Title
How Confident Are We about Observational Findings in Healthcare: A Benchmark Study.

Journal
Harvard data science review, 2(1)

ISSN
2688-8513

Authors
Schuemie, Martijn J
Cepeda, M Soledad
Suchard, Marc A
et al.

Publication Date
2020

DOI
10.1162/99608f92.147cc28e

Peer reviewed

HHS Public Access

Author manuscript

Harv Data Sci Rev. Author manuscript; available in PMC 2020 December 22.

Published in final edited form as:

Harv Data Sci Rev. 2020 ; 2(1): . doi:10.1162/99608f92.147cc28e.

How Confident Are We about Observational Findings in Healthcare: A Benchmark Study

Martijn J. Schuemie[*,1,2,3], **M. Soledad Cepeda**[1,2], **Marc A. Suchard**[1,3,6,7], **Jianxiao Yang**[1,6], **Yuxi Tian**[1,6], **Alejandro Schuler**[1,8], **Patrick B. Ryan**[1,2,4], **David Madigan**[1,5], **George Hripcsak**[1,4,9]

[1]Observational Health Data Sciences and Informatics

[2]Epidemiology Analytics, Janssen Research and Development

[3]Department of Biostatistics, University of California, Los Angeles

[4]Department of Biomedical Informatics, Columbia University

[5]Department of Statistics, Columbia University

[6]Department of Biomathematics, University of California, Los Angeles

[7]Department of Human Genetics, University of California, Los Angeles

[8]Center for Biomedical Informatics Research, Stanford University

[9]Medical Informatics Services, New York-Presbyterian Hospital

Abstract

Healthcare professionals increasingly rely on observational healthcare data, such as administrative claims and electronic health records, to estimate the causal effects of interventions. However, limited prior studies raise concerns about the real-world performance of the statistical and epidemiological methods that are used. We present the “OHDSI Methods Benchmark” that aims to evaluate the performance of effect estimation methods on real data. The benchmark comprises a gold standard, a set of metrics, and a set of open source software tools. The gold standard is a collection of real negative controls (drug-outcome pairs where no causal effect appears to exist) and synthetic positive controls (drug-outcome pairs that augment negative controls with simulated causal effects). We apply the benchmark using four large healthcare databases to evaluate methods commonly used in practice: the new-user cohort, self-controlled cohort, case-control, case-crossover, and self-controlled case series designs. The results confirm the concerns about these methods, showing that for most methods the operating characteristics deviate considerably from nominal levels. For example, in most contexts, only half of the 95% confidence intervals we calculated contain the corresponding true effect size. We previously developed an “empirical calibration” procedure to restore these characteristics and we also evaluate this procedure. While no one method dominates, self-controlled methods such as the empirically calibrated self-controlled case series perform well across a wide range of scenarios.

[*]Corresponding author: schuemie@ohdsi.org.

Plain Language Summary:

Existing healthcare data such as insurance claims and electronic health records are used to determine what the effects, both good and bad, of medical treatments are. However, concerns have been raised about whether the results are reliable. One challenge that must be overcome is that people who get a treatment may differ from those that do not, and if we do not adjust for that appropriately we may draw incorrect conclusions. For example, studying a chemotherapy drug, one might erroneously conclude the drug causes cancer, because patients taking the drug have cancer more often than those that do not take the drug.

We have created a benchmark to measure the performance of various methods for dealing with this and related issues. We use “control questions,” i.e., questions where we know the answer, and evaluate whether the different methods produce the expected results. Running this benchmark on four large healthcare databases covering millions of lives, we observe that most methods are not reliable. For example, more often than not, the known “true” answer lies outside the confidence interval, despite the fact that such confidence intervals are typically designed to include that true answer 95% of the time.

Our results therefore confirm the concerns about using healthcare data to determine the effect of treatments, but also show a way forward: when performing a study using a particular method, researchers should also perform a similar, although smaller, experiment like we did, to measure how reliable the method is in that context. These performance characteristics can then be taken into account when interpreting the results. We call this ‘empirical calibration.’ When using this calibration, we show that some methods tend to work rather well across the different scenarios we test.

Keywords

evaluation; methods; causal effect estimation; observational research

1 Introduction

Observational healthcare data, such as administrative claims and electronic health records, offer opportunities to generate real-world evidence about the effect of treatments that can meaningfully improve the lives of patients. Even though healthcare researchers have had access to large-scale observational databases for at least two decades, the literature still abounds with divergent opinions about the value of observational studies. Many past observational study results have failed to show concordance with randomized trials (Rush, Campbell, Jhund, Petrie, & McMurray, 2018) and have failed to replicate upon subsequent investigation (Overhage, Ryan, Schuemie, & Stang, 2013). A valid criticism of the entire observational study enterprise remains its historical lack of reproducibility: any researcher with a hypothesis about a potential causal effect of an exposure on an outcome can choose any observational dataset that captures the exposure and the outcome, choose from a wide array of alternative analytical designs, produce an effect estimate and, then, rationalize the clinical interpretation of the findings, whatever they might be. A different researcher with the same question could choose to study different data or apply different methods and may well reach different conclusions. In the face of conflicting evidence, decision-makers are

faced with making the subjective determination of which study results to trust; many decide to dismiss observational evidence completely. Little empirical evidence exists to guide decisions about when and how to use observational studies. If the field of observational research is to mature from an artisanal pursuit devoid of any established performance characteristics into a true data science, further methodological work is required to quantify the reliability of the generated evidence. Our proposed benchmark aims to help fill this void.

The performance of effect estimation methods will likely vary from use-case to use-case. We therefore recommend that practitioners always perform an evaluation within the study setting of interest, for example, by including negative (Dusetzina, Brookhart, & Maciejewski, 2015; Prasad & Jena, 2013) and positive controls (Schuemie, Hripcsak, Ryan, Madigan, & Suchard, 2018; Schuemie, Ryan, Hripcsak, Madigan, & Suchard, 2018) to estimate residual bias. Nonetheless, characterizing how a method performs across a wide range of settings also adds value. This understanding can serve as a prior when a study-specific evaluation is not (yet) available, and may aid development of novel methodology. We establish the Observational Health Data Science and Informatics (OHDSI) (Hripcsak et al., 2015) Methods Benchmark that seeks to measure the performance and operating characteristics of observational analysis methods against disparate observational data for the task of population-level effect estimation. We subsequently apply this benchmark to a wide range of commonly used study designs and analysis approaches as implemented in the OHDSI Methods Library (https://ohdsi.github.io/ MethodsLibrary), an open source collection of R packages.

