Title

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Permalink

https://escholarship.org/uc/item/5sj9x8tw

Journal

Proceedings of the National Academy of Sciences of the United States of America, 115(11)

ISSN

0027-8424

Authors

Schuemie, Martijn J
Hripcsak, George
Ryan, Patrick B
et al.

Publication Date

2018-03-13

DOI

10.1073/pnas.1708282114

Peer reviewed

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California, Los Angeles, CA 90095

Edited by Victoria Stodden, University of Illinois at Urbana–Champaign, Champaign, IL, and accepted by Editorial Board Member Susan T. Fiske October 26, 2017 (received for review June 15, 2017)

Observational healthcare data, such as electronic health records and administrative claims, offer potential to estimate effects of medical products at scale. Observational studies have often been found to be nonreproducible, however, generating conflicting results even when using the same database to answer the same question. One source of discrepancies is error, both random caused by sampling variability and systematic (for example, because of confounding, selection bias, and measurement error). Only random error is typically quantified but converges to zero as databases become larger, whereas systematic error persists independent from sample size and therefore, increases in relative importance. Negative controls are exposure–outcome pairs, where one believes no causal effect exists; they can be used to detect multiple sources of systematic error, but interpreting their results is not always straightforward. Previously, we have shown that an empirical null distribution can be derived from a sample of negative controls and used to calibrate *P* values, accounting for both random and systematic error. Here, we extend this work to calibration of confidence intervals (CIs). CIs require positive controls, which we synthesize by modifying negative controls. We show that our CI calibration restores nominal characteristics, such as 95% coverage of the true effect size by the 95% CI. We furthermore show that CI calibration reduces disagreement in replications of two pairs of conflicting observational studies: one related to dabigatran, warfarin, and gastrointestinal bleeding and one related to selective serotonin reuptake inhibitors and upper gastrointestinal bleeding. We recommend CI calibration to improve reproducibility of observational studies.

observational studies | systematic error | calibration

Observational healthcare data, such as electronic health records and administrative claims, offer the potential to estimate the effects of various medical product exposures on many health-related outcomes of interest. Population-level effect estimation has many applications throughout healthcare, including safety surveillance by product manufacturers and regulatory agencies and evaluation of comparative effectiveness for payers and providers.

A critical challenge limiting the acceptance of observational data as part of any causality assessment is the inherent uncertainty around how much we can trust the results from nonrandomized experiments. There have been many observational studies that proved to be nonreproducible (1). Failure to reproduce results likely stems from error in the original study, in the replication, or in both. Error, the difference between true and estimated effect sizes, can be decomposed in two components: random error and systematic error. Random error arises from sampling variability and is commonly reflected in most study statistics through an estimate of variance and some calculation of a confidence interval (CI) around the point estimate of the average treatment effect. Systematic error can manifest from multiple sources, including confounding, selection bias, and measurement error. While there is widespread awareness of the potential for systematic error in observational studies and a large body of research that examines how to diagnose and statistically adjust for specific sources of bias, there has been comparatively little work in devising approaches to empirically estimate the magnitude of systematic error or clinical applications that show how to integrate this error into effect estimation methods.

The acuity of this problem is only exacerbated as the size of observational databases grow: random error (the only component that is typically quantified) converges to zero as sample size increases, but systematic error persists independent from sample size. Some sources of systematic error may potentially increase if expanding the size of a data source comes with compromise in the depth or quality of the data captured. Therefore, the hype of "big data" has brought with it an increased number of studies with vanishingly narrow CIs, while our collective uncertainty about the accuracy of any given observational estimate has steadily increased. While we expect an accurate 95% CI to have a 95% coverage probability—the proportion of time that an interval contains the true value of interest—we have little empirical evidence to support that observational estimates exhibit this basic, nominal operating characteristic.

A promising development toward better explication of systematic error has been recent proposals and examples to apply negative controls as a diagnostic tool or "falsification hypothesis" (2–4). Negative controls are exposure–outcome pairs where one believes no causal effect exists. Executing a study on negative controls and determining whether the results indeed show no effect

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: M.J.S., G.H., P.B.R., D.M., and M.A.S. designed research; M.J.S. performed research; M.J.S. and M.A.S. contributed new reagents/analytic tools; M.J.S. and M.A.S. analyzed data; and M.J.S., G.H., P.B.R., D.M., and M.A.S. wrote the paper.

Conflict of interest statement: M.J.S. and P.B.R. are full-time employees and shareholders of Janssen Research & Development.

This article is a PNAS Direct Submission. V.S. is a guest editor invited by the Editorial Board.

Data deposition: The ISAC application forms have been deposited on GitHub and are available at https://github.com/OHDSI/StudyProtocols/blob/master/CiCalibration/extras/ISACapplication.zip.

[1]To whom correspondence should be addressed. Email: schuemie@ohdsi.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1708282114/-/DCSupplemental.

