Commentary

Statistics in Medicine

(wileyonlinelibrary.com) DOI: 10.1002/sim.6977 Published online in Wiley Online Library

Robust empirical calibration of *p*-values using observational data

Martijn J. Schuemie,[a,e*†] **George Hripcsak,**[b,e] **Patrick B. Ryan,**[a,b,e] **David Madigan**[c,e] **and Marc A. Suchard**[d,e]

In our previous paper [1], we proposed empirical calibration of *p*-values as a strategy for mitigating risk of systematic error when estimating average treatment effects from observational studies. By estimating the effect of exposure on outcomes across a collection of settings where the exposure is not believed to cause the outcome (negative controls), one can estimate an empirical null distribution of the exposure effect and compute calibrated *p*-values that take both random and systematic error into account. Gruber and Tchetgen [2] recently published a simulation that is intended to demonstrate the theoretical scenario in which empirical calibration may not be recommended. We welcome a thoughtful debate about the theoretical and empirical underpinnings of empirical calibration and share an enthusiasm to develop practical solutions that can be made broadly applicable to all observational analyses as a means of generating more reliable evidence. However, as we explain in more detail later, we would like to highlight and challenge the premise of some of the concerns raised by Gruber and Tchetgen and demonstrate how our empirical findings support empirical calibration as a robust approach to Type I error control. We believe their simulations are not realistic: the simulated estimates of negative controls showed severe bias with an odds ratio (OR) of 3, and perhaps more important, this bias was also simulated to be unusually homogeneous across the sample of negative controls. No one would continue their analysis after observing such estimates for their negative controls, and in all our real-world experiments using calibration, we have never seen such bias or homogeneity of negative controls. They confirm our findings that nominal *p*-values are strikingly and perhaps dangerously optimistic but recommend that we continue to use them. We believe that *p*-value calibration should be carried out, and the results should be reported.

We first would like to make the following points of clarification:

(1) Empirical calibration deals with bias in observational research estimates, which is the bias that remains after any measures taken to adjust for bias and confounding inherent in observational data, such as propensity score adjustment or using a self-controlled design.
(2) Empirical calibration does not require that the sources and structure of this bias are the same for all negative controls. Instead, the calibration approach assumes that the bias observed for each negative control draws from a common distribution of biases that can plague the specific analysis at hand. For example, if we suspect strong confounding by indication for our exposure-outcome pair of interest, it is not necessary for all negative controls to have the same confounding by indication.
(3) We believe bias inherent in an analysis can be approximated using a Gaussian distribution for our purposes.

[a] *Janssen Research and Development LLC, Titusville, NJ, U.S.A.*
[b] *Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, U.S.A.*
[c] *Department of Statistics, Columbia University, New York, NY, U.S.A.*
[d] *Department of Biomathematics and Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, U.S.A.*
[e] *Observational Health Data Sciences and Informatics (OHDSI)*
**Correspondence to: Martijn J. Schuemie, Janssen Research and Development LLC, Titusville, NJ, U.S.A.*
[†] *E-mail: schuemie@ohdsi.org*

Previously, we performed a leave-one-out validation study that provides strong support for these points. In this analysis, for each negative control, we computed the empirical null distribution using all other controls and calibrated the *p*-value of the held-out control using that distribution. Results showed good calibration properties (e.g. approximately 5% of the negative controls had calibrated $p < 0.05$ as expected) despite the fact that negative controls represented a wide range of drug-outcome pairs, with potentially very different sources of bias. (Note that we recommend that people always perform this leave-one-out analysis, because in specific cases the previously mentioned assumptions could be violated.)

2. Exchangeability

The empirical calibration approach requires an exchangeability assumption that there is some level of exchangeability between the negative controls and the drug-outcome hypothesis of interest. Our leave-one-out evaluation shows that there is exchangeability within the negative controls; for any negative control, the others can be used to calibrate its *p*-value accurately. But we have no evidence that this exchangeability holds for drug-outcome pairs outside the set of negative controls.

Gruber and Tchetgen's main challenge is to this assumption. They argue that theoretically all negative controls could show strong bias in one direction, while the hypothesis of interest shows no bias or the other way around, in which case empirical calibration will fail. They use simulations to demonstrate this point. In their 'medium' simulation, all negative controls show strong bias, with a mean OR of 3, whereas the hypotheses of interest all have small bias, as shown in Figure 1.