1.1 Population-level effect estimation

Observational healthcare data can support multiple analytic use-cases such as clinical charac-terization of populations of interest, patient-level prediction (Reps, Schuemie, Suchard, Ryan, & Rijnbeek, 2018), and population-level effect estimation. In this manuscript we focus on population-level effect estimation, that is, the estimation of average causal effects of medical interventions on specific health outcomes of interest. In what follows, we consider two different estimation tasks:

- Direct effect estimation: estimating the effect of an exposure on the risk of an outcome, as compared to no exposure.
- Comparative effect estimation: estimation the effect of one exposure (the target exposure) on the risk of an outcome, as compared to another exposure (the comparator exposure).

In both cases, the patient-level causal effect contrasts a factual outcome, e.g., what happened to the exposed patient, with a counterfactual outcome, that is, what would have happened had the exposure not occurred (direct) or had a different exposure occurred (comparative). Since any one patient reveals only the factual outcome (the fundamental problem of causal inference), the various effect estimation methods employ analytic devices to shed light on the counterfactual outcomes.

Use-cases for population-level effect estimation include treatment selection, safety surveillance, and comparative effectiveness. Methods can test specific hypotheses one-at-a-

time (e.g., ‘signal evaluation’) or explore multiple-hypotheses-at-once (e.g., ‘signal detection’). In all cases, the objective remains the same: to produce a high-quality estimate of the causal effect.

1.2 Prior work

Many authors have employed simulation to evaluate the general usefulness of specific observational study designs yet concerns always remain about real world relevance. Others use just one or two real world examples, raising concerns about generalizability. Substantial literature compares results from observational studies to those from randomized controlled trials (RCTs) (Anglemyer, Horvath, & Bero, 2014). Indisputably, RCTs provide the most credible evidence about causal effects of medical interventions. However, for myriad reasons, RCTs themselves can fail to replicate (Ioannidis, 2005) or can yield answers that are simply wrong or irrelevant to the populations of actual interest (Deaton & Cartwright, 2018; Frieden, 2017).

The EU-ADR (Exploring and Understanding Adverse Drug Reactions) project performed the first attempt at systematically evaluating a wide range of population-level estimation methods (Schuemie et al., 2012). The project constructed a reference set (Coloma et al., 2013) consisting of 50 negative controls (drug-outcome pairs where no causal association is believed to exist) and 44 positive controls (drug-outcome pairs where the drug is known to cause the outcome). The project applied ten different methods to estimate the effects for the negative and positive controls, using data from seven databases across three countries in Europe comprising over 20 million subjects. The project evaluated each method on whether positive controls tended to have higher estimates than negative controls. In that experiment, two particular analytic methods, the case-control design and the Longitudinal Gamma Poisson Shrinker (Schuemie, 2011), provided the best performance.

The Observational Medical Outcomes Partnership (OMOP) performed a similar evaluation in the U.S. (Ryan et al., 2012). OMOP evaluated eight analytic methods on a set of 44 negative controls and 9 positive controls in ten databases comprising over 130 million subjects. Although no specific method demonstrated superior performance across the board, a propensity score-based new-user cohort method achieved the highest performance. OMOP also performed a second, much larger experiment (DuMouchel, Ryan, Schuemie, & Madigan, 2013; Madigan, Schuemie, & Ryan, 2013; Noren et al., 2013; Overhage et al., 2013; Reich, Ryan, & Schuemie, 2013; Ryan & Schuemie, 2013; Ryan, Schuemie, Gruber, Zorych, & Madigan, 2013; Ryan, Schuemie, & Madigan, 2013; Ryan, Schuemie, Welebob, et al., 2013; Ryan, Stang, et al., 2013; Schuemie, Madigan, & Ryan, 2013; Suchard, Zorych, et al., 2013). This experiment evaluated hundreds of different variants of seven main analytic methods on a set of 234 negative controls and 165 positive controls in five databases comprising 73 million subjects. Schuemie, Gini, et al., (2013) also replicated the experiment in the EU-ADR network. The results of these experiments suggested higher performance for self-controlled methods, but also revealed that for all methods, the coverage of, for example, 95% confidence intervals, was substantially less than the nominal 95%.

All prior evaluations relied on reference sets of manually crafted negative and positive controls. These sets require onerous work to create, and, even after meticulous manual

review, arguments arose over the true status of controls (Hennessy & Leonard, 2015; Overhage, Ryan, Schuemie, & Stang, 2015). More importantly, whereas we can assume the true relative risk is 1 for the negative controls, the true magnitude of the effect is never known with acceptable precision for the positive controls. This is the main reason why all evaluations primarily focused on the ability to distinguish positive from negative controls and not on the ability to accurately estimate the effect size. Another important limitation of positive controls is the fact that, by design, little or no controversy surrounds their existence. Physicians know of these effects, and, in the case of adverse outcomes, may well attempt to mitigate these known risks, for example, by careful monitoring to prevent the adverse outcome, or by selectively prescribing only to those who have not experienced the outcome before. The latter behavior might lead to bias in the evaluation of methods, favoring self-controlled methods (Noren, Caster, Juhlin, & Lindquist, 2014). A further limitation of these earlier evaluations is their failure to include some important analytical choices. For example, in the OMOP experiments, the new-user cohort design did not consider time-to-event models; the multiple self-controlled case series design (Simpson et al., 2013) failed to include a variant that disabled shrinkage on the estimate of the effect of interest, thus only evaluating SCCS estimates with a built-in bias towards the null.

2 Analytical Methods

This section highlights the OHDSI Methods Benchmark, the analytical methods included in our evaluation, and the data sources used. More details can be found in the protocol provided online, along with the full source code used to execute this study, at https://github.com/ohdsistudies/MethodsLibraryPleEvaluation.

2.1 Notation

A key concept in our methodology is that of a cohort. We define a cohort c, $c = 1, \ldots, C$ as a group of subjects that satisfy one or more criteria for a duration of time. For example, a cohort could comprise individuals newly diagnosed with hypertension, with one year's observation prior to cohort entry, prescribed a beta blocker at cohort entry, and followed thereafter for two years. A subject can belong to multiple cohorts at the same time, and belong to the same cohort multiple times. For example, a subject could drop in and out of a hypertension cohort according to whether they are taking or not taking a particular drug. We refer to each period of time a subject is in a cohort as an "entry." We denote by N_c the number of entries in cohort c and by d_{ci} the duration (in days) for entry i in cohort c. Finally, $N_c(t)$ denotes the number of entries in cohort c that span day t.

We use cohorts to study associations between interventions and "outcomes." An outcome (e.g., stroke) occurs at a discrete moment in time and may or may not have a duration. We denote by y_{ci} the number of outcome events observed for entry i in cohort c and by $y_{ci}(t)$ the number of outcome events observed for entry i in cohort c on day t.

