authors appear to ignore other possible sources of departure from uniformity such as bias, model misspecification, and measurement error.

In a recent experiment, we investigated the distribution of p-values when the null hypothesis is true for real-world studies using large-scale longitudinal observational databases (Schuemie *and others*, 2013). More specifically, we developed a set of "negative controls" in a systematic evaluation of the association between hundreds of pharmaceutical products and a set of clinically relevant adverse drug reactions. Our negative controls were drug–outcome pairs for which an expert panel deemed that no evidence of a causal association exists, i.e. we believe the true effect size is zero. We then applied multiple epidemiological

*To whom correspondence should be addressed.

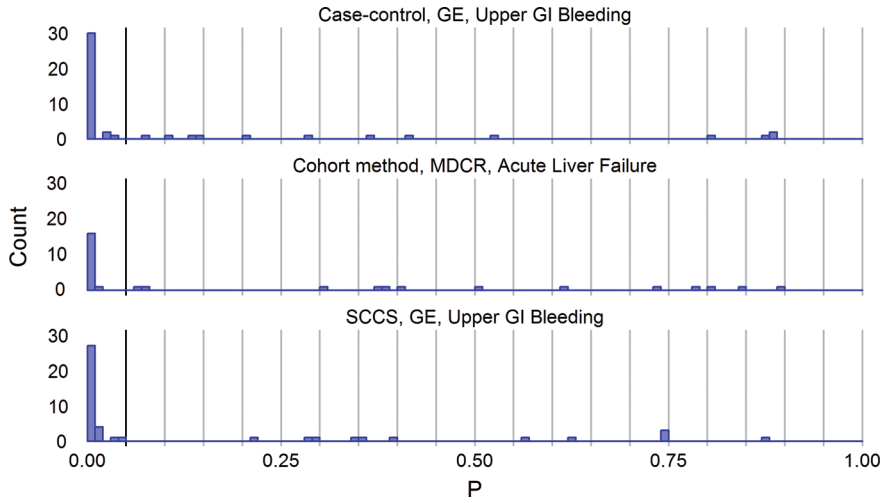

Fig. 1. Histograms of p-values in analyses of observational data where we believe the null hypothesis is true. SCCS, self-controlled case series; GE, general electric centricity database; MDCR, MarketScan medicare supplemental beneficiaries database).

designs to generate estimated effect sizes and associated p-values for these negative controls across different large-scale patient-level healthcare databases. Figure 1 shows the histograms for three study designs that are used throughout the medical literature: case–control studies, studies using a cohort method, and self-controlled case series (SCCS) (Farrington, 1995). Our experiments cover a broad range of outcomes and data sources, but in our discussion here we focus on the application of case–control and SCCS designs to a set of drugs in relation to upper gastro-intestinal bleeding within the general electric (GE) centricity database, a large electronic health record database, and the cohort method to drugs in relation to acute liver failure within the MarketScan Medicare Supplemental Beneficiaries database, a large administrative claims database of retirees who purchase additional insurance to augment their Medicare benefits.

In Figure 1, we note that the distribution of observed p-values for our negative controls shows severe departures from uniformity, even though we performed no p-value hacking or rounding. This finding suggests that the study estimates themselves are severely biased. Despite the use of advanced epidemiological techniques such as propensity score adjustment (Schneeweiss *and others*, 2009) and self-controlled designs (Farrington *and others*, 1996), it seems that we remain unable to sufficiently correct for confounding and other measurement errors.

In order to understand to what extent this violation of the uniformity assumption affects the estimates of the FDR, we adapted Jager and Leek's simulation as shown in their Figure 5(a). In their simulation, p-values for studies where the alternative hypothesis is true draw from a beta distribution and p-values for studies where the null hypothesis is true draw from a uniform distribution. In our adaptation, we sampled with replacement the p-values for studies where the null hypothesis holds from the empirical p-value distribution we observed in our experiment (see Figure 1). Figure 2 presents the result of our simulation study, indicating that the estimated FDR is not informative about the true FDR. We have added the R code and data to reproduce this figure as supplementary material available at *Biostatistics* online.

In summary, the scope of Jager and Leek's analyses includes "p-hacking" and rounding, but fails to address other threats to the validity of the core uniformity assumption. Our experimental results strongly suggest that these threats substantially impact real-world FDRs.

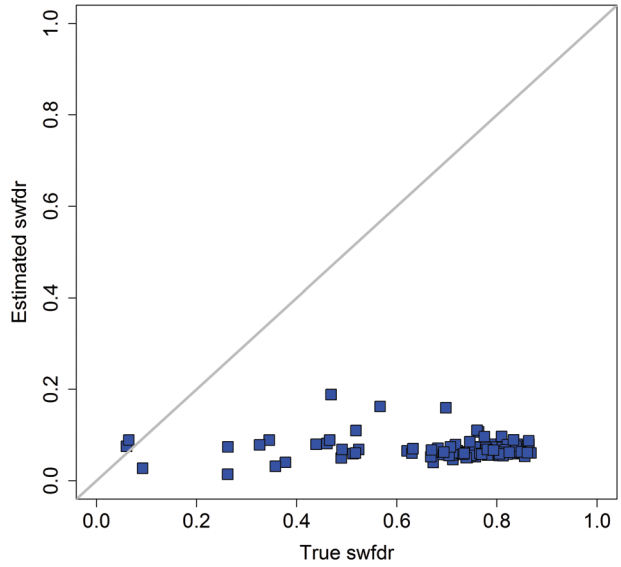

Fig. 2. Estimated versus SWFDR. We created 100 simulated journals where the p-values reported were all the p-values <0.05. When the alternative hypothesis is true, we draw the p-values from a beta distribution. When the null hypothesis is true, we resampled with replacement the p-values from the p-values observed in real observational analyses, as shown in Figure 1.

Supplementary material

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

Acknowledgments

Conflict of Interest: Drs Schuemie and Ryan are employees of Janssen Research and Development LLC.

Funding

The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development LLC, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc., Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi, Schering-Plough Corporation, and Takeda.

References

Farrington, C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* **51**(1), 228–235.

Farrington, C. P., Nash, J. and Miller, E. (1996). Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American Journal of Epidemiology* **143**(11), 1165–1173.

SCHNEEWEISS, S., RASSEN, J. A., GLYNN, R. J., AVORN, J., MOGUN, H. AND BROOKHART, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**(4), 512–522.

SCHUEMIE, M. J., RYAN, P. B., DUMOUCHEL, W., SUCHARD, M. A. AND MADIGAN, D. (2013). Interpreting observational studies: why empirical calibration is needed to correct *p*-values. *Statistics in Medicine*. http://onlinelibrary.wiley.com/doi/10.1002/sim.5925/abstract;jsessionid=600EF19932F674BCC9E2DEBF20110E76.d03t01.