Published in final edited form as:

Stat Med. 2016 September 30; 35(22): 3869–3882. doi:10.1002/sim.6936.

Limitations of empirical calibration of p-values using observational data

Susan Gruber^{a,*} and Eric Tchetgen Tchetgen^b

^aReagan-Udall Foundation for the Food and Drug Administration, Washington, DC

^bDepartments of Biostatistics and Epidemiology, Harvard School of Public Health, Boston, MA

Abstract

Controversy over non-reproducible published research reporting a statistically significant result has produced substantial discussion in the literature. P-value calibration is a recently proposed procedure for adjusting p-values to account for both random and systematic error that addresses one aspect of this problem. The method's validity rests on the key assumption that bias in an effect estimate is drawn from a normal distribution whose mean and variance can be correctly estimated. We investigated the method's control of type-I and type-II error rates using simulated and real world data. Under mild violations of underlying assumptions control of the type-I error rate can be conservative, while under more extreme departures it can be anti-conservative. The extent to which the assumption is violated in real world data analyses is unknown. Barriers to testing the plausibility of the assumption using historical data are discussed. Our studies of the type-II error rate using simulated and real-world electronic healthcare data demonstrated that calibrating p-values can substantially increase the type-II error rate. The use of calibrated p-values may reduce the number of false positive results, but there will be a commensurate drop in the ability to detect a true safety or efficacy signal. While p-value calibration can sometimes offer advantages in controlling the type-I error rate, its adoption for routine use in studies of real-world healthcare datasets is premature. Separate characterizations of random and systematic errors provides a richer context for evaluating uncertainty surrounding effect estimates.

Keywords

p-value calibration; OMOP; safety; pharmacovigilance; p-value

1. Introduction

Controversy over non-reproducible published research that reports a statistically significant result is currently a hot topic in the literature. In an essay titled *Why Most Published Research Findings Are False*, John Ioannidis suggested that over 50% of published scientific findings are actually incorrect.[1] Vigorous discussion ensued, and continues to this day. [2, 3, 4, 5] Although methods for quantifying the proportion of false findings remain under debate, with some estimates as low as 14% [3], there is agreement that multiple factors

*Correspondence to: Susan Gruber, Reagan-Udall Foundation for the FDA, 1025 Connecticut Ave., N.W., Suite 1000, Washington, D.C. 20036. sgruber@reaganudall.org.

contribute to the problem. These include publication policies that favor manuscripts reporting statistically significant results, and p-value hacking, where an investigator artificially manipulates p-values by, for example, reporting results from a model cherry-picked to give a statistically significant result, or reporting the lone statistically significant result from a suite of analyses, neither acknowledging nor adjusting for multiple testing. In addition to these behaviors, mathematical properties of statistical analyses can play a key role. A study's statistical power, and the prevalence of true and false null hypotheses in the field under study can each affect the proportion of false positive results in the literature. Of particular interest in a recent paper by Schuemie and co-authors (SRDSM) is the fact that residual bias in a study result will impact the reported p-value, and can lead to an erroneous conclusion regarding the statistical significance of a study result. [6]

SRDSM proposed to adjust p-values to account for both random and systematic error using a novel method they call p-value calibration. SRDSM presented the method in the context of secondary analysis of observational studies of electronic healthcare data. These data are typically collected for health or reimbursement purposes, but are also used to conduct comparative effectiveness research and safety analysis. Challenges in using these data include misclassification of exposures and outcomes, lack of information on key confounders, and other sources of bias. SRDSM documents an unexpectedly large number of statistically significant findings where no exposure-outcome association is thought to exist. The authors propose a calibrated p-value approach to significance testing that addresses this problem by imposing an alternative model for the null distribution of the test statistic that attempts to account for residual bias in effect estimates under the null hypothesis of no causal effect of exposure on the outcome variable. SRDSM assumes the bias in effect estimates is normally distributed, and use a reference set of drug-outcome pairs to estimate the mean and variance that parameterize the distribution. SRDSM produced evidence that the use of calibrated p-values can provide excellent control of the type-I error rate at nominal level α . A large majority of the drugs used in their analyses have been on the market for many years. Factors that affect sources of bias evolve over time, as does the availability of data on identified confounders. 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. Some of these changes will tend to reduce residual bias in study results, while others may tend to increase residual bias. It is not yet clear whether SRDSM's promising findings hold when p-value calibration is applied in studies of recently marketed drugs. An important open question is how p-value calibration performs when residual biases in effect estimates for the reference set of negative controls differ from bias in the estimate for the novel drug-outcome pair under study. We carried out simulation studies to investigate this question. (Section 2.1).

A second open question centers on the ability to detect a safety or efficacy signal truly present in the data. SRDSM did not evaluate the impact of p-value calibration on the type-II error rate, the probability of correctly rejecting a false null hypothesis. We simulated data to compare the number of correct rejections from hypothesis tests based on calibrated p-values and non-calibrated (raw) p-values when the method's assumptions are met, and when they are violated (Section 3.1). We also present type-II error rates in the application of p-value calibration to real world data generated by the Observational Medical Outcomes

Partnership[7] from cohort and case control studies (Section 3.2). For both study designs the use of calibrated p-values greatly increased the type-II error rate. Our findings indicate that while the use of calibrated p-values can often reduce the number of incorrect rejections of true null hypotheses, there is a concomitant reduction in correct rejections of false null hypotheses. We review the p-value calibration procedure to gain insight into the difficulty in obtaining a statistically significant result, even when a true signal is present.

1.1. Review of p-value calibration

SRDSM's p-value calibration method requires the analysis of a reference set of drug-outcome associations in addition to the analysis of interest. The reference set consists of n known negative controls, defined by SRDSM as drug-outcome pairs where risk for the outcome is unchanged by exposure to the drug. These risk estimates and standard errors are used to model the distribution of residual bias under the null. Estimated parameters of the null distribution are incorporated into the calculation of a revised test statistic used for hypothesis testing.