An "exposure cohort" is a cohort where all entries are exposed to a particular treatment x, $x = 1, \ldots, X$.As such, $y_{ci}(x = j)$ denotes the number of outcome events for subject i in exposure cohort c associated with treatment $x = j$. We can also consider a counterfactual cohort, identical in every way, except each subject is unexposed to treatment j at all times

while in the cohort. Here. $y_{ci}(x = \neg j)$ denotes the number of outcome events for patient i in this counterfactual cohort.

We define the causal effect of $x = j$ on Y, within some exposure cohort c defined by exposure to $x = j$, as the incidence rate ratio:

$$\mu_{cx} = \frac{\sum_i y_{ci}(x = j) \,/\, \sum_i d_{ci}}{\sum_i y_{ci}(x = \neg j) \,/\, \sum_i d_{ci}}.$$

Alternatively, we can also formulate the effect as the hazard ratio:

$$h_{cx} = E\left(\lim_{\Delta t \to 0} \frac{\sum_i y_{ci}(x = j, [t, t + \Delta t]) \;/\; N_c(t)}{\sum_i y_{ci}(x = \neg j, [t, t + \Delta t]) \;/\; N_c(t)}\right).$$

Note that these quantities estimate the average treatment effect in the treated (ATT).

Finally, $y_{ci}(x = j, t)$ denotes the number of outcome events on day t for subject i in exposure cohort c associated with treatment $x = j$ and $y_{ci}(x = \neg j, t)$ denotes the corresponding quantity in the counterfactual unexposed cohort.

2.2 The OHDSI Methods Benchmark

The OHDSI Methods Benchmark consists of a gold standard (i.e., a set of causal facts), and a set of metrics to characterize a method's performance in estimating the answers.

2.2.1 Gold standard—The gold standard comprises 800 controls, with each item specifying a target treatment, comparator treatment, outcome, nesting cohort, and true effect size. Of the total set, 200 are "neg-ative controls" and Table 1 shows four examples. For each negative control, neither the target treatment nor the comparator treatment are believed to cause the corresponding outcome. Therefore the true effect sizes for the direct causal effect of the target treatment on the outcome, the direct causal effect of the comparator on the outcome, and the comparative effect of the target treatment versus the comparator treatment on the outcome are all 1. For example, considering the first row of Table 1, and letting y denote the outcome "acute pancreatitis" and $x = j$ denote the treatment brinzolamide, we have

$$y_{ci}(x = j, t) = y_{ci}(x = \neg j, t), i = 1, \ldots, N_c, t = 1, \ldots, d_{ci}.$$

As a consequence, both $\mu_{cx} = 1$ and $h_{cx} = 1$.

We selected these negative controls by first picking four outcomes (acute pancreatitis, gastrointestinal bleeding, inflammatory bowel disease, and stroke) and four pharmaceutical treatments (ciprofloxacin, diclofenac, metformin, and sertraline) representing chronic, acute, rare, and prevalent outcomes and treatments. For each of the four outcomes, we created 25 entries with target and comparator treatments that we do not believe cause the outcome. For example, the top two rows of Table 1 consider the outcome of acute pancreatitis and, collectively, four treatments that do not cause acute pancreatitis. Similarly, for each of the

four treatments, we selected 25 comparator treatment-outcome pairs such that neither the target treatment nor the comparator treatment cause the corresponding outcome. For example, for the bottom two rows of Table 1 where the target treatment is diclofenac, we selected celecoxib as a comparator treatment and in one case selected “acute stress disorder” as the outcome and, in the other case, “ingrowing nail.” Neither diclofenac nor celecoxib cause either acute stress disorder or ingrowing nail.

To create these entries, we deployed an automated procedure (Voss et al., 2017) to generate candidate lists of negative controls for each of the four main outcomes and four main treatments, drawing on literature, product labels, and spontaneous reports. We used these candidates to construct target-comparator-outcome triplets where neither the target nor the treatment causes the outcome, and the target and comparator were either previously compared in a randomized trial per ClinicalTrials.gov, or both had the same four-digit ATC code (same indication) but not the same five-digit ATC code (different class). We ranked the candidate negative controls on prevalence of the treatments and outcome and manually reviewed from the top until we established 25 controls per initial outcome or treatment that passed review, considering both lack of casual associations between treatments and outcomes as well as the suitability of the comparator.

We associated a “nesting cohort” with each negative control. The nesting cohort identifies a more homogeneous subgroup of the population, for example, people diagnosed with arthralgia. Often, retrospective case-control studies are nested in such a subgroup rather than the general population captured in a database to make exposed and unexposed more comparable. Defining the nesting cohort thus allows us to evaluate such a nested case-control design. We selected nesting cohorts by manually reviewing the most prevalent conditions and procedures on the first day of the target or comparator treatment. Supplementary Table 1 provides the full list of negative controls.

The remaining 600 entries comprise positive controls. To avoid the aforementioned shortcomings of “real” positive controls, we chose to generate synthetic positive controls (Schuemie, Hripcsak, et al., 2018). We used an automated procedure to derive these from the 200 negative controls by adding simulated additional outcomes in the target treatment cohort until a desired incidence rate ratio was achieved. For example, assume that, during treatment with diclofenac, m occurrences of ingrowing nail were observed. None of these were caused by diclofenac since this is one of our negative controls. If we were to add an additional m simulated occurrences during treatment with diclofenac, we would have doubled the observed effect size. Since this was a negative control, and since only the treatment cohort received new ingrowing nails and not the counterfactual cohort, the observed relative risk which was one becomes two.

More specifically, let θ denote the target effect size. Currently we use $\theta = 1.5$, $\theta = 2$ and $\theta = 4$ to generate 3 positive controls from every negative control. We increase outcome count y_{ci} $(x = j)$ in the target treatment (j) cohort to $y^{*}_{ci}(x = j)$ to approximate the desired θ. To avoid issues due to low sample size, we generate positive controls only when $y_{ci} \geq 25$.

The steps in the “injection” process are as follows:

1. Within the target treatment cohort c, we fit an l1-regularized Poisson regression model (Suchard et al., 2013) where the outcome $y_{ci}\,(x = j)$ represents the subject-level dependent variable and Z_{ci} represents the independent variables. The independent variables include demographics, as well as all recorded diagnoses, drug exposures, measurements, and medical procedures all measured prior to cohort entry ("baseline covariates"). We use 10-fold cross-validation to select the regularization hyperparameter. Let $\hat{\lambda}_{ci} = E(y_{ci} \mid Z_{ci})$ denote the predicted Poisson event rate for entry i in treatment cohort c.