Gruber and Tchetgen provide no evidence that such conditions happen in real applications. They do, however, motivate these conditions by noting that when using negative exposure controls for a newly marketed drug that 'for example, uptake of a newly marketed drug, prevalence of off-label drug use, availability of alternative treatment options and physician prescribing behaviors typically changes in the years following drug approval.' To frame this in their simulation, Gruber and Tchetgen assume that a particular study design (e.g. a new-user cohort design using propensity score adjustment) would be almost unbiased for a newly marketed drug (the hypothesis of interest) but would systematically produce an OR of 3 for negative control drugs that have been on the market longer. They do not discuss the consequence that a reasonable researcher would become suspicious when observing OR = 3 for all negative controls and might decide to discard the analytic approach.

In practice, negative controls are inherently heterogeneous. Figure 2 shows the distribution of negative control estimates from a new-user cohort study [3] comparing the risk of gastrointestinal (GI) bleeds in celecoxib versus diclofenac users, where we use negative outcome controls (i.e. outcomes not believed to be causally related to either drug) (see the Supporting information for details of the study). In this particular analysis, we make no adjustments for confounding, and we observe large bias as expected.

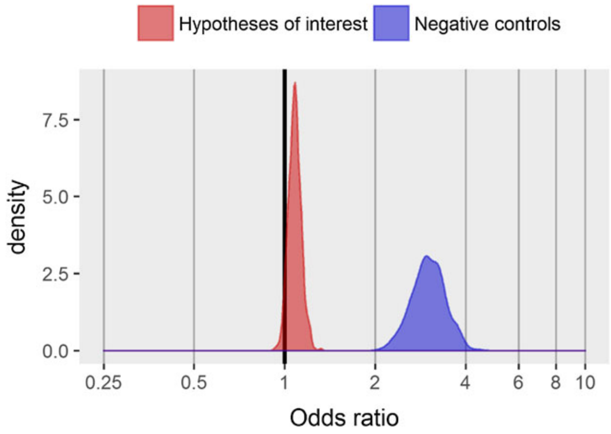

Figure 1. Density plot of the point estimates generated in the 'medium' simulation by Gruber and Tchetgen, corresponding to their Study Ib.

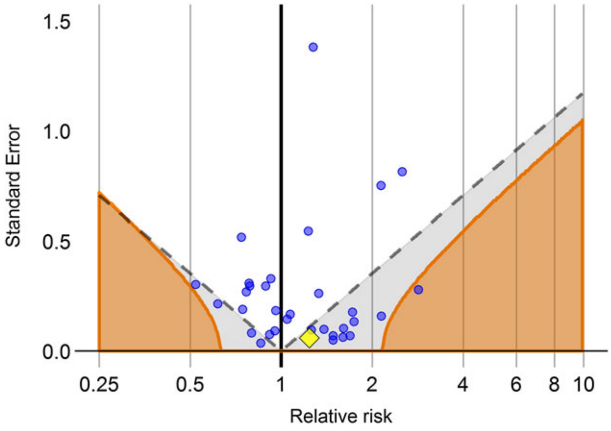

Figure 2. Estimates from a crude comparison between celecoxib and diclofenac users. Estimates below the dashed line (gray area) have $p < 0.05$ using traditional p-value calculation. Estimates in the orange areas have $p < 0.05$ using the calibrated p-value calculation. Blue dots indicate negative controls, and the yellow diamond indicates the outcome of interest: GI bleeds.

In this real-world example the distribution of negative controls looks nothing like the simulation. Overall, negative controls are positively biased as one would expect; celecoxib tends to be prescribed to a frailer population than diclofenac. But much more pronounced is the large variability in bias, both in positive and negative direction.

Figure 3 shows the results of the leave-one-out cross-validation of the negative controls, demonstrating superior calibration for the empirically calibrated p-value compared with the traditional p-value. Even though negative controls are very different from one another (negative controls includes 'ingrowing nail', 'allergic rhinitis' and also 'obsessive compulsive disorder'), we achieve better calibration than when using traditional p-value computation, suggesting the exchangeability assumption holds within the set of negative controls.