Published online March 12, 2018.

can help detect bias inherent to the study design or data used. This can include bias from multiple sources, such as confounding (5), selection bias, and measurement error (6). For example, in one study (7) investigating the relationship between childhood diseases and later multiple sclerosis, the authors include three negative controls that are not believed to cause multiple sclerosis: a broken arm, concussion, and tonsillectomy. Two of these three controls produce statistically significant associations with multiple sclerosis, suggesting that the study may be biased. However, an open question is how to interpret findings from these negative controls. In this example, should we consider all study results to be invalid? Or, as the authors in this case do, should only effect sizes greater than those observed for the controls be considered to be true effects? Alternatively, is it possible that rejection of the null hypothesis for the two negative controls is caused by random chance alone and that the study is, in fact, unbiased?

One path forward is to incorporate the error observed for negative controls into the estimates of observational studies, in effect calibrating the estimates in a way similar to how one would calibrate a scale using objects of known weight. Other researchers have proposed such calibration methods using a single negative control (8, 9), but these approaches require the negative control to have identical systematic error as the effect under study, which is unlikely to be true for any negative control and will always be unknowable. In our prior work, we have shown that estimates from a sample of negative controls can be used to derive an empirical null distribution that can be applied to unknown effect estimates to calibrate P values in a manner that accounts for both random and systematic error (10, 11). In effect, our empirical calibration aims to restore nominal operating characteristics (for example, having only 5% of negative controls return a calibrated P value < 0.05).

In this paper, we extend our notion of empirical calibration to improve the validity of CIs in observational studies. We describe a statistical procedure for CI calibration that uses negative controls as well as positive controls. Positive controls are exposure–outcome pairs, where a causal effect of known magnitude is believed to exist. For reasons explained in *Materials and Methods*, we use synthetic positive controls constructed by modifying real negative controls. We show the procedure using three large observational databases to evaluate internal validity and show how the procedure restores the nominal operating characteristics, such that a 95% CI achieves a 95% coverage probability. We further illustrate the impact of the procedure by replicating two pairs of conflicting observational studies and examining the extent to which calibration identifies bias within each design and resolves the apparent inconsistency reported in the literature. The two pairs of studies are as follows.

i) Dabigatran vs. warfarin drug exposure for the risk of gastrointestinal (GI) bleeding outcomes as performed by Southworth et al. (12) compared with the study by Graham et al. (13). Both studies used a new user cohort design, but only Graham et al. (13) used propensity scores to adjust for potential confounding between exposure and outcome. The incidence rate ratio implied by Southworth et al. (12) was 1.6/3.5 = 0.46. The hazard ratio (95% CI) reported by Graham et al. (13) was 1.28 (1.14–1.44).

ii) Selective serotonin reuptake inhibitor (SSRI) drug exposure for the risk of upper GI bleeding outcomes in two studies performed by Tata et al. (14). The first study used a case–control (CC) design, and the second used a self-controlled case series (SCCS) design. The CC analysis produced an odds ratio (95% CI) of 2.38 (2.08–2.72). The SCCS produced an incidence rate ratio (95% CI) of 1.71 (1.48–1.98).

Results

Negative Controls. For each study, we select 50 negative control outcomes using a semiautomated process as detailed in *Materials and Methods*. One of the negative controls for the dabigatran studies is "ingrowing nail," because we firmly believe that neither dabigatran nor warfarin exposure cause ingrowing nails, and we, therefore, believe that the true hazard ratio when comparing these two drugs for this outcome should equal one. We can now apply our observational study designs to confirm whether they produce estimates close to the truth. Note that both drugs could still be noncausally associated with ingrowing nails, which is acceptable, since such an association could be a source of systematic error in our effect estimates and therefore, an opportunity to measure a study's ability to account for such error. We explore this example further below.

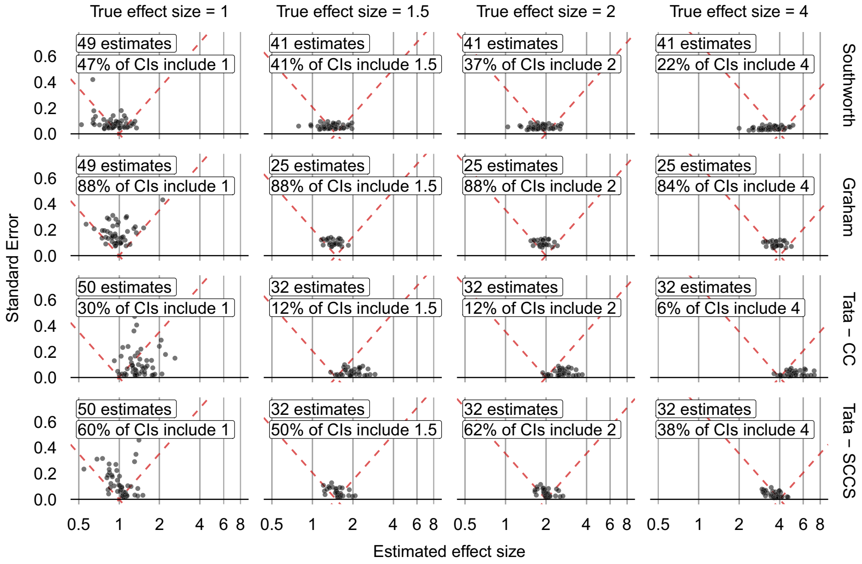

Fig. 1. Uncalibrated estimates and corresponding SEs for the negative and positive controls in the four studies. The estimates are stratified by true effect size. The areas above the red dashed lines indicate where the CIs include the true effect size. Note that, because of limitations in sample size, not all negative controls could be used to synthesize positive controls.