2. For every entry in the target treatment cohort, sample n from a Poisson distribution with parameter $(\theta - 1)\hat{\lambda}_{ci}$ and set $y^*_{ci}(x = j) = y_{ci}(x = j) + n$.

3. Repeat step 2 until $\frac{\sum_i y^*_{ci}(x = j) \,/\, \sum_i d_{ci}}{\sum_i y_{ci}(x = j) \,/\, \sum_i d_{ci}} - \theta = \epsilon$, where ϵ is currently set to 0.01.

Figure 1 depicts this process. Assuming the synthetic outcomes have the same measurement error (same positive predictive value and sensitivity) as the observed outcomes, this process creates data that mimic a true marginal effect size of θ. Because we sample new outcomes from a large-scale predictive model, we also mimic the conditional effect (conditional on Z). We note that the altered data can capture effects due to measured confounding but not unmeasured confounding. Since all outcomes we consider are rare, post-injection odds ratios are all but identical to the corresponding relative risks.

We define exposures as exposure to any drug containing the ingredient specified in the gold standard. We merge consecutive exposures if the gap between exposures is less than 30 days. We defined the four main outcomes of interest (acute pancreatitis, gastrointestinal bleeding, inflammatory bowel disease, and stroke) using manually crafted rule-based definitions including various diagnosis codes (see Supplementary Data). The outcome occurs if we observe the outcome concept or any of its descendants. The nesting cohorts comprise the group of people that have any occurrence of the diagnosis code or any of its descendants. We define the cohort start date as the day of the first such diagnosis, and the cohort end date as the end of observation.

2.2.2 Metrics—For every database-method-control combination, we generate an effect size estimate that takes the form of either a relative risk, an odds ratio, or a hazard ratio, together with an indication of the uncertainty associated with the estimate (either a 95% confidence interval or a standard error). We also assume that the method computes a two-sided p-value for the null hypothesis of no effect.

Based on the estimates of a particular method for the 800 negative and positive controls, we can then compute the following metrics:

- AUC: The ability to discriminate between positive controls and negative controls based on the point estimate of the effect size.
- Coverage: How often the true effect size is within the 95% confidence interval.

- Mean precision: Precision is computed as 1 / (standard error)^2, higher precision means narrower confidence intervals. We use the geometric mean to account for the skewed distribution of the precision.
- Mean squared error (MSE): Mean squared error between the log of the effect size pointestimate and the log of the true effect size.
- Type 1 error: For negative controls, how often the null was rejected (at $\alpha = 0.05$). This is equivalent to the false positive rate and 1 - specificity.
- Type 2 error: For positive controls, how often the null was not rejected (at $\alpha = 0.05$). This is equivalent to the false negative rate and 1 - sensitivity.
- Non-estimable: How many of the controls the method was unable to produce an estimate. There can be various reasons why an estimate cannot be produced, for example, because there were no subjects left after propensity score matching, or because no subjects remained posessing the outcome.

The benchmark computes these metrics both overall, as well as by true effect size, by each of the four initial outcomes and four initial exposures, and by amount of data as reflected by the minimum detectable relative risk (MDRR) that we compute using a standard approach (Armstrong, 1987). When a method cannot estimate an effect, it returns an estimate of 1 with an infinite confidence interval.

2.3 Empirical calibration

In prior work, we described a method for empirically calibrating p-values (Schuemie, Ryan, Du-Mouchel, Suchard, & Madigan, 2014). Briefly, when evaluating a particular analytical method, the calibration procedure applies the method not only to the target-comparator-outcome of interest but also to all the negative controls. This generates draws from an "empirical" null distribution. By contrast with the theoretical null distribution (typically a Gaussian centered on 1), the empirical null distribution does not assume that the estimated effect size provides an unbiased estimate of the true effect. Instead the location and dispersion of the empirical null distribution reflects both random error and systematic error. "Calibrated" or "empirical" p-values use the empirical null distribution in place of the theoretical null distribution when computing p-values.

More formally, let $\hat{\theta}_i$ denote the estimated log effect estimate (relative risk, odds or incidence rate ratio) from the ith negative control, and let $\hat{\tau}_i$ denote the corresponding estimated standard error, $i = 1, \ldots, n$. Let θ_i denote the true log effect size (assumed 0 for negative controls), and let β_i denote the true (but unknown) bias associated with pair i, that is, 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 p-value computation, we assume that $\hat{\theta}_i$ is normally distributed with mean $\theta_i + \beta_i$ and variance $\hat{\tau}_i^2$. Note that in traditional p-value calculations, β_i is assumed to be equal to zero for all i. Instead we assume the β_i's arise from a normal distribution with mean v and variance σ^2. This represents the null (bias) distribution. We estimate v and σ^2 via maximum likelihood. In summary, we assume the following:

$$\beta_i \sim N(\upsilon, \sigma^2) \text{ and } \hat{\theta}_i \sim N(\theta_i + \beta_i, \tau_i^2),$$

and we estimate v and σ^2 by maximizing:

$$\prod_{i=1}^{n} \int p(\hat{\theta}_i \mid \beta_i, \theta_i, \hat{\tau}_i)\, p(\beta_i \mid \upsilon, \sigma^2) d\beta_i$$

yielding maximum likelihood estimates $\hat{\upsilon}$ and $\hat{\sigma}^2$. We compute a calibrated p-value that uses the empirical null distribution. Let $\hat{\theta}_{n+1}$ denote the log of the effect estimate for the outcome of interest, and let $\hat{\tau}_{n+1}$ denote the corresponding (observed) estimated standard error. Assuming β_{n+1} arises from the same null distribution, we have that:

$$\hat{\theta}_{n+1} \sim N(\hat{\upsilon}, \hat{\sigma}^2 + \hat{\tau}_{n+1}),$$

and the p-value calculation follows naturally. Our prior work has shown that, unlike standard p-values, calibrated p-values maintain type I error rates at or close to the desired level, e.g., 5%.

Schuemie, Hripcsak, et al. (2018) used positive controls to extend the concept of calibrated p-values to calibrated confidence intervals. These intervals reflect actual accuracy on negative and positive controls and, like calibrated p-values, capture both random and systematic error. We assume that β_i, the bias associated with control i, again comes from a Gaussian distribution, but this time using a mean and standard deviation that are linearly related to θ_i, the true effect size:

$$\beta_i \sim N(\upsilon(\theta_i), \sigma^2(\theta_i))$$

where:

$$\upsilon(\theta_i) = a + b\theta_i \text{ and } \sigma^2(\theta_i) = c + d \times \mid \theta_i \mid .$$

We estimate a, b, c and d by maximizing the marginalized likelihood in which we integrate out the unobserved β_i:

$$\prod_{i=1}^{n} \int p(\hat{\theta}_i \mid \beta_i, \theta_i, \hat{\tau}_i^2)\, p(\beta_i \mid a, b, c, d,]\theta_i) d\beta_i$$

yielding $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$. We compute a calibrated CI that uses the systematic error model. Let $\hat{\theta}_{n+1}$ again denote the log of the effect estimate for the outcome of interest, and let $\hat{\tau}_{n+1}$ denote the corresponding (observed) estimated standard error. Then:

$$\hat{\theta}_{n+1} \sim N(\theta_{n+1} + \hat{a} + \hat{b} \times \theta_{n+1}, \hat{c} + \hat{d} \times \mid \theta_{n+1} \mid + \hat{\tau}^2_{n+1}),$$

and the calibrated confidence interval follows.