We have no evidence that these performance characteristics apply for exposure-outcome pairs outside the set of negative controls. However, given the large heterogeneity in the types of controls, ranging from infectious diseases to mental disorders and accompanying heterogeneity in underlying confounding, it is likely the bias of the hypothesis of interest is represented to some extent by the negative controls. Furthermore, the observed empirical null distribution is far wider than those simulated by Gruber and

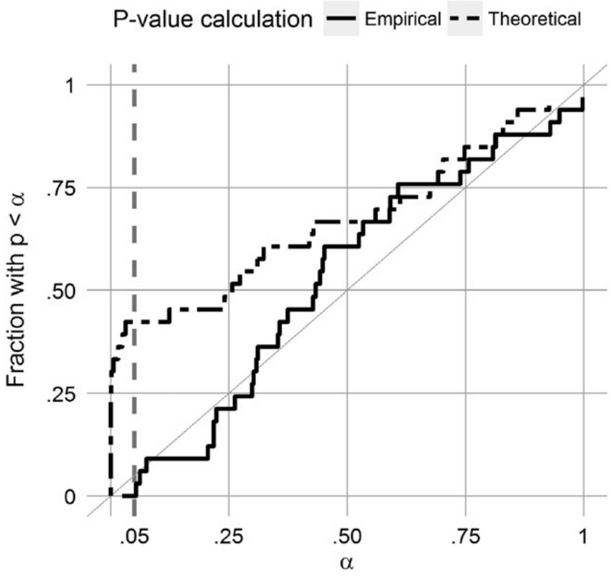

Figure 3. Calibration plot of the crude analysis. For the calibrated p-value, we employ a leave-one-out design with the negative controls and plot the fraction of p-values greater than nominal type I error rate α versus α.

Tchetgen, making it more likely there is at least some overlap with the bias distribution for the hypotheses of interest. Even though the calibration performance may not be perfect, it is unlikely to be as bad as simulated by Gruber and Tchetgen, where the two distributions were far apart as shown in Figure 1.

3. Type II error

Gruber and Tchetgen also lament that correcting type I error by *p*-value calibration can drastically increase type II error. As noted in our original paper: 'The method proposed here aims to correct the type I error (erroneously rejecting the null hypothesis) level, most likely at the cost of vastly increasing the number of type II errors (erroneously rejecting the alternative hypothesis).' [1] There is no free lunch: we cannot have both low type I and type II error in the face of a strongly biased estimator, such as the one depicted in Figure 2. In this plot, it is clear that no matter how large the sample size, no estimate between 0.6 and 2.1 will ever reach statistical significance when using *p*-value calibration. Even when the null hypothesis is not true, we will most likely fail to reject it.

This trade-off is not a drawback of the calibration method, but instead exposes a limitation of the observational data and the analysis design being employed. With a less biased estimator, we could achieve both lower type I and type II error. For example, Figure 4 reports the same results as Figure 2 but now using an adjusted analysis using propensity score matching.

In this analysis, there is little bias, and the calibrated *p*-value does not differ much from the traditional *p*-value. Figure 5 shows that in a leave-one-out validation, both *p*-values are indeed comparable. It is worthwhile to note that without examining negative controls in the crude analysis, we would not have unearthed the limitations of that design.

4. Discussion

Gruber and Tchetgen challenge the assumption of exchangeability, arguing that newly marketed drugs may have inherently different bias than exposure controls, which are typically older drugs. However, in their simulations, they underestimate the variability in the bias observed for negative controls in the real world. Given the width of the estimated empirical null distribution, it is likely that the bias of the hypothesis of interest is included in this distribution. Therefore, even if the exchangeability assumption is violated, and Gruber and Tchetgen have provided no evidence that this is the case, this most likely would mean the calibrated *p*-value would err on the conservative side, leading to larger *p*-values and making it harder to reject the null. As the mean and standard deviation of the empirical null distribution approach zero, as can be seen in Figure 4, we interpret this as meaning that the method is becoming insensitive to confounding in the data. A consequence is that the calibrated *p*-value is becoming less sensitive to the exchangeability assumption. In other words, if large bias and variability

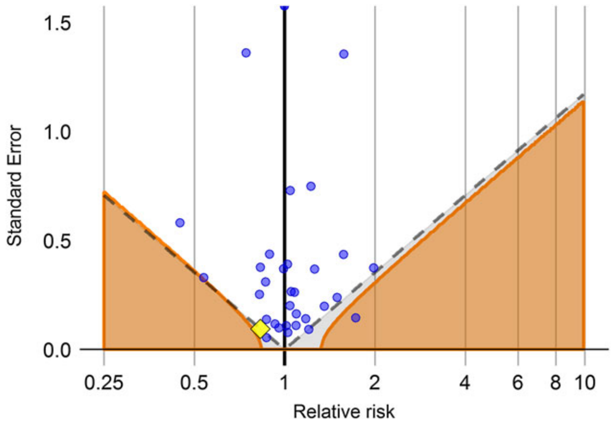

Figure 4. Estimates from a propensity-score matched comparison between celecoxib and diclofenac users. Estimates below the dashed line (gray area) have $p < 0.05$ using traditional *p*-value calculation. Estimates in the orange areas have $p < 0.05$ using the calibrated *p*-value calculation. Blue dots indicate negative controls, and the yellow diamond indicates the outcome of interest: GI bleeds.