The method assumes that under the null each risk estimate on the linear scale is normally distributed, with an estimate-specific mean and variance. In the authors' notation, $y_i \sim N(\theta_i, \tau_i^2)$, θ_i can be interpreted as the true bias in the i th effect estimate, y_i . The bias is assumed to follow a normal distribution parameterized by common mean, μ , and variance σ^2 . The p-value calibration procedure estimates μ and σ^2 by maximum likelihood, defined in

SRDSM as $L(\mu, \sigma | \theta, \tau) = \prod_{i=1}^n \int p(y_i | \theta_i, \tau_i) p(\theta_i | \mu, \sigma) d\theta_i$.

Consider an effect estimate of a novel hypothesis, y_{n+1} . Let γ be the limiting value of y_{n+1} under the null hypothesis of no treatment effect. Without loss of generality, we follow SRDSM in assuming that under the null, $\gamma = 0$. A standard Wald-type test statistic is calculated as, $t_{n+1} = y_{n+1} / \tau_{n+1}$. A 2-sided p-value is calculated as $2\Phi(-|t_{n+1}|)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Contrast this with the proposed approach. Under the null, y_{n+1} is assumed to be distributed as $N(\mu, \sigma^2 + \tau_{n+1}^2)$. To assess statistical significance the calibrated 2-sided p-value is given by $2\Phi(-|t'_{n+1}|)$, where the revised test statistic, t'_{n+1} , is calculated by performing shifting and scaling operations,

$$t'_{n+1} = \frac{y_{n+1} - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \tau_{n+1}^2}}.$$

We can gain insight into p-value calibration by examining the equation for the revised test statistic. First consider the variance terms in the denominator. Both $\hat{\sigma}^2$ and τ_{n+1}^2 are greater than 0, thus the denominator of t'_{n+1} is always larger than the denominator of t_{n+1} . Inflating the denominator always tends to shift the test statistic toward the null. In contrast, the scaling operation in the numerator can tend to shift the test statistic in either direction. When bias on average is not in a specific direction, i.e., $\hat{\mu} \approx 0$, t'_{n+1} will necessarily be closer to the null than t_{n+1} , and the calibrated p-value will always be larger than the raw p-value, $p_{cal} > p_{raw}$.

When estimated mean bias under the null is non-zero the situation is more nuanced. Consider a situation where bias is in the positive direction, $\hat{\mu} > 0$, and y_{n+1} is also positive.

Case 1: $y_i \approx \hat{\mu}$: The numerator of t'_{n+1} shifts towards 0, and thus $p_{cal} > p_{raw}$. Using calibrated p-values instead of raw p-values will result in fewer rejections of 2-sided tests of the null hypothesis.

Case 2: $y_i \ll \hat{\mu}$: The numerator of t'_{n+1} shifts to the left of 0, becoming negative. Using calibrated p-values instead of raw p-values will result in fewer tests of one-sided hypotheses of no increase in risk. However, the affect on 2-sided tests of null hypotheses is unclear, since a sufficiently large shift in the numerator could produce a revised test statistic with an extreme negative value.

Case 3: $y_i \gg \hat{\mu}$: The numerator of t'_{n+1} will shift slightly towards 0. For small enough shifts, raw and calibrated p-values could reject approximately the same number of false null hypotheses, particularly when $\hat{\sigma}^2$ is small.

Parallel arguments hold when bias is in the negative direction, and when y_{n+1} is negative. In summary, when $\hat{\mu}$ and y_{n+1} have the same sign, shifting the numerator of the test statistic moves its value towards 0. When $\hat{\mu}$ and y_{n+1} have opposite signs, shifting the numerator moves its value away from 0. The inflated denominator of the revised test statistic t'_{n+1} always tends to move its value closer to the null than t_{n+1} . The use of calibrated p-values can be expected to typically produce fewer 2-sided rejections than the use of raw p-values, regardless of whether there is a true effect.

2. Impact of p-value calibration on the type-I error rate

The set of negative controls used to model the null distribution of bias could either contain drugs known not to be causally associated with the outcome of interest, or, alternatively, a set of outcomes known to be causally unassociated with the drug, or exposure, of primary interest. SRDSM argues in favor of relying on negative exposure controls, because exposures are more easily and accurately ascertained from data than outcomes. For monitoring drug safety, the typical concern is whether a newly marketed drug increases risk for a particular adverse outcome, Y . Calibrating p-values necessitates analyzing a set of drugs for which exposure is not associated with a change in risk for Y . The method's success rests on the key assumption that under the null, residual bias in the effect estimate is drawn from the same distribution as the residual bias in the set of negative controls. If p-value calibration is to be used in practice, it is important to understand its robustness under violations of this assumption, and how often the assumption holds in real world data analyses.

The extent to which the distribution of any residual bias in effect estimates is similar to the bias in the effect estimate of a newly marketed drug is unknown. For example, factors such as selection bias and exposure misclassification might differentially impact analyses of established and novel drugs. Consider that drugs in the set of negative controls will necessarily have been on the market long enough for evidence to accumulate. If in that time an alternative treatment option emerge that is superior for a sub-population defined by

covariates that are not measured, differential selection into the exposure group could bias an effect estimate. Exposures to an over-the-counter counterpart of the drug would not be captured in administrative data, and therefore difficult to adjust for in a statistical analysis. One could construct more examples. The important open question is the extent to which the distribution of residual bias in the effect estimates from the set of negative controls differs from the truth. A reasonable approach to investigating this question could use historical real world data on known negative controls. The historical data would be collected within the first few years following a drug's approval at time t_0 , before its status as a negative control for the outcome was firmly established. This drug-outcome pair would serve as the novel hypothesis under study. Data on negative controls whose status had been established prior to t_0 could be used to estimate the distribution of bias under the null, and the calibrated p-value could be calculated. Carrying out this procedure for multiple drugs would yield a large number of calibrated p-values. The type-I error rate could be calculated as the proportion of calibrated p-values $\leq \alpha$. Unfortunately, sufficient real world data are not available for this exercise. Many large electronic healthcare datasets extend back only to the early 2000s. [8] However, only six of the 31 drugs identified by SRDSM as negative controls for acute liver injury have been approved since then (Appendix A). Being able to distinguish between a 1% and a 2% rejection rate requires a minimum of 100 hypotheses. With only six novel hypotheses, the smallest observable (non-zero) rejection rate would be 17%. This investigation would only provide coarse information on the type-I error rate, and is deferred until more suitable data become available. We can, however, use simulation studies to verify that p-value calibration provides effective control of the type-I error rate when assumptions are met, and to learn what can happen when they are violated.