Our prior work has also shown that, unlike standard confidence intervals, calibrated confidence intervals maintain coverage at or close to the desired level, e.g., 95%. Typically, but not necessarily, the calibrated confidence interval is wider than the nominal confidence interval, reflecting the problems unaccounted for in the standard procedure (such as unmeasured confounding, selection bias and measurement error) but accounted for in the calibration.

In this paper, we are using controls both for calibration and for evaluation. To avoid an over-optimistic evaluation, we use a leave-one-out approach: for each control (positive or negative) we use all the controls *except* the control being calibrated and its siblings. By siblings we mean the set containing a negative control and the positive controls derived from that negative control.

2.4 Implementation

To facilitate others in executing the Methods Benchmark on their own data and methods, we have created an open-source R package (https://github.com/OHDSI/MethodEvaluation). This package works with any observational database in the OMOP Common Data Model (Overhage, Ryan, Reich, Hartzema, & Stang, 2012). The package will construct the various exposures, outcomes, and nesting cohorts, as well as perform the positive control synthesis. The package also computes the various performance metrics described above. Note that negative controls are application-specific and implementation requires a de novo effort to develop negative controls for each new application domain.

2.5 The OHDSI Methods Library

The OHDSI Methods Library comprises a collection of open source R packages designed to work directly on observational data in the OMOP Common Data Model. The library supports a wide array of technical platforms including traditional database systems (PostgreSQL, Microsoft SQL Server, and Oracle), parallel data warehouses (Microsoft APS, IBM Netezza, and Amazon RedShift), as well as Big Data platforms (Hadoop through Impala, and Google Big-Query). With a few lines of R code and predefined exposures and outcomes of interest, the library allows one to execute a full observational study, producing effect size estimates as well as study diagnostics and additional information such as population characteristics. The Methods Library implements a wide range of population-level estimation methods, was primarily developed by the authors of this paper, and has already been used extensively in published clinical and methodological studies (Duke et al., 2017; Ramcharran, Qiu, Schuemie, & Ryan, 2017; Ryan et al., 2018; Ryan, Schuemie, Ramcharran, & Stang, 2017; Schuemie, Hripcsak, Ryan, Madigan, & Suchard, 2016; Schuemie, Hripcsak, et al., 2018; Schuemie et al., 2014; Schuemie, Ryan, et al., 2018; Suchard et al., 2019; Tian, Schuemie, & Suchard, 2018; Vashisht et al., 2018; Wang et al., 2017; Weinstein et al., 2017; Yuan et al., 2018; Suchard et al., 2019).

Below are descriptions of the five packages included in our evaluation, representing five well-known population-level estimation methods. For each package, we also list the analytic choices within each method that we evaluate separately.

2.5.1 Cohort method—The new-user cohort method attempts to emulate a randomized clinical trial (Hernan & Robins, 2016). Subjects that are observed to initiate one treatment (the target exposure cohort j with treatment $x = j$) are compared to subjects initiating another treatment (the comparator exposure cohort k with treatment $x = k$) and are followed for a specific amount of time following treatment initiation, for example, the time they stay on the treatment. Figure 2 provides an illustration.

We compute the hazard ratio between the two cohorts:

$$h_{(j\ /\ k)} = E\left(\lim_{\Delta t \to 0} \frac{\sum_i y_{ji}(x = j, [t, t + \Delta t]) \ /\ N_j(t)}{\sum_i y_{ki}(x = k, [t, t + \Delta t]) \ /\ N_k(t)}\right).$$

One crucial difference with a randomized trial is that there is no randomization, and therefore there might be systematic differences between the target and comparator populations, making the comparator a poor approximation of the counterfactual. Without adjusting for these differences, estimates are likely to be confounded. A popular mechanism for adjusting for confounding is the use of Propensity Scores (PS). The PS is the probability of a subject re-ceiving one treatment instead of the other, conditional on baseline characteristics (Rosenbaum & Rubin, 1983):

$$e_{cijk} = \Pr(x = j \mid x = j \text{ or } x = k, Z_{ci}).$$

A model – typically a logistic regression – is fitted using the observed treatment assignments (target or comparator), then the model is used to produce the PS for each subject. In the past, PS's were computed based on manually selected characteristics, and although the CohortMethod package can support such practices, we use large-scale regularized regression using many generic characteristics. Tian et al. (2018) provide empirical evidence indicating the superiority of such an approach. These characteristics include demographics, as well as all diagnoses, drug exposures, measurement, and medical procedures observed prior to treatment initiation, and exclude the target and comparator treatment. A model typically involves 10,000 to 100,000 unique characteristics.

The advantage of the PS is that, when there are no unmeasured confounders, the treatment assignment is independent of the potential outcomes, conditional on the PS:

$$(y_{ci}(x = j), y_{ci}(x = k)) \perp\!\!\!\perp x_{ci} \mid e_{cijk}.$$

In other words, absent unmeasured confounding, conditional on the PS, the comparator can serve as a counterfactual, and we can compute an unbiased hazard ratio.

One way to take advantage of this property is by performing stratification on the PS, or matching, which can be considered very fine stratification. Another way is to use inverse probability of treatment weighting (IPTW), where each observation is weighted by $w_{ci\,jk} = p_{cj}/e_{ci\,jk}$ if $x = j$, and by $w_{ci\,jk} = (1 - p_{cj})/(1 - e_{ci\,jk})$ if $x = k$, where p_{cx} is the proportion of c having $x = j$ (Xu et al., 2010).

Another strategy for adjusting for differences between the two groups is to include additional variables in the outcome model. One major limitation of this approach is that, whereas there often is a wealth of data to fit a propensity model (with thousands of people in both treatment groups), the outcomes we study tend to be somewhat rare, causing a paucity of data when trying to fit elaborate models with the outcome as the dependent variable. One approach is to use both a PS and add the same variables that were used in the propensity model in the outcome model, thus adjusting for the same variables twice, but in different ways.