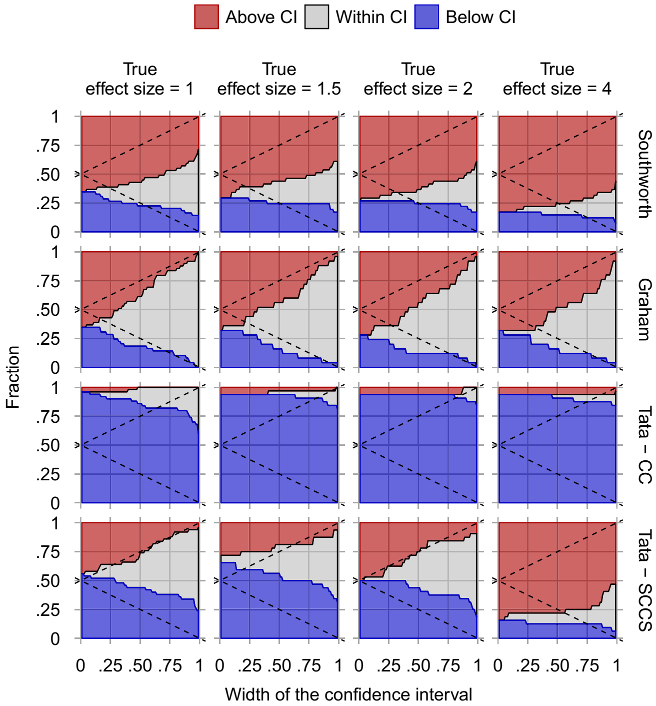

Fig. 2. The fraction of controls where the true hazard ratio is above, within, or below the CI for various widths of the CI. The dashed lines indicate the boundaries of a perfectly calibrated and centered estimator.

Synthetic Positive Controls. We exploit the negative controls to construct synthetic positive controls by injecting simulated outcomes during exposure. For example, assume that, during exposure to dabigatran, n occurrences of ingrowing nail were observed. If we now add an additional n simulated occurrences during exposure, we have doubled the risk. Since this was a negative control, the relative risk compared with warfarin was one, but after injection, it becomes two. Using this process, we have generated positive controls based on each negative control with relative risks of 1.5, 2, and 4. To preserve measured confounding, we sample the simulated outcomes based on the predicted probabilities of the outcome for each subject that we generated by a model fitted for each negative control outcome. We provide these models in Dataset S1. For example, the largest predictors in the model fitted for ingrowing nail are prior diagnosis of "onychomycosis due to dermatophyte," "gender = FEMALE," and prior use of piperazine derivatives. The accuracy of this model as measured using the area under the receiver–operator characteristics curve for predicting occurrence of ingrowing nail in the first year after index is 0.71.

Negative and Positive Control Estimates. In our replication of the study by Southworth et al. (12), the estimated incidence rate ratios for ingrowing nail and derived positive controls are 0.89 (0.77–1.03), 1.33 (1.18–1.50), 1.75 (1.57–1.95), and 3.30 (3.01–3.62) for real and synthetic incidence rate ratios 1, 1.5, 2, and 4, respectively. Fig. 1 reports effect size estimates for all negative and positive controls across the four studies as well as the percentage of CIs containing the true effect size. (All effect size estimates reported in this study can be found in *SI Appendix*.) Note that, for most studies, the 95% CIs do not show nominal characteristics; they do not contain the true effect size 95% of the time. Our replication of the study by Southworth et al. shows large bias in both positive and negative directions, probably because of the fact that the study design does not adjust for any confounding. In contrast, the Graham et al. (13) study replication shows little to no bias. Both the CC and SCCS designs in our Tata et al. (14) replications identify bias that tends to be positive.

A slightly different perspective on these results is provided in Fig. 2, where we have plotted the fractions of controls where the true effect size is above, within, or below the CI for various widths of the CI.

CI Calibration. Table 1 shows the maximum likelihood estimates for the systematic error model parameters described in *Materials and Methods*. In brief, a and b are the intercept and slope, respectively, of a model for estimating the mean, and c and d are the intercept and slope, respectively, of a model for estimating the logarithm of the SD. Note that an unbiased observational study would have $\hat{a} = 0$, $\hat{b} = 1$, $\hat{c} = -\infty$, and $\hat{d} = 0$.

Fig. 3 reveals effect size estimates for all negative and positive controls after calibration, showing that the coverage of the 95% CIs is much closer to the nominal 95%.

Internal Validity. To validate our CI calibration procedure, we apply a leave-one-out cross-validation approach. For each negative control and the positive controls derived from that negative control, we fit systematic error models using all other controls and compute calibrated CIs for the left-out controls across a wide range of widths. We subsequently check how often the true hazard ratio was within, above, or below the CI as shown in Fig. 4. These results show the calibrated CIs showing near-optimal coverage.