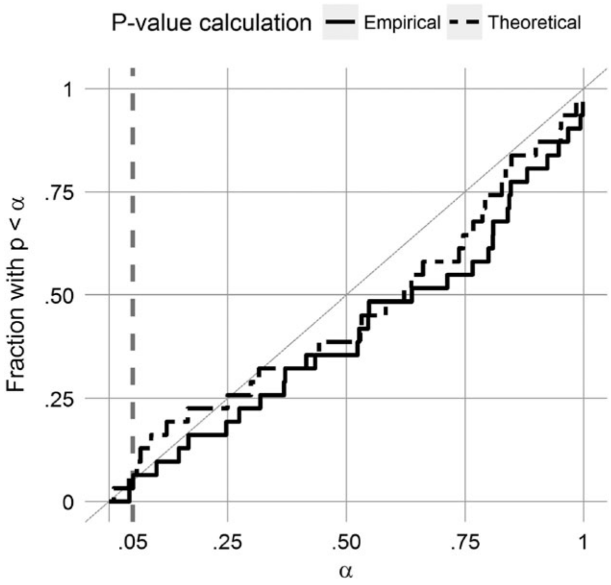

Figure 5. Calibration plot of the adjusted analysis. For the calibrated *p*-value, a leave-one-out design was used.

in bias is observed for the negative controls, one should be aware there is a possibility one's calibrated *p*-values are too conservative. If small or no bias is observed, we should feel confident the calibrated *p*-value is approximately correct. It should also be noted that Gruber and Tchetgen's criticism only applies to exposure controls for new drugs, in which case one could elect to use outcome controls instead.

We fully agree with Gruber and Tchetgen that empirical calibration can lead to a large increase in type II error when using a strongly biased estimator. It seems Gruber and Tchetgen suggest that if one is truly concerned with only type II error, one should forego calibration and accept that the *p*-value does not represent any meaningful statistic. But we think that in this case a better approach would be to still use calibrated *p*-values and increase one's nominal α to decrease type II error. In that case, at least the increase in type I error would be quantifiable. We argue that the best solution is to use designs that have small to no error, as achieved here in this example using a propensity score adjusted new-user cohort design. For such designs the empirical calibration has little effect, and both type I and type II error are, in some sense, optimal.

We do not agree with Gruber and Tchetgen that our recommendation that 'observational studies always include negative controls to derive an empirical null distribution and use these to compute calibrated *p*-values' is premature. Our findings so far suggest that the use of negative controls is a good way to evaluate the bias inherent in a study design, as demonstrated here by comparing a crude and adjusted analysis. The leave-one-out evaluations show that the calibrated *p*-values provide superior (in the case of large bias) to equal (in the case of almost no bias) calibration properties compared with traditional *p*-values. We also do not see viable alternatives. For example, the control outcome calibration approach proposed by Gruber and Tchetgen requires a strong exchangeability assumption that the structure and magnitude of the bias is exactly the same for the control and hypothesis of interest. Even if this assumption were to hold true, it remains unknowable to the researcher.

Given the large sample size in most observational studies, even small bias can 'distort the *p*-values beyond repair'. [4] Interpretation of results of observational studies is impossible without quantifying the bias, and we currently see empirical calibration as the only viable means to do this. We believe empirical calibration should always be used and presented in observational studies. To facilitate this, we have created an R package called EmpiricalCalibration (available in Comprehensive R Archive Network (CRAN)) and are working on a tool to quickly identify candidate negative controls.

5. Conflicts of interest and source of funding

Drs Schuemie and Ryan are employees of Janssen Research and Development, LLC. Drs Madigan and Suchard are partially supported by National Science Foundation grant IIS 1251151.

References

1. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine* 2014; **33**:209–218.
2. Gruber S, Tchetgen Tchetgen E. Limitations of empirical calibration of p-values using observational data. *Statistics in Medicine* 2016; **35**(22):3869–3882.
3. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Safety* 2013; **36**(Suppl 1):S59–S72.
4. Bruns SB, Ioannidis JP. p-Curve and p-hacking in observational research. *PLoS One* 2016; **11**:e0149144.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.