The new-user cohort method inherently is a method for comparative effect estimation, comparing one treatment to another. It is difficult to use this method to compare a treatment against no treatment, since it is hard to define a group of unexposed people that is comparable with the exposed group. If one wants to use this design for direct effect estimation, one way is to select a comparator treatment for the same indication as the exposure of interest, where the comparator treatment is believed to have no effect on the outcome. Unfortunately, such a comparator might not always be available. In our gold standard, the comparators were specifically selected to have no effect, so we can also evaluate the cohort method's performance for direct effect estimation.

Evaluation settings.: In our evaluation we focus on differences between the various adjustment strategies. All evaluations require 365 days of continuous observation prior to treatment initiation, capture a large set of covariates in the year prior to exposure, use a Cox proportional hazards model, and follow subjects from the day of treatment initiation (so including outcomes occurring on the day of treatment initiation) until treatment discontinuation (end of exposure) or end of observation, whichever is first. Subjects having both target and comparator exposures are removed. PS are computed using large-scale regularized regression (Suchard, Simpson, et al., 2013), where the regularization hyperparameter is selected by optimizing the out-of-sample likelihood in a 10-fold cross-validation. All PS matching uses a caliper of 0.2 on the standardized logit scale (Austin, 2011). Stratification applies five equally-sized strata based on the PS distribution in the study population. A full outcome model including all covariates that are also included in the PS is fitted using a large-scale Cox regression with regularization on all variables except the treatment variable, again applying 10-fold cross-validation to select the regularization hyperparameter.

Table 2 lists the variants of the new-user cohort method included in the evaluation. A stratified outcome model is conditioned on the matched sets or PS strata and is required when using PS stratification or variable ratio matching. Variable ratio matching allows for more than one comparator subject to be selected for each target subject, as long as the matches stay within the predefined caliper (Rassen et al., 2012). Trimming is a common

practice when performing IPTW to counter the effect of extreme weights (Brookhart, Wyss, Layton, & Stürmer, 2013). Here we trim the most extreme 5% of each exposure group

2.5.2 Self-controlled cohort—Figure 3 illustrates the self-controlled cohort (SCC) design (Ryan, Schuemie, & Madigan, 2013) that compares the rate of outcomes during exposure (A) to treatment *j* to the rate of outcomes in the time just prior to the exposure (B):

$$\mu(A+B) = \frac{\sum_i y_{Ai}(x=j) \,/\, \sum_i d_{Ai}}{\sum_i y_{Bi}(x \neq j) \,/\, \sum_i d_{Bi}(x \neq j)}.$$

Because this is a self-controlled design (subjects are used as their own comparator), and because of the proximity in time (the control cohort entry immediately precedes the target cohort entry), an assumption is made that the comparator cohort is a good approximation of the counterfactual:

$$\frac{\sum_i y_{Ai}(x=\neg j)}{\sum_i d_{Ai}} = \frac{\sum_i y_{Bi}(x \neq j)}{\sum_i d_{Bi}}.$$

However, the method is vulnerable to differences between different time periods.

Evaluation settings.: All evaluations of the SCC compare time while exposed to time prior to exposure, require 365 days of continuous observation prior to the exposure, and 183 days of continuous observation after the exposure start. Where possible, the pre-exposure window is set to the same length as the exposure window. All exposures are included, not just the first per person. Confidence intervals of the incidence rate ratio are computed using an exact test (Lehmann, 2005).

Table 3 shows the analysis variations included in our evaluation. We vary the definition of the time-at-risk to be either the entire time the subject was exposed, or just the first 30 days after exposure start. In all analyses, the pre-exposure window is set to be the same length as the corresponding exposure window. In some variants, the date when the exposure started was included in the time-at-risk, in others it was ignored. Sometimes the amount of observation time prior to exposure is shorter than the time-at-risk window. By default, the pre-exposure window is then truncated to the available observation time, but in some analyses (marked “require full observation”) subjects were removed from the analyses if the pre-exposure observation time was too short.

2.5.3 Case-control—Figure 4 illustrates the case-control design. Case-control (Vandenbroucke & Pearce, 2012) studies consider questions of the form: “Are persons with a specific disease outcome exposed more frequently to a specific agent than those without the disease?” Thus, the central idea is to compare “cases,” i.e., subjects that experience the outcome of interest, with “controls,” i.e., subjects that did not experience the outcome of interest. Because in our case-control designs tested here we consider only exposure on the index date (not prior), we can frame the case-control design as defining four cohorts having length = 1 day for all entries:

A: exposed cases, defined as any day when an outcome occurs ($y_{Ai}(t) = 1$) and the subject is exposed ($x(t) = 1$)

B: exposed controls, defined as any day when an outcome does not occur ($y_{Bi}(t) = 0$) and the subject is exposed ($x(t) = 1$)

C: unexposed cases, defined as any day when an outcome occurs ($y_{Ci}(t) = 1$) and the subject is not exposed ($x(t) = 0$)

D: unexposed controls, defined as any day when an outcome does not occur ($y_{Di}(t) = 1$) and the subject is not exposed ($x(t) = 0$). Typically, controls (B and D) are reduced to some small sample, and matched to cases on some variables.

The case-control design computes the odds ratio:

$$\mathrm{OR}_{(A+B+C+D)} = \frac{N_A / N_B}{N_C / N_D}.$$

Because essentially all the outcomes we consider are rare, the odds ratio is almost identical to the rate ratio:

$$\mathrm{OR}_{(A+B+C+D)} \sim \mathrm{RR}_{(A+B+C+D)} = \frac{N_A / (N_A + N_B)}{N_C / (N_C + N_D)}.$$

Although rarely stated, an assumption in the case-control design is that the unexposed cases and controls form a good counterfactual for the exposed cases and controls:

$$E(y_{(A+B)x=\neg j}) = E(y_{(C+D)x=\neg j}).$$

Often, one matches controls to cases based on characteristics such as age and sex to make them more comparable. Another widespread practice is to nest the analysis within a specific subgroup of people, for example, people that have all been diagnosed with one of the indications of the exposure of interest.

Evaluation settings.: In all evaluations of the case-control design we match randomly selected controls to cases on age with a two-year caliper and sex, set the index date of the cases to the date of the outcome and use the same calendar date as the index date for the matched controls. We require 365 days of continuous observation prior to the index date, exclude cases from being controls for another case, and consider cases and controls to be exposed if they are exposed on the index date itself, including when the treatment is initiated on the index date. The outcome model uses logistic regression conditioned on the matched sets.

Table 4 enumerates the variants we evaluate. We either select up to two or ten matched controls per case, and optionally nest within the population corresponding to the indication.

2.5.4 Case-crossover—

Figure 5 illustrates the case-crossover design.