External Validity. Figs. 5 and 6 report effect size estimates for the outcomes of interest in the original studies as well as our replications both before and after calibration. Fig. 5 shows that, for our replication of the study by Southworth et al. (12), CI calibration leads to vastly wider CIs to account for the residual bias in this unadjusted design. In contrast, for the study by Graham et al. (13), our calibration generates little effect. To account for the strong positive bias observed in the Tata et al. (14) CC replication, Fig. 6 shows that the calibrated CI not only is made wider but also, moves toward lower effect sizes.

Discussion

Evidence from observational studies can only be trusted to the extent to which we have confidence in the validity of the statistics generated as part of the studies. In this paper, we discuss an empirical approach to calibrating CIs as a means of improving the value of the evidence produced from observational analyses. We showed CI calibration in the replication of two observational healthcare study pairs and evaluated internal and external validity of calibration. The procedure can be applicable to observational estimates generated from any study design as illustrated here with examples of cohort, CC, and SCCS studies from the literature. In all cases, the internal validation establishes that CI calibration restores nominal characteristics for various widths of the CI. Most importantly, the 95% calibrated CI contains the truth ∼95% of the time. The external validation indicates that accounting for potential residual bias inherent in a study design reduces the disagreement between conflicting observa-

Table 1. Estimated parameters for the systematic error models for the four studies

	Mean		SD	
Study	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$
Southworth et al. (12)	−0.07	0.93	−1.67	0.02
Graham et al. (13)	−0.03	0.99	−2.92	0.29
Tata et al. (14): CC	0.32	0.95	−1.69	−0.07
Tata et al. (14): SCCS	0.07	0.89	−2.11	−0.13

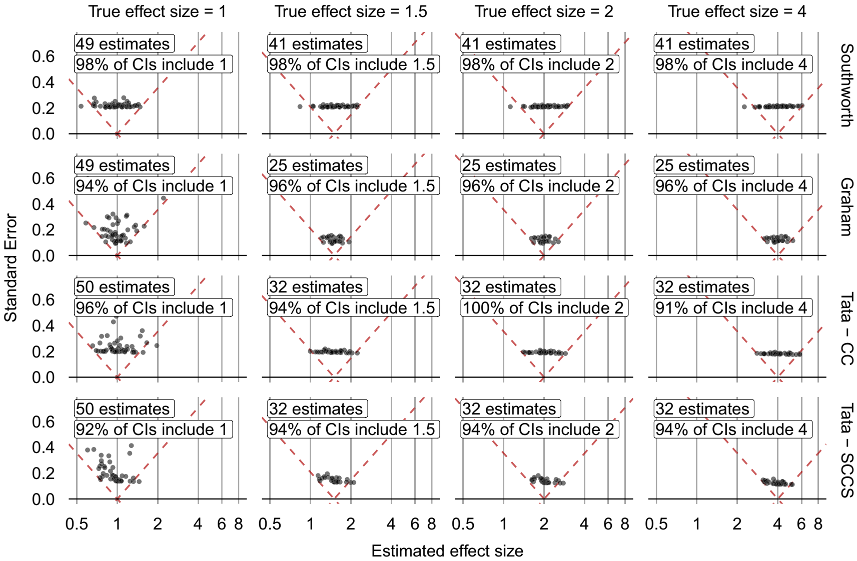

Fig. 3. Calibrated estimates and corresponding SEs for the negative and positive controls in the four studies. The estimates are stratified by true effect size. The areas above the red dashed lines indicate where the CIs include the true effect size.

tional studies. Not all disagreement was removed, but we would also not expect so. There are other differences between the studies for which calibration would not adjust. For example, the studies by Southworth et al. (12) and Graham et al. (13) use different populations of different ages, which could explain some of the remaining differences in estimated effect sizes.

In some analyses, like our Southworth et al. (12) study replication, the calibrated CI is much larger than the uncalibrated one. Some may object to this apparent loss of precision in the estimate, but the unadjusted study design is highly susceptible to residual bias in both positive and negative directions, and our calibrated CI merely reflects the uncertainty caused by this systematic error. In contrast, for other analyses, such as our Graham et al. (13) replication, the calibrated CI is virtually identical to the uncalibrated one. Although some might argue that, in this case, our calibration does not contribute anything, that would be incorrect. In the face of systematic error, the nominal characteristics of the uncalibrated CI remain unknown until our extensive empirical evaluation. After evaluation, we know that the calibrated CI contains the truth about 95% of the time.

Significance. This work is complementary to prior work in P value calibration, although we expect it to have a broader impact in our ability to interpret observational study results. Whereas the P value statistic has utility in the context of hypothesis testing and has commonly been used to gauge "statistical significance" at $P < 0.05$, the CI provides a richer collection of information: it not only estimates and bounds the magnitude of the effect, but it also expresses the extent of uncertainty attributable to both random and systematic error. While additional work should be conducted to further evaluate and validate the procedure, we believe that the evidence already provided suggests that CI calibration to quantify systematic error and produce more realistic statistics should be considered as part of standard practice in retrospective observational studies moving forward.