2.1. Simulation Study I

Simulation study I was designed to compare control of the type-I error rate provided by calibrated p-values and raw p-values when the method's assumptions are met (study Ia) and when they are violated (study Ib). The diagram in Fig. 1 shows the causal structure of the data for study Ia. In the diagram, L is the set of baseline covariates predictive of both exposure and outcome, A is the drug of interest, B_{known} represents any drug in the set of known negative controls used to estimate the null distribution of the bias, and Y is the adverse outcome. The absence of an arrow from B_{known} into Y indicates that the outcome is unaffected by exposures B_{known} . U_1 represents one or more unmeasured confounders of the associations between B_{known} and Y . U_1 also confounds the association between A and Y . U_2 is a proxy for one or more additional unmeasured covariates that confound the relationship between B_{known} and Y . In this first simulation study U_2 also confounds the relationship between A and Y . The arrows from U_1 and U_2 into A indicate that each causally effects the exposure of interest. The arrows directly from U_1 and U_2 into Y indicate that each has a direct causal effect on the outcome. The dashed line from A to Y represents the causal hypotheses under investigation, i.e., the associations for which we will obtain effect estimates, standard errors, p-values, and calibrated p-values. There is no causal association between B and Y , nor between A and Y in the data generated under this DAG for the simulation study.

Heterogeneity in residual bias across the set of negative controls is a function of U_1 and U_2 . U_1 is normally distributed with variance = 1, and U_2 follows a Uniform distribution bounded by (a, b) , with variance = $1/12(b - a)^2$. The random draw from these distributions for each drug in the set B_{known} gives rise to heterogeneity in the bias in the set of negative controls. The DAG makes clear that residual bias in the estimate of the effect of drug A on Y stems from the same sort of confounding by U_1 and U_2 as estimates concerning the effect of B_{known} on Y . Thus, B_{known} provides an ideal negative control set for investigating the effect of A on Y .

All data analyses adjust for the effect of the known confounder, but not for unmeasured covariates U_1 and U_2 . We used a main terms logistic regression of Y on L and a binary treatment indicator to obtain biased log odds ratio estimates and associated standard errors for 50 known negative exposure controls in the set B_{known} , based on 50 independent samples of size of $n = 10,000$. These values were used to model the reference null distribution necessary for calibrating the p-values. P-value calibration was carried out using the R code available as online supplementary materials for SDRSM. Briefly, their procedure invokes a non-linear optimization function to find the maximum likelihood estimates $(\hat{\mu}, \hat{\sigma})$ from data consisting of the log odds ratios and associated standard error estimates from each analysis of the drug-outcome pairs in the set of negative controls. The effect of novel drug A on the identical outcome, Y , was estimated by a logistic regression of Y on A and L . To assess the type-I error rate a single set of 50 known controls was used to estimate the parameters of the null distribution of the bias, and 1000 $A - Y$ pairs were generated. Raw and calibrated p-values were calculated as described in subsection 1.1.

The data were generated according to the following simulation scheme,

$$L \sim N(1, 1)$$

$$U_1 \sim N(0, 1)$$

$$U_2 \sim \text{Unif}(x_1, x_2)$$

$$\text{logit}[P(B=1|L, U_1, U_2)] = 1 - 0.1L - 0.4U_1 - U_2$$

$$\text{logit}[P(A=1|L, U_1, U_2)] = 1 - 0.1L - 0.4U_1 - U_2$$

$$\text{logit}[P(Y=1|L, U_1, U_2)] = -1 + 0.2L + U_1 - U_2$$

Mild bias was simulated by bounding U_2 between $(x_1, x_2) = (1, 2)$. A second set of results were obtained for moderate to extreme bias by setting the bounds on U_2 to $(x_1, x_2) = (1, 6)$. The observed finite sample bias in the set of known negative controls used to estimate the null bias distribution averaged 0.16 on the log scale in study Ia (min = 0.02, max = 0.28) and 1.08 in study Ib (min = 0.74, max = 1.31). On the odds ratio scale this is equivalent to a mean bias of 0.18 for study Ia (min = 0.01, max = 0.32), and a mean bias of 1.96 for study Ib (min = 1.10, max = 2.73).

Logistic regression of Y on A and L was used to estimate the log odds ratio for all drug-outcome pairs. As in SRDSM, estimates and standard errors on the log scale were used to calculate raw and calibrated p-values. Rejection rates based on hypothesis tests at level $\alpha = 0.05$ are shown in Table 1. Source code for all simulations is available as a supplementary web appendix.

For study Ib the causal structure of the data was slightly modified so that A and B_{known} are not subject to identical sources of unmeasured confounding (Fig. 2). Although there is still an arrow from U_2 into B_{known} , the arrow from U_2 into A has been removed. The data were again generated such that there is no true causal association between B_{known} and Y , or between A and Y . The dashed line from A to Y represents the causal hypothesis under investigation. The only modification to the data generating process described above is that the probability of receiving the novel drug, A , is no longer a function of U_2 . Instead, $\text{logit}[P(A = 1 | L, U_1, U_2)] = 1 - 0.1L - 0.4U_1$. The residual bias in an estimate of the association between A and Y is not drawn from the same distribution as the residual bias affecting the associations between B_{known} and Y . Therefore, B_{known} is no longer an ideal negative control for investigating the effect of A on Y .

Logistic regression of Y on A and L was used to estimate the log odds ratio for 1000 novel drug-outcome pairs. Rejection rates based on hypothesis tests at level $\alpha = 0.05$ are shown in Table 1.

Results—Table 1 shows the proportion of p-values below the cutoff, $\alpha = 0.05$, for each study. Ideally, this proportion would equal 0.05. In study Ia calibrated p-values provided better control of type-I errors than raw p-values, coming close to the nominal rejection rate. In study Ib raw p-values performed better than they did in study Ia. This is because the novel drugs tested in study Ib were subject to less unmeasured confounding than those tested in study Ia. In contrast, calibrated p-values were anti-conservative. Their use resulted in poor control of the type-I error rate, worsening to 100% when there was medium bias due to U_2 . For Study Ia where assumptions are met, the average numerator of the calibrated test statistics is quite close to 0. In Study Ib, where assumptions are not met, the average numerator of the calibrated test statistic is further away from zero than the average numerator of the raw test statistics. In contrast, the denominator of the test statistics are identical to the third decimal place. In this simulation study the shift in the numerator explains nearly all of the differences between raw and calibrated p-values.

3. Impact of p-value calibration on the type-II error rate

Recall that the revised test statistic t' is obtained by subtracting the estimated mean bias, $\hat{\mu}$, from effect estimate y_i in the numerator, and dividing by a standard error estimate that is inflated to account for the estimated variance of $\hat{\mu}$. The shift in the numerator could move the test statistic either closer to or further away from 0. The inflated denominator always tends to move the revised test statistic closer to 0 than the original. A calibrated p-value that is greater than its raw counterpart is less likely to result in correct rejection of a false null hypothesis. In this case we'd expect to see an increase in the type-II error rate. A simulation study confirms that this increase can occur when assumptions are met as well as when they are violated (subsection 3.1. We also present results when p-value calibration is applied to analyses of real-world data (subsection 3.2). A dramatic decrease in the rejection rate was observed across nearly all of the outcomes studied by OMOP using cohort or case control study designs.

3.1. Simulation Study II

The data generating procedure used for this simulation study is nearly the same as above. The only difference is that in this study all the simulated novel null hypotheses are false. Thus, every failure to reject the null hypothesis contributes to the type-II error rate. L , U_1 , U_2 , A , and B_{known} were generated as above. In this study of positive control test cases, since A affects Y , the outcome was generated as $\text{logit}[P(Y = 1 | L, U_1, U_2)] = -1 + 0.2L + U_1 - U_2 + \beta A$. The value of beta was set to approximate the mean value of the bias in the negative control analyses under both mild bias ($\beta = 0.16$) and medium bias ($\beta = 0.9$). This corresponds to true odds ratios of 1.17 and 2.46, respectively. Confounding of the effect of B_{known} on Y is the same as in study I. For study IIa the assumptions underlying p-value calibration were met. The 1000 test associations between A and Y were subject to the same unmeasured confounding as the set of negative controls. For study IIb, the 1000 test drug-outcome associations were not subject to the same sources of bias. U_2 did not confound the effect of these test drugs on Y . In both study IIa and IIb, logistic regression of Y on A and L was used to estimate the log odds ratio for 1000 novel drug-outcome pairs. The same analysis was applied to the 50 negative controls (B_{known} - Y pairs). As in SRDSM, estimates and standard errors on the log scale were used to calculate raw and calibrated p-values. We evaluated rejection rates based on two-sided tests of the null hypothesis of no treatment effect. We also evaluated rejection rates based on one-sided tests of the null hypothesis that treatment does not increase risk for the outcome.

3.1.1. Results—Type-II error rates for one- and two-sided hypothesis tests are shown in Table 2. Higher values are better, with 1 corresponding to correctly rejecting all of the null hypotheses. Hypothesis testing based on raw p-values consistently achieved the highest rejection rate. Simulation studies IIa and IIb confirm that in some circumstances p-value calibration can substantially reduce rejection rates (mild bias, one- and two-sided tests, regardless of whether assumptions are met (study IIa) or violated (study IIb).) When the effect size is small and bias is in the positive direction, as it is here, a one-sided hypothesis test will typically reject more null hypotheses than a two sided-test. As anticipated from the mathematical analysis of the revised test statistic, when the mean estimated bias is much

larger than the effect size estimate, y_i , the revised test statistic can take on a sizable negative value. Simulation study IIb demonstrated that when the bias is large and the novel drug is not subject to all the same sources of bias, a one-sided test using calibrated p-values can fail to reject all false null hypotheses.

3.2. Application to real-world data

We applied p-value calibration to estimates obtained from OMOP cohort studies and case control studies of Truven Marketscan Medicaid data [7]. OMOP defined variants of four outcomes, acute liver failure (ALF), acute renal failure (ARF), myocardial infarction (MI), and hospitalization for upper gastrointestinal bleed (UGI Hosp). OMOP defined sets of negative controls for each outcome, and sets of *positive* controls, drugs that increase risk for the outcome of interest.

Our procedure for calibrating p-values was the following. For a fixed study design and for each outcome in turn, relative risk and variance estimates from all the negative control analyses were used to estimate the parameters of the bias distribution under the null. Next, calibrated 2-sided p-values were calculated for the drugs labeled by OMOP as *positive* controls. These are drugs identified as increasing risk for the outcome of interest. Figure 3(a) allows us to compare the proportion of p-values $\alpha = 0.05$ for 15 variant outcome definitions established by OMOP [9], ranging from narrow to more broadly defined. The number of positive and negative controls for each study design and outcome for which data were available is summarized in Table 3. A detailed list of controls for each of the four outcomes is provided in Appendix B. Figure 3(b) shows rejection rates when testing the one-sided null hypothesis that exposure increases risk for the outcome. Rejection rates when the same procedure was applied to case-control study results using the same data are shown in Figure 3(c) and (d). A rejection is the hoped-for result when testing a hypothesis concerning a positive control. Therefore, the procedure that rejects more null hypotheses is exhibiting better performance. The plots show stark differences in rejection rates. Rates using raw p-values are substantially higher than using calibrated p-values. In fact, hypothesis testing based on calibrated p-values rejected approximately 5% of all false null hypotheses among cohort study results and among case control study results. For some outcomes performance was better, rejecting nearly 20% of false null hypotheses among case-control study results for Acute Liver Failure, definition 4. Within each study design there were some outcomes where p-value calibration failed to reject any false null hypotheses. In contrast, rejection rates based on raw p-values were always above 15%, and reached as high as 50%. The type-II error rate is calculated as 1 minus the rejection rate. A small type-II error rate is desirable. The type-II error rate was higher for calibrated p-values than for raw p-values in 59 out of 60 cases.