The case-crossover (Maclure, 1991) design is very similar to the case-control design, except the control cohorts are restricted to the same subjects as the case cohorts, and the control dates are restricted to dates falling in a specific interval before the case dates. It tries to determine whether there is something special about the day the outcome occurred. Since cases serve as their own control, it is a self-controlled design, and should therefore be robust to confounding due to between-person differences. One concern is that, because the outcome date is always later than the control date, the method will be positively biased if the overall frequency of exposure increases over time (or negatively biased if there is a decrease). To address this, the case-time-control design (Suissa, 1995) was developed, which adds matched controls to the case-crossover design to adjust for exposure trends.

Evaluation settings.: In all evaluations of the case-crossover design we require 365 days of continuous observation prior to the outcome date and consider subjects to be exposed if they are exposed on the outcome or control date itself, including when the treatment is initiated on the outcome or control date. The outcome model is a logistic regression conditioned on the persons.

Table 5 lists the variants of the case-crossover design included in our evaluation. When nested, the cases and, for the case-time-control extension, the controls, are selected from the group of people having the indication.

2.5.5 Self-controlled case series—

Figure 6 illustrates the self-controlled case series design.

The Self-Controlled Case Series (SCCS) design (Farrington, 1995; Whitaker, Farrington, Spiessens, & Musonda, 2006) compares the rate of outcomes during exposure to treatment j (cohort A) to the rate of outcomes during all unexposed time (cohort B), both before, between, and after exposures. It is a Poisson regression that is conditioned on the person:

$$\mu(A+B)x = \frac{\sum_i y_{Ai}(x=j) \,/\, \sum_i d_{Ai}}{\sum_i y_{Bi}(x \neq j \,/\, \sum_i d_{Bi}} \mid (s_{Ai} = s_{Bi})$$

where s_{ci} denotes the subject corresponding to entry i in cohort c. Thus, SCCS seeks to answer the question: “Given that a patient has the outcome, is the outcome more likely during exposed time compared to non-exposed time?” The assumption behind the SCCS is that the unexposed time of a subject forms a good counterfactual for the exposed time for that same subject. Like other self-controlled designs, the SCCS is robust to confounding due to between-person differences, but vulnerable to confounding due to time-varying effects. Several adjustments are possible to attempt to account for these.

Evaluation settings.: In all evaluations, we follow subjects from their start of observation (e.g., start of enrollment for insurance claims) to their end of observation, but disregard the first 365 days in the analysis except to determine the exposure status right after those initial 365 days. For example, if a 60-day prescription is started on day 340 after observation start, the subject is considered exposed on days 366-400. Unless stated otherwise, the time-at-risk is assumed to start the day after exposure start, and end when exposure stops. Only the first

occurrence of the outcome is considered, recurrent outcomes are ignored. A Poisson regression conditioned on the person estimates the incidence rate ratio.

Table 6 shows the variations of the SCCS included in the evaluation. We evaluate the effect of including the first day of exposure in the risk window, since this day could have many unrelated diagnoses being recorded while visiting the doctor. One frequent practice in SCCS designs is to set aside the time just prior to exposure to adjust for time-varying effects such as the contra-indications. We further evaluate adjusting for age and season by assuming a constant effect of age and season within each calendar month and using 5-knot cubic splines to model the effect across months. One important assumption underlying the SCCS is that the observation period end is independent of the date of the outcome. Because for some outcomes, especially ones that can be fatal such as stroke, this assumption can be violated. An extension to the SCCS has been developed that corrects for any such dependency (Farrington et al., 2011). A final refinement of the SCCS is to include not just the exposure of interest, but all other exposures to drugs recorded in the database (Simpson et al., 2013), potentially adding thousands of additional variables to the model. L1-regularization using cross-validation to select the regularization hyperparameter is applied to the coefficients of all exposures except the exposure of interest.

2.6 Data sources

For our evaluation we use the four databases listed below. These databases are converted to the OMOP Common Data Model (Overhage et al., 2012), which not only imposes a standard data structure, but also a standard encoding of the information. This allows for the same analysis code to be executed against each database without modification. Figure 7 shows summary descriptives of the four databases.

CCAE.—The IBM MarketScan® Commercial Claims and Encounters (CCAE) database represents data from individuals enrolled in United States employer-sponsored insurance health plans. The data include adjudicated health insurance claims (e.g., inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. Additionally, it captures laboratory tests for a subset of the covered lives. This administrative claimsdatabase includes a variety of fee-for-service, preferred provider organizations, and capitated health plans. The major data elements contained within this database are outpatient pharmacy dispensing claims (coded with National Drug Codes), and inpatient and outpatient medical claims, which provide procedure codes (coded in CPT-4, HCPCs, ICD-9-CM or ICD-10-PCS) and diagnosis codes (coded in ICD-9-CM or ICD-10-CM). The data also contain selected laboratory test results (those sent to a contracted third-party laboratory service provider) for a nonrandom sample of the population (coded with LOINC codes).

PanTher.—The Optum Pan-Therapeutic (PanTher) Electronic Health Records (EHR) dataset contains medical record data primarily from United States Integrated Delivery Networks. These include clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and

information derived from clinical notes using Natural Language Processing (NLP). PanTher integrates provider data from different EHR platforms (i.e., Cerner, Epic, GE, McKesson, etc.) and different versions of the same EHR platform.

JMDC.—The Japan Medical Data Center (JDMC) database consists of data from sixty societymanaged health insurance plans covering workers aged 18 to 65 and their dependents (children younger than 18 years old and elderly people older than 65 years old). JMDC data include membership status of the insured people and claims data provided by insurers under contract (e.g., patient-level demographic information, inpatient and outpatient data inclusive of diagnosis and procedures, and prescriptions as dispensed claims information). Claims data are derived from monthly claims issued by clinics, hospitals and community pharmacies. For claims only the month and year are provided; however, prescriptions, procedures, admission, discharge, and start of medical care are associated with a full date.

MDCR.—The IBM MarketScan® Medicare Supplemental Database (MDCR) represents health services of retirees in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. These data include adjudicated health insurance claims (e.g., inpatient, outpatient, and outpatient pharmacy). Additionally, it captures laboratory tests for a subset of the covered lives.

3 Results

We execute all 28 design variants of the five estimation methods on all 800 controls against the four databases, both with and without empirical calibration, thus producing a total of 179,200 effect size estimates. From these we derive a large set of performance metrics, which vary depending on choices of which controls and data to include in the evaluation. Below we walk through several examples of our results, starting with a single control, single database and two analysis variants, and gradually increasing the complexity. However, it is infeasible to cover the full set of results in this paper, and instead we refer the reader to our R Shiny-based (Chang, Cheng, Allaire, Xie, & McPherson, 2018) application: http://data.ohdsi.org/MethodEvalViewer/. This application, as shown in Figure 8, serves up our complete method evaluation results. Importantly, the app allows readers to inject their own sorting, and to filter the results to specific sets of controls, for example those for a specific outcome or exposure, or those with a specific true effect size.