Limitations. We require that our negative controls are truly negative, meaning that the true effect size is exactly zero. Our process for selecting negative controls requires there to be no evidence for drug–outcome pairs where both the drug and the outcome have sufficient evidence, but lack of evidence does not necessarily equate to evidence of a lack of an effect. However, an analysis of effect sizes of negative controls appearing in randomized trials suggests that the null is consistently and completely true for our negative controls (*SI Appendix*).

We furthermore require that our negative and positive controls are, to some extent, exchangeable with the outcomes of interest. Note that we do not require the controls to have exactly the same magnitude and structure of confounding as the outcome

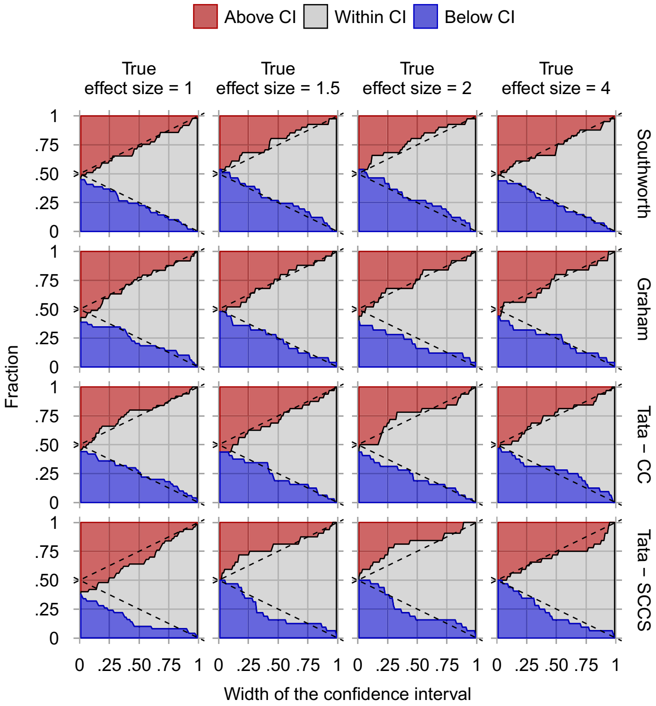

Fig. 4. The fraction of controls where the true hazard ratio is above, within, or below the calibrated CI for various widths of the CI. The dashed lines indicate the boundaries of a perfectly calibrated and centered estimator. Fractions were computed using leave-one-out cross-validation.

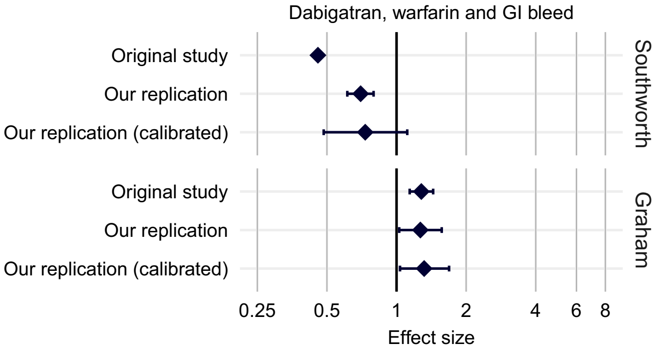

Fig. 5. Estimates from the original studies and our reproduction of the studies by Southworth et al. (12) and Graham et al. (13) both before and after calibration.

of interest as other proposed approaches do (8, 9) but rather, assume that they draw from the same distribution. Our leave-one-out cross-validation provides evidence that this assumption holds in general.

Negative controls represent both unmeasured and measured confounding, but our injection of outcomes to synthesize positive controls can only maintain measured confounding. As a consequence, positive controls may have less unmeasured confounding than what exists in reality, which could thereby lead to the calibration procedure underestimating the magnitude of systematic error. Nonetheless, we argue that calibration based on the systematic error that can be explicated, even if that comes with its own measurement error, is always preferable to ignoring systematic error entirely.

The uncalibrated estimates for some of our replications were not in agreement with the results of the original studies. For the study by Southworth et al. (12), this is most likely because we did not have access to the same database; for the SCCS study by Tata et al. (14), we suspect that this is because we adjusted for age using spline functions, and the original study may have used a step function instead. These discrepancies warrant caution if one wants to extrapolate our findings to the original studies, but we argue that our "replications" are still valid studies in their own right.

Here, we use 50 negative controls for each study. While an exploration of the required number (*SI Appendix*) suggests that perhaps that number could be lowered to 30 with little impact, selecting these controls may still require a significant amount of work. We do have a semiautomated process that can ease this burden as detailed in *Methods and Materials* and strive to make tools supporting this process available to the public soon.

Our process for computing calibrated CIs is computationally expensive. For each negative control, we must fit an outcome model, and we must perform the study itself not just for the outcome of interest but also, for each control. For example, for the Graham et al. (13) study replication, we need to estimate the hazard ratio for 50 negative control and 150 positive control outcomes in addition to the outcome of interest. However, as we show here, performing these computations is feasible and mostly requires computer time, which is cheap. In matters of public health significance, we would argue that computational complexity is acceptable when it leads to more reliable inference. To support the community in applying this practice, we have developed standardized tools to run these analyses efficiently and make them available as open source R packages.