4. Alternative assessment of uncertainty

A p-value assesses the probability that under the posited model a value at least as extreme as an observed test statistic will be observed when the null hypothesis holds. SRDSM rightly point out that when systematic error contributes to the value of an observed test statistic, basing a hypothesis test on a raw p-value will typically not control the type-I error rate

(limited exceptions exist [10]). However, blending bias and variance into a single metric, a calibrated p-value, may be misguided. Separately characterizing systematic error and random error provides a richer context for understanding uncertainty in effect estimates. When there is residual bias, analytic standard estimates and bootstrap approaches can accurately estimate random error, but the estimate itself lacks interpretability. [11, 12]. Because variance shrinks as the number of observations grows, in 'big data' analyses, residual bias swamps variance as a primary concern. Understanding the likely magnitude and direction of the bias provides useful information for drawing a substantive conclusion from analytical results. Nuances are erased when bias and variance are encapsulated within a single scalar.

Relevant external knowledge can aid in detecting residual bias in effect estimates. Negative control outcomes can be used to detect suspected unobserved confounding. At times, such variables may be used in a formal counterfactual-based approach to correct causal effect estimates for bias due to unobserved confounding [13]. The control outcome calibration approach (COCA) is based on the simple observation that the exposure-free potential outcome is a perfect surrogate measure of the degree of unobserved confounding. Therefore under a rank preserving causal model for a continuous outcome, or a similar assumption for a binary outcome, the parameter for the model can be identified by obtaining a corresponding prediction of the exposure-free potential outcome, which, once conditioned on in the negative control outcome regression on exposure and observed confounders, recovers a null association with exposure. [13] The validity of this approach rests on the ability to identify one or more negative controls that are subject to identical sources of residual bias as the study pair. This requirement is more stringent than that imposed by SRDSM. When underlying assumptions are met, the estimated bias could be used to correct the initial estimate. [14, 15]

A necessary component of a valid negative control is its similarity to the target exposure of interest with respect to the sources of uncontrolled bias. Lipsitch, et. al. identified sufficient conditions for the use of such negative controls using causal directed acyclic graphs, and discussed their potential to improve epidemiologic inference. [14] The causal diagram in Fig. 4 shows an ideal negative exposure control, B_{known} , for studying the causal relationship between exposure A and outcome Y [14, Fig. 3]. Arrows depict potential causal associations, U is an unmeasured confounder, and L is a baseline covariate. The line between U and L indicates they are correlated. B_{known} is an ideal negative exposure control because all nodes that have arrows into A also have arrows into B_{known} . The absence of any other arrows into B_{known} indicates that no other confounder links B_{known} and Y . Unmeasured confounder U is a parent of both B_{known} and Y . An analyst interested in estimating the effect of B_{known} on Y could not adjust for U directly since it is unmeasured. However, adjusting for A would block the backdoor path from B to Y through U . Thus, the analysis of each negative control should include an adjustment for the exposure of interest.

When adequate negative controls are not available, sophisticated sensitivity analyses can bound effect sizes. Another way to examine the stability of the estimate is to investigate how strong the sources of bias would have to be to impact the substantive conclusion. An estimators robustness can be evaluated by performing hypothesis tests across levels of a

sensitivity parameter, that represents an overall degree of violation from statistical assumptions (e.g., unmeasured confounding, misclassification, and lack of overlap between exposed and unexposed patient characteristics). This approach can be applied either parametrically or non-parametrically across a broad range of study designs and effect parameters. [16, 17] For matching analyses, a similar approach is to establish Rosenbaum bounds that assess the strength of confounding required to undermine the conclusions about causal effects.[18]. Unlike p-value calibration these approaches cannot be fully automated. Nevertheless, their use may be justified in studies designed to produce actionable evidence to support regulatory or drug-development decision making.

5. Discussion

SRDSM presented p-value calibration as a method for improving control of the type-I error rate. In this paper we explored the performance of p-value calibration across a variety of settings. The goal was to better understand how departures from the untestable assumption underlying the method could affect control of type-I and type-II error rates.

Simulation study Ia confirmed that p-value calibration improves control of the type-I error rate when the method's key assumption is met. Simulation study Ib demonstrated that when this assumption is not met, when the bias in the estimate under study is not drawn from the same distribution as the bias in the negative control analyses, p-value calibration can provide anti-conservative control of the type-I error rate. Calibrated p-values were especially problematic when the novel drug-outcome analysis was not subject to the same sources of bias as the set of negative controls used to model the null distribution. When the key assumption is violated, calibrated p-values are not necessarily more valid than their raw counterparts. How often these departures occur in practice has not yet been established. Simulation studies IIa and IIb demonstrated that the type-II error rate will often increase when calibrated p-values are used for hypothesis testing. This finding was confirmed in our analysis of real-world data.

This paper did not address implementation issues that could affect performance, even when assumptions are met. Care must be taken to ensure that the optimization routine used to maximize the likelihood when estimating (μ, σ^2) converges, and that sensitivity to starting values is assessed. Another important component of the method is careful ascertainment of a sufficiently large set of true negative controls. The authors established the status of the known negative controls by looking at drug labels and published results. However, a drug that poses only a slight increase in risk for an outcome may not have been the subject of safety warnings, or multiple published studies. This does not rule out the possibility that the true relative risk is, although close to 1, nonetheless non-null, e.g., 1.1. A second issue is the number of negative controls used to establish the parameters of the normal null distribution. Appendix E of SRDSM suggests that as few as 25 negative controls is sufficient for estimating these parameters, and report a lack of sensitivity to relaxing the normality constraint. Empirically, 25 data points cannot serve to distinguish p-values in steps smaller than 1/25, or 0.04. At this sample size, finer gradations in calibrated p-values are entirely model driven. While a larger set of negative controls would help with this problem, in the context of drug safety, it's not clear how to go about finding a large set of appropriate

negative controls. That is, negative controls where the sources of bias are sufficiently like those that influence the target drug-outcome association. It is worth noting that in other settings this may be much less problematic. For example, negative control genes have been used to correct for batch effects in microarray expression studies, and spike-in controls have been used to normalize RNA sequencing data to remove unwanted effects introduced during sample processing. [19, 20]

Hypothesis testing when there is residual bias in an effect estimate is a complex undertaking. Calibrated p-values arguably provide better control of the type-I error rate than raw p-values, but weaker control of the type-II error rate. We suggest that separate assessments of systematic error and random error can provide more complete information. Under restricted conditions where sources of systematic error affecting controls and the drug-outcome pair under consideration can plausibly be deemed similar, accurate automated bias adjustment is possible. The SRDSM approach may provide better type I and II error control than raw p-values, particularly if the heterogeneity of bias is large (e.g. the true value of $\mu = 0$ and τ is large). We are not aware of any general, fully-automated method for assessing systematic error.

P-value calibration shows promise in reducing the type-I error rate. However, appropriate application of the method and careful interpretation of the results is necessary. The idea of using a reference set of negative controls to detect whether bias is likely to be a major concern has merit. Whether this approach can aid in discriminating between true and false safety signals is still unknown. SRDSM's recommendation that *observational studies always include negative controls to derive an empirical null distribution and use these to compute calibrated p-values* is premature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Tchetgen Tchetgen was supported by NIH grant AI104459.

The authors would like to thank George Hripcsak, Patrick Ryan, Christian Reich, Sara Dempster, Lisa Kammerman, Adler Perotte, Marc Suchard, and David Madigan for helpful discussions. We also thank anonymous reviewers, whose critiques of an earlier draft motivated important changes to the manuscript.

Appendix A

FDA Drug Approval Dates

FDA approval dates for 31 drugs listed as negative controls for Acute Liver Injury by SRDSM (source: OMOP vocabulary accessed in the IMEDS Research Lab, July 7, 2015). Entries in the table are sorted by approval date. Six drugs have been approved since mid-1999. However one of these, Noestigmine, has been used in the United States since 1931 and it's properties are well known to the medical community. [21]

Table A1

FDA Approval Date.

Drug Name	Approval Date	Drug Name	Approval Date
Ergotamine	26-Nov-1948	Sucralfate	30-Oct-1981
Dicyclomine	11-May-1950	Tetrahydrocannabinol	31-May-1985
Phentolamine	30-Jan-1952	Adenosine	30-Oct-1989
Propantheline	2-Apr-1953	Fluticasone	14-Dec-1990
Primidone	8-Mar-1954	Salmeterol	4-Feb-1994
Benzonatate	10-Feb-1958	Amylases	9-Dec-1996
Phentermine	4-May-1959	Lipase	9-Dec-1996
Methenamine	5-Jul-1967	Ketotifen	2-Jul-1999
Paromomycin	24-Mar-1969	Almotriptan	7-May-2001
Flavoxate	15-Jan-1970	Gatifloxacin	28-Mar-2003
Droperidol	11-Jun-1970	Tinidazole	17-May-2004
Miconazole	8-Jan-1974	Sitagliptin	16-Oct-2006
Oxybutynin	16-Jul-1975	Cosyntropin	21-Feb-2008
Lactulose	25-Mar-1976	Griseofulvin	8-Oct-2010
Scopolamine	31-Dec-1979	Neostigmine	31-May-2013
Lithium citrate	23-Dec-1980		

Appendix B**Controls used to Calculate Type-II Error Rates**

Drugs and outcomes shown in Table B1 were used as negative and positive controls in at least one cohort or case-control study analyzed by OMOP using Truven MarketScan Medicaid data.[7]

Table B1

Negative and positive drug-outcome control pairs used to calibrate p-values and calculate type-II error rates for at least one variant definition of each outcome.

Acute Liver Failure*negative controls*

Scopolamine	Tetrahydrocannabinol	fluticasone	oxybutynin	sitagliptin
Ketotifen	almotriptan	benzonatate	Amylases	Phentolamine
Lactulose	Paromomycin	Ergotamine	Hyoscyamine	Droperidol
Sodium Phosphate, Monobasic	Penicillin V	Miconazole	Dicyclomine	Methenamine
Lipase	Endopeptidases	gatifloxacin	ferrous gluconate	Primidone
Adenosine	Propantheline	Cosyntropin	salmeterol	Scopolamine
Phentermine	Flavoxate	Neostigmine	lithium citrate	Ketotifen
Sucralfate	Griseofulvin	Benzocaine	Tinidazole	Lactulose

positive controls

Methotrexate	trandolapril	abacavir	celecoxib	tolcapone
Methyldopa	Flutamide	efavirenz	Piroxicam	Ofloxacin
Lisinopril	Tamoxifen	terbinafine	Carbamazepine	gemcitabine
moexipril	Thioguanine	Erythromycin	Valproate	Nitrofurantoin
Nifedipine	Methimazole	Fluconazole	felbamate	pioglitazone
bosentan	Niacin	darunavir	Nevirapine	Zidovudine
Diltiazem	Propylthiouracil	Rifampin	Stavudine	Dacarbazine
quinapril	Itraconazole	Allopurinol	isoniazid	tipranavir
Ramipril	posaconazole	Ibuprofen	Ciprofloxacin	Didanosine
bortezomib	voriconazole	Indomethacin	Sulfisoxazole	Levofloxacin
Captopril	gemifloxacin	Etodolac	Thiabendazole	Busulfan
Nortriptyline	Caspofungin	Sulindac	Cyclosporine	Lamivudine
Interferon beta-1a	Norfloxacin	imatinib	alatrofloxacin	Clozapine
Disulfiram	Zalcitabine	zafirlukast	Penicillamine	Ketorolac
Tolmetin	Acetazolamide	Naproxen	lamotrigine	orlistat
Enalaprilat	infliximab	oxaprozin	nefazodone	Methotrexate

Acute Myocardial Infarction*negative controls*

Scopolamine	posaconazole	salmeterol	darifenacin	Tetrahydrocannabinol
Ketoconazole	Nelfinavir	Droperidol	Hyoscyamine	Thiabendazole
Ketotifen	Didanosine	Prochlorperazine	Temazepam	Chlorambucil
Lactulose	Paromomycin	lithium citrate	Penicillin V	Vitamin A
Sodium Phosphate, Monobasic	Dicyclomine	metaxalone	Propantheline	oxybutynin
Lipase	Acetazolamide	ramelteon	Thiothixene	bromfenac
Pemoline	Prilocaine	Chlorazepate	Amylases	Nevirapine
Chlorothiazide	Flavoxate	Methenamine	ferrous gluconate	Cosyntropin
Clindamycin	Sulfasalazine	Urea	Tinidazole	Mebendazole
Sucralfate	tipranavir	Miconazole	terbinafine	Primidone
Flutamide	fluticasone	gatifloxacin	entecavir	Scopolamine
Methimazole	Loratadine	Sulfisoxazole	Simethicone	Ketoconazole
Acarbose	zafirlukast	Penicillamine	Endopeptidases	Ketotifen
sitagliptin	benzonatate	Methocarbamol	Stavudine	Lactulose

positive controls

moexipril	Fenoprofen	Epoetin Alfa	Sumatriptan	Nortriptyline
Nifedipine	rizatriptan	almotriptan	Imipramine	Sulindac
Bromocriptine	Flurbiprofen	nabumetone	Dipyridamole	darbepoetin alfa
Tolmetin	Indomethacin	zolmitriptan	estropipate	Ketorolac
Enalaprilat	Ketoprofen	oxaprozin	Desipramine	moexipril
Factor VIIa	frovatriptan	naratriptan	Estrogens, Conjugated (USP)	Nifedipine
Estradiol	eletriptan	Diffunisal	Amlodipine	Bromocriptine
Piroxicam	Etodolac	Salsalate	Amoxapine	Tolmetin

Acute Renal Failure

negative controls

Scopolamine	Paromomycin	Imipramine	Hyoscyamine	Nelfinavir
Simethicone	Penicillin V	metaxalone	Thiothixene	almotriptan
Ketotifen	Endopeptidases	ramelteon	Benzocaine	Acarbose
Lactulose	Prilocaine	Chlorazepate	gatifloxacin	Urea
Sodium Phosphate, Monobasic	Flavoxate	Clozapine	rizatriptan	Temazepam
Adenosine	darunavir	Methenamine	Dicyclomine	Ketoconazole
Dacarbazine	Griseofulvin	Miconazole	Lipase	infliximab
Phentolamine	frovatriptan	Thiabendazole	benzonatate	Cosyntropin
Tetrahydrocannabinol	eletriptan	Vitamin A	Loratadine	Mebendazole
Flutamide	darbepoetin alfa	Methocarbamol	Prochlorperazine	orlistat
Chlorambucil	zafirlukast	Neostigmine	Nortriptyline	Primidone
ferrous gluconate	Ergotamine	darifenacin	entecavir	Scopolamine
Tinidazole	Disulfiram	Amylases	bromfenac	Simethicone

positive controls

Lisinopril	Acyclovir	oxaprozin	Hydrochlorothiazide	telmisartan
Olmesartan medoxomil	Allopurinol	Cyclosporine	Diflunisal	moexipril
Captopril	Ibuprofen	Naproxen	Piroxicam	Ketorolac
Enalaprilat	Etodolac	Ketoprofen	Fenoprofen	Lisinopril
candesartan	Mefenamate	meloxicam	Chlorothiazide	Olmesartan medoxomil

Hospitalization for Upper GI Bleed*negative controls*

Scopolamine	abacavir	darifenacin	sitagliptin	Paromomycin
Dacarbazine	Epoetin Alfa	Dicyclomine	Methocarbamol	Loratadine
Phentolamine	salmeterol	Simethicone	Nitrofurantoin	Neostigmine
Phentermine	Ergotamine	Stavudine	Chlorazepate	Lactulose
ferrous gluconate	Disulfiram	Chlorambucil	Lipase	terbinafine
pioglitazone	Droperidol	fluticasone	Benzocaine	Nevirapine
Acarbose	lithium citrate	Amylases	Sucralfate	Tetrahydrocannabinol
rosiglitazone	ramelteon	Hyoscyamine	entecavir	Lamivudine
Itraconazole	Temazepam	Ketoconazole	metaxalone	Cosyntropin
Zidovudine	Urea	oxybutynin	Tinidazole	Mebendazole
Penicillin V	Griseofulvin	Adenosine	Prochlorperazine	orlistat
Endopeptidases	Thiabendazole	benzonatate	bromfenac	Scopolamine
Prilocaine	Vitamin A	moexipril	Pemoline	Dacarbazine
Propantheline	Thiothixene	Miconazole	Ketotifen	Phentolamine

positive controls

clopidogrel	Ibuprofen	Naproxen	meloxicam	oxaprozin
Clindamycin	Indomethacin	Sertraline	Flurbiprofen	Ketoprofen
Tolmetin	Mefenamate	Fluoxetine	valdecoxib	Etodolac
Piroxicam	Sulindac	Citalopram	Diflunisal	Ketorolac
Fenoprofen	nabumetone	Potassium Chloride	Escitalopram	clopidogrel