3.1 Example analyses, control, and database

We use one example negative control, diclofenac - ingrowing nail, to illustrate our evaluation procedure. This is a negative control, because we firmly believe diclofenac does not cause ingrown nails, and we therefore assume the true effect size is 1.

When applying the cohort method in the CCAE database, we identify 967,086 new users of diclofenac, with at least 365 days of prior observation, no prior diclofenac exposure, and no prior diagnosis of ingrowing nails. We compare these to 774,063 new users of celecoxib, another negative control for ingrowing nail identified using similar criteria and included in the gold standard. We construct 98,159 baseline covariates based on data observed prior to

treatment initiation, including demographics, drug exposures, procedures, and diagnoses, and use these to fit a propensity model and compute the PS. For this example, we perform 1-on-1 matching on the PS, leaving 466,622 subjects in both the diclofenac and the celecoxib group. In the time until exposure end or observation end (whichever comes first), we observe 1,180 ingroing nail diagnoses in the diclofenac group, and 758 in the celecoxib group. A Cox regression produces a hazard ratio of 1.08 (95% confidence interval: 0.99-1.19).

When using the nested case-control design in the CCAE database, we first identify 25,054,470 subjects with a diagnosis of arthralgia, the nesting cohort listed in the gold standard. Within this nesting cohort we identify 793,153 cases who had their first ingrown nail diagnosis after their arthralgia diagnosis. We select up to 10 controls for each case from the nesting cohort, matching on age and sex, and requiring the index date (outcome date of the matched case) to be after their arthralgia diagnoses. After matching, there are 407,386 cases and 4,073,857 controls. We observe 5,736 cases and 35,330 controls to be on diclofenac on their index date. A logistic regression produces an odds ratio of 1.64 (95% confidence interval: 1.59-1.68).

3.2 Extending the example to all analyses

We similarly apply all other analysis variants discussed in Section 2.5 to produce the effect size estimates for our example negative control and show the results in Figure 9.

3.3 Extending the example to all controls

Similarly we apply all other analysis variants to all other negative and positive controls in our gold standard. Figure 10 shows the estimates for the two exemplar analysis variants discussed above. Note that this plot does not include all 800 controls, because some variants failed to produce an estimate.

To provide some sense of the scale of these analyses, we list median counts of some key quantities for the various analyses across the 200 negative controls in Table 7. We use the estimates for the gold standard to compute the metrics described earlier and shown in Figure 11. For example, for our exemplar control we observe that the confidence interval produced by the specific cohort method analysis (0.99-1.19) contains the true effect size (1.00). In the ‘Coverage’ column of Figure 11 we note that for this particular design using 1-on-1 PS matching, the 95% confidence interval contains the true effect size for 73% of our controls. The confidence interval of the nested case-control design (1.59-1.68) does not contain the true effect size. Figure 11 informs us that the 95% confidence interval of this particular design contains the true effect size only 22% of the time. We filter the set of controls to those having MDRR > 1.25 to ensure the different methods have somewhat comparable numbers of non-estimable controls. In general, all methods report lower coverage, and this is the case both for negative and positive controls, as can be seen in our web app. We compute the same metrics after applying empirical calibration of the confidence intervals and p-values, as shown in Figure 12.

3.4 Extending the example to all databases, stratifying by exposure and outcome

We compute similar metrics in the PanTher, JMDC, and MDCR databases, and provide these as supplementary materials. These metrics provide an overall evaluation of the performance of the various methods across all controls. However, we maybe interested in the performance in a given context, for example, when faced with a rare and acute outcome such as acute pancreatitis. As described earlier, our set of controls is constructed by first selecting four exposures and four outcomes, and generating controls for each of these. We can therefore stratify our controls accordingly. For example, there are 100 controls with acute pancreatitis as the outcome, and we can evaluate how well each method performs on this subset of controls. Figure 13 evaluates all analysis variants across all control strata, across all databases. To reduce visual complexity, we show only one metric: mean precision (1/SE2) after empirical calibration. Our Shiny app allows users to generate this graph interactively for the other metrics.

4 Discussion

Overall, we observe that most methods have low coverage and high type I error across all evaluated scenarios. For the methods we evaluate, the true effect size is generally more often outside the 95% confidence interval than within, and the methods reject the null hypothesis more often than not when the null hypothesis is in fact true. Many sources of systematic error threaten the validity of observational studies including selection bias, confounding, model misspecification, and measurement error. Since standard methods assume no systematic error exists, this poor performance perhaps should not surprise us. Fortunately, empirical calibration largely restores the nominal characteristics (i.e., 95% confidence interval coverage and 5% type I error rate). We believe these findings warrant consideration when interpreting findings of observational studies using these designs, and demonstrate the value of empirical calibration. This warrants particular scrutiny in large-scale observational research, where, we contend, textbook methods focus on the wrong type of error. Uncalibrated confidence intervals solely reflect random error. However random error approaches zero as sample size increases while systematic error, by stark contrast, remains stubbornly immune to more data. Empirical calibration represents an attempt to capture both sources of error.

4.1 How did the methods perform?

Answering the question “How did the methods perform?” depends on the use case and setting in which the method is used. If we are interested in establishing the magnitude of an effect, we may choose a method with low MSE, but we must also take into consideration how well a method expresses uncertainty. If a method with low MSE also has a low coverage of the 95% confidence interval it may still mislead us about the true magnitude of the effect. It is important to realize we can trade off performance on various metrics. We can use confidence interval calibration to ensure all methods have roughly nominal coverage and select the method with the highest precision after calibration to find which method reduces uncertainty the most. In that respect, the SCCS method adjusting for age and season, and the SCCS method adjusting for all drugs perform best in CCAE across all controls, with coverages after calibration of 94%, and 95% respectively, and mean precision after

calibration of 22.15 and 21.98, respectively. These same methods also comprise the top two in JMDC and MDCR, and are in the top five for Pan-Ther. If we drill down further into specific use cases, we do see different methods performing better under certain circumstances, as shown in Figure 13. For example, for our controls related to diclofenac, the self-controlled cohort performs best in three out of four databases, and for stroke the case-control design achieves the highest mean precision after calibration in the two US insurance claims databases (CCAE and MDCR).

If we are interested in the ability to distinguish effects from non-effects, we may look to type I and type II error. Again, we can trade off between these two errors, for example, by choosing a different alpha threshold. Alternatively, we might focus on the AUC, which summarizes the