Conclusions

As observational databases grow larger, the uncertainty caused by random error diminishes. As a consequence, the relative importance of quantifying nonrandom error mushrooms. Nowadays, observational studies report ever tighter CIs, but these do not truly capture the uncertainty in effect size estimates. We recommend producing calibrated CIs in all observational studies and presenting these calibrated intervals alongside the noncalibrated intervals to provide insight into the uncertainty inherent in evidence caused by systematic error.

Materials and Methods

Here, we provide an overview of the materials and methods used in this paper. More details can be found in *SI Appendix*, and the full code for executing the analyses described here is available as open source R package (https://github.com/OHDSI/StudyProtocols/tree/master/CiCalibration).

Negative Control Selection. We use the standardized process described elsewhere (15) for selecting negative controls. In brief, information from literature, product labels, and spontaneous reporting is automatically extracted and synthesized using a logistic regression model that we fit and evaluate on existing reference sets of negative and positive controls. For each of our drug exposures, we use the fitted model to predict negative control status for all outcomes having some data not related to the exposure. We rank-order the probable negative controls by prevalence in the observational database and manually review these in order until the target number of controls is selected For the Southworth et al. (12) and Graham et al. (13) replication studies, we selected 50 negative control outcomes that are assumed not to be causally related to either dabigatran or warfarin. For the replication studies by Tata et al. (14), we selected 50 negative control outcomes that are assumed not to be causally related to any SSRI. The full lists of negative controls can be found in *SI Appendix*.

Synthesizing Positive Controls. To understand the behavior of a method when the true relative risk is smaller or greater than one requires the use of positive controls, where the null is believed to not be true. Unfortunately, real positive controls for observational research tend to be problematic for three reasons. First, in most research contexts (for example, when comparing the effect of two treatments), there is a paucity of positive controls relevant for that specific context. Second, even if positive controls are available, the magnitude of the effect size may not be known with great accuracy and often depends on the population in which one measures it. Third, when treatments are widely known to cause a particular outcome, this shapes the behavior of physicians prescribing the treatment (for example, by taking actions to mitigate the risk of unwanted outcomes), thereby rendering the positive controls useless as a means for evaluation (16). We, therefore, use synthetic positive controls created by modifying a negative control through injection of additional simulated occurrences of the outcome during the time at risk for the exposure. One issue that stands important is the preservation of confounding. The negative controls may show strong confounding, but if we inject additional outcomes randomly, these outcomes will not be confounded, and we may, therefore, be optimistic in our evaluation of our capacity to deal with confounding for positive controls. To preserve confounding, we want the outcomes to show similar associations with baseline subject-specific covariates as the original outcomes. To achieve this, we fit large-scale predictive models

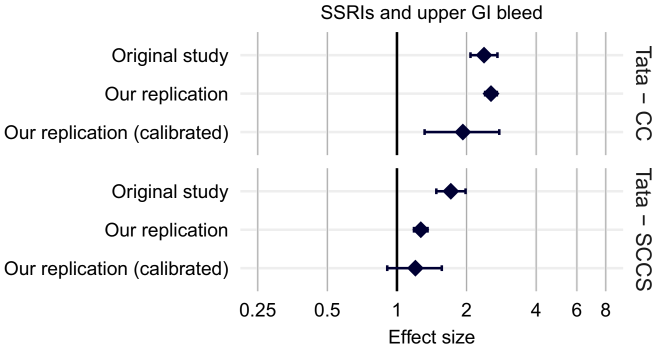

Fig. 6. Estimates from the original studies and our reproduction of the studies by Tata et al. (14) both before and after calibration.

for each negative control using L_1 regularized survival regression (17, 18). We insert outcomes by drawing from the per-subject predicted probabilities within the exposed population until we achieve the desired incidence rate ratio. The target incidence rate ratios are 1.5, 2, and 4.

When fitting the predictive models, we include covariates for demographics (age, gender, race, ethnicity, year of index date, month of index date), all diagnose codes, groups of diagnose codes (Medical Dictionary for Regulatory Activities as well as Systematized Nomenclature of Medicine groups), all drug exposure codes, groups of drugs [ATC (Anatomical Therapeutic Chemical Classification System) groups], all procedure codes, all observation codes, all measurement codes, and several risk scores [Charleston; Diabetes Complications Severity Index; CHADS2 (congestive heart failure, hypertension, age $\geq$75, diabetes mellitus, stroke or transient ischemic attack); CHADS2VASc (congestive heart failure, hypertension, age $\geq$75, diabetes mellitus, stroke or transient ischemic attack or thromboembolism, vascular disease, age $\geq$65, sex category)]. The four replication studies return between 5,482 and 26,097 potential covariates for prediction. The full details are in the Observational Health Data Science and Informatics FeatureExtraction R package (https://github.com/OHDSI/FeatureExtraction).

CI Calibration. For CI calibration, we build on our previous work in calibrating *P* values (10). Using the computed effect size estimates for the negative and positive controls, we observe to what extent random error alone explains the difference between the estimates and their true effect sizes. Systematic error explains any additional difference. We fit a systematic error model using the effect size estimates for the controls and subsequently use this model to compute calibrated CIs for the effect sizes of interest. The model assumes that systematic error follows a Gaussian probability distribution around the true effect size. We have found that a Gaussian distribution provides a good approximation, and more complex models, such as mixtures of Gaussians and nonparametric density estimation, do not improve results. Let $\hat{\theta}_i$ denote the computed log effect estimate (e.g., hazard ratio) from the *i*th negative or positive control, and let $\hat{\tau}_i$ denote its corresponding estimated SE for $i = 1, \ldots, n$. Let θ_i denote the true log effect size, and let β_i denote the asymptotic bias associated with pair *i*: specifically, the difference between the log of the true effect size and the log of the estimate that the study would have returned for control *i* had it been infinitely large. As in the standard CI computation, we assume that $\hat{\theta}_i$ is normally distributed with mean $\theta_i + \beta_i$ and variance $\hat{\tau}_i^2$. Note that the traditional CI calculation always assumes $\beta_i = 0$ but that we assume that β_i for all *i* arises from a normal distribution with mean $\mu(\theta_i)$ and SD $\sigma(\theta_i)$ that follow linear models, after appropriate transformation, with unknown intercepts *a* and *c* and slopes *b* and *d*, respectively. Specifically, we model

$$\begin{aligned} \beta_i &\sim N(\mu(\theta_i), \sigma^2(\theta_i)) \text{ and} \\ \hat{\theta}_i &\sim N(\theta_i + \beta_i, \hat{\tau}_i^2), \end{aligned} \qquad [1]$$

where

$$\begin{aligned} \mu(\theta_i) &= a + b \times \theta_i \text{ and} \\ \log \sigma(\theta_i) &= c + d \times \theta_i. \end{aligned} \qquad [2]$$

Also, $N(\cdot, \cdot)$ denotes a Gaussian distribution characterized by its mean and variance. We estimate *a*, *b*, *c*, and *d* by maximizing the marginalized likelihood in which we integrate out the unobserved β_i:

$$l(a, b, c, d | \theta, \hat{\theta}, \hat{\tau}) \propto \prod_{i=1}^{n} \int p(\hat{\theta}_i | \beta_i, \theta_i, \hat{\tau}_i) p(\beta_i | a, b, c, d, \theta_i) d\beta_i, \qquad [3]$$

where fixed $\theta = (\theta_1, \ldots, \theta_n)$, measured $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$, and $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_n)$, yielding maximum likelihood estimates $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$.

We compute a calibrated CI that uses the systematic error model. Let $\hat{\theta}_{n+1}$ denote the log of the effect estimate for a new outcome of interest, and let $\hat{\tau}_{n+1}$ denote the corresponding estimated SE. From the assumptions above and assuming that β_{n+1} arises from the same systematic error model, we have

$$\hat{\theta}_{n+1} \sim N\left(\theta_{n+1} + \hat{a} + \hat{b} \times \theta_{n+1}, e^{2(\hat{c} + \hat{d} \times \theta_{n+1})} + \hat{\tau}_{n+1}^2\right). \qquad [4]$$

We find the lower bound of the calibrated 95% CI by solving this equation for θ_{n+1}:

$$\Phi\left(\frac{\theta_{n+1} + \hat{a} + \hat{b} \times \theta_{n+1} - \hat{\theta}_{n+1}}{\sqrt{e^{2(\hat{c} + \hat{d} \times \theta_{n+1})} + \hat{\tau}_{n+1}^2}}\right) = 0.025, \qquad [5]$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. We find the upper bound similarly for probability 0.975. We define the calibrated point estimate by using probability 0.5.

The R code for estimating calibrated CIs is included in *SI Appendix* and is also implemented in the EmpiricalCalibration R package (https://cran.r-project.org/web/packages/EmpiricalCalibration).

Study Replication. In this section, we provide a short overview of how we replicate the four observational studies. Full details can be found in *SI Appendix*.

Southworth et al. (12) replication. This Southworth et al. (12) study is a new user cohort design that compares new users of dabigatran with new users of warfarin for the outcome of GI hemorrhage. Subjects are required to have 183 d of continuous observation before initiating treatment, a prior diagnosis of atrial fibrillation, and no prior exposure to either dabigatran or warfarin. The study computes an incidence rate ratio without any adjustment for confounders. Time at risk is defined as the time on the drug. The original study used the "Mini-Sentinel Database," a network of private payer claims databases. For our replication, we use the OptumInsight's deidentified Clinformatics Datamart (Optum), a private payer claims database that is part of the Sentinel network. We analyze 5,982 dabigatran-exposed and 19,155 warfarin-exposed subjects.