References

1. Ioannidis J. Why most published research findings are false. *PLoS Med.* 2005; 2:e124. [PubMed: 16060722]
2. Goodman SN. Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics.* 2013 kxt035.
3. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics.* 2014; 15(1):1–12. [PubMed: 24068246]
4. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science.* 2014; 1(3):140 216.
5. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. *PLoS Biol.* 2015; 13(3):e1002 106.
6. 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(2): 209–218. [PubMed: 23900808]
7. OMOP. Observational medical outcomes partnership research 2013. [Accessed: 1/31/2016] URL <http://omop.org/Research>.
8. Mini-Sentinel Data Core. MINI-SENTINEL DISTRIBUTED DATABASE YEAR 4 SUMMARY REPORT Version 1.1. 2014 Aug. [Accessed: 1/31/2016] URL http://mini-sentinel.org/data_activities/distributed_db_and_data/default.aspx.
9. OMOP. Observational medical outcomes partnership 2013. [Accessed: 1/31/2016] URL <http://omop.org>.
10. Bross I. Misclassification in 2×2 tables. *Biometrics.* 1954; 10:478–486.
11. Efron, B.; Tibshirani, RJ. *An Introduction to the Bootstrap.* Chapman & Hall/CRC; 1993.
12. Freedman DA. On the so-called Huber sandwich estimator and robust standard errors. *The American Statistician.* 2006; 60(4):299–302.
13. Tchetgen Tchetgen E. The control outcome calibration approach (COCA) for unobserved confounding. *American Journal of Epidemiology.* 2013; 179:633–640. [PubMed: 24363326]
14. Lipsitch M, Tchetgen Tchetgen E, Cohen T. The use of negative controls to detect confounding and other sources of error in experimental and observational science. *Epidemiology.* 2010:383–388. [PubMed: 20335814]
15. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative exposure controls in epidemiologic studies. *Epidemiology.* 2012:351–352.
16. Rotnitzky A, Scharfstein D, Su S, Robins J. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics.* 2001; 57:103113.
17. Diaz I, van der Laan MJ. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *International Journal of Biostatistics.* 2013; 9:149160.
18. Rosenbaum, P. *Observational Studies.* 2nd. New York: Springer; 2002.
19. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics.* 2012; 13:539–552. [PubMed: 22101192]
20. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology.* 2014; 32:896–902.
21. [Accessed: 09/02/2014] Neostigmine methylsulfate bloxiverz clinical review. <http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DevelopmentResources/UCM361414.pdf>

Figure 1. Causal diagram showing causal relationships among measured and unmeasured covariates, (L, U_1, U_2) , known negative exposure control, B_{known} , used to estimate the null distribution, and negative exposure test case, A for study Ia.

Figure 2. Causal diagram showing causal relationships among measured and unmeasured covariates, (L, U_1, U_2) , known negative exposure control, B_{known} , used to estimate the null distribution, and negative exposure test case, A for study Ib. The absence of an arrow between U_2 and A indicates there is no causal association.

Figure 3.

Comparison of rejection rates for hypothesis tests at level $\alpha = 0.05$ based on calibrated and raw p-values. All test cases represent false null hypotheses. Rates are shown for testing 2-sided hypotheses concerning cohort study results (a), 1-sided hypotheses concerning cohort study results (b), 2-sided hypotheses concerning case-control results (c), and 1-sided hypotheses concerning case-control studies (d).

Figure 4.
Causal diagram showing negative exposure control set, B_{known} , for investigating the effect of exposure A on outcome Y .

Table 1

Simulation study I results using standard and calibrated p-values: observed Type-I error rates at nominal level $\alpha = 0.05$, and (minimum, mean, maximum) observed values of the numerator and denominator of the test statistics

Rejection Rate		Numerator (min, mean, max)		Denominator (min, mean, max)	
raw	calibrated	raw	calibrated	raw	calibrated
Study Ia (assumptions are met)					
<i>mild bias</i>	0.579	0.059	(-0.052, 0.15, 0.394)	(-0.210, -0.008, 0.236)	(0.066, 0.070, 0.074)
<i>med bias</i>	1.000	0.061	(0.714, 1.101, 1.512)	(-0.432, -0.046, 0.365)	(0.115, 0.128, 0.143)
Study Ib (assumptions are not met)					
<i>mild bias</i>	0.346	0.399	(-0.079, 0.077, 0.284)	(-0.237, -0.081, 0.126)	(0.047, 0.048, 0.049)
<i>med bias</i>	0.346	1.000	(-0.079, 0.077, 0.284)	(-1.225, -1.069, -0.862)	(0.047, 0.048, 0.049)

Table 2

Rejection rates using one and two-sided raw and calibrated p-values ($\alpha = 0.05$), for small and medium effect sizes with mild and moderate bias.

	Rejection rate of 2-sided test		Rejection rate of 1-sided test	
	raw	calibrated	raw	calibrated
<i>Study IIa</i> (same sources of bias)				
<i>mild bias</i>				
OR = 1.17	0.995	0.585	0.997	0.694
OR = 2.46	1	1	1	0.983
<i>med bias</i>				
OR = 1.17	1	0.753	1	0.829
OR = 2.46	1	1	1	0.997
<i>Study IIb</i> (different sources of bias)				
<i>mild bias</i>				
OR = 1.17	1	0.349	1	0.466
OR = 2.46	1	1	1	0.945
<i>med bias</i>				
OR = 1.17	1	1	1	0
OR = 2.46	1	0.154	1	0

Number of negative and positive controls used to calibrate p-values and calculate rejection rates for each study design and outcome.

Table 3

Outcome	Definition	Number of Controls			
		Cohort Study Design		Case Control Design	
		negative	positive	negative	positive
Acute Liver Failure	1	36	37	33	73
	2	35	57	31	72
	3	30	20	28	63
	7	32	35	30	69
	8	34	36	33	73
Acute Myocardial Infarction	1	63	36	59	34
	2	62	42	55	33
	3	62	46	55	34
	5	62	39	56	33
	1	59	23	59	21
Acute Renal Failure	2	54	22	52	21
	4	59	25	58	23
	5	59	28	60	21
	1	63	33	61	24
	3	65	32	62	24
Hospitalization for Upper GI Bleed					