Graham et al. (13) replication. The Graham et al. (13) study is also a new user cohort design that compares new users of dabigatran with new users of warfarin for the outcome of GI hemorrhage. Subject are required to have 183 d of continuous observation before initiating treatment, be at least 65 y old at index date, and have no prior exposure to warfarin or dabigatran (or any other anticoagulant). Additionally, subjects are required have a prior diagnosis of atrial fibrillation or flutter and no prior diagnosis of other indications. Propensity scores are generated by fitting a model for predicting treatment assignment based on baseline patient characteristics and are used to perform one-on-one matching. Hazard ratios are estimated through a Cox regression on the matched population. Time at risk is defined as starting on the day after initiating treatment and stopping when treatment is stopped, the outcome occurs, or observation time ends, whichever comes first. The original study uses the Centers for Medicare & Medicaid Services Medicare database. For our replication, we use the Truven MarketScan Medicare Supplementary Beneficiaries database. We analyze 15,796 dabigatran-exposed and 15,796 warfarin-exposed subjects.

Tata et al. (14) CC replication. The Tata et al. (14) CC study matches cases of upper GI bleeding to up to six controls on age, gender, and general practice. Only cases and controls ages 18 y old or older are included. Conditional logistic regression is used to estimate the odds ratio for the first upper GI bleed associated with exposure to any SSRI in the 30 d preceding the index date. The original study uses the The Health Improvement Network (THIN) database (19). For our replication, we use the Clinical Practice Research Datalink (CPRD) database (20), since both databases are United Kingdom general practice databases. We analyze 30,987 cases and 184,775 controls.

Tata et al. (14) SCCS replication. The Tata et al. (14) SCCS study uses a conditional Poisson regression to estimate relative incidence of upper GI bleeding compared with within-person control periods. Time at risk is defined as the time when exposed to any SSRI. Also included in the model are subject age using a spline model and exposures to nonsteroidal antiinflammatory drugs and tricyclic antidepressants. Patient time is restricted to time when the patient is at least 18 y old. To account for possible contraindication of antidepressants shortly after a GI bleed, the 30 d before SSRI exposure are excluded from the analysis. The original study used the THIN database (19). For our replication, we use the CPRD database (20). We analyze 31,386 cases.

The use of Optum and Truven Marketscan databases was reviewed by the New England Institution Review Board (IRB) and was determined to be exempt from broad IRB approval, as this research project did not involve human subjects research. The use of the CPRD for this study has been approved by the CPRD Independent Scientific Advisory Committee (ISAC) as protocol number 17_017R.

ACKNOWLEDGMENTS. This work was partially supported through National Science Foundation Grant IIS 1251151 and NIH Grant R01 LM006910. Infrastructure to carry out the project was funded, in part, by Janssen Research and Development.

1. Overhage JM, Ryan PB, Schuemie MJ, Stang PE (2013) Desideratum for evidence based epidemiology. *Drug Saf* 1(36 Suppl):S5–S14.
2. Prasad V, Jena AB (2013) Prespecified falsification end points: Can they validate true observational associations? *JAMA* 309:241–242.
3. Dusetzina SB, Brookhart MA, Maciejewski ML (2015) Control outcomes and exposures for improving internal validity of nonrandomized studies. *Health Serv Res* 50:1432–1451.
4. Arnold BF, Ercumen A (2016) Negative control outcomes: A tool to detect bias in randomized trials. *JAMA* 316:2597–2598.
5. Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* 21:383–388.
6. Arnold BF, Ercumen A, Benjamin-Chung J, Colford JM (2016) Brief report: Negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology* 27:637–641.
7. Zaadstra BM, Chorus AM, vanBuuren S, Kalsbeek H, vanNoort JM (2008) Selective association of multiple sclerosis with infectious mononucleosis. *Mult Scler* 14:307–313.
8. Tchetgen Tchetgen E (2014) The control outcome calibration approach for causal inference with unobserved confounding. *Am J Epidemiol* 179:633–640.
9. Flanders WD, Strickland MJ, Klein M (2017) A new method for partial correction of residual confounding in time-series and other observational studies. *Am J Epidemiol* 185:941–949.
10. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D (2014) Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Stat Med* 33:209–218.
11. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA (2016) Robust empirical calibration of p-values using observational data. *Stat Med* 35:3883–3888.
12. Southworth MR, Reichman ME, Unger EF (2013) Dabigatran and postmarketing reports of bleeding. *N Engl J Med* 368:1272–1274.
13. Graham DJ, et al. (2016) Stroke, bleeding, and mortality risks in elderly medicare beneficiaries treated with dabigatran or rivaroxaban for nonvalvular atrial fibrillation. *JAMA Intern Med* 176:1662–1671.
14. Tata LJ, et al. (2005) Does concurrent prescription of selective serotonin reuptake inhibitors and non-steroidal anti-inflammatory drugs substantially increase the risk of upper gastrointestinal bleeding? *Aliment Pharmacol Ther* 22:175–181.
15. Voss EA, et al. (2016) Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform* 66:72–81.
16. Noren GN, Caster O, Juhlin K, Lindquist M (2014) Zoo or Savannah? Choice of training ground for evidence-based pharmacovigilance. *Drug Saf* 37:655–659.
17. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 73:267–288.
18. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D (2013) Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans Model Comput Simul* 23:1–17.
19. van Staa TP, Parkinson J (2008) Response to: Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research by Lewis et al. *Pharmacoepidemiol Drug Saf* 17:103–104.
20. Herrett E, et al. (2015) Data resource profile: Clinical practice research datalink (CPRD). *Int J Epidemiol* 44:827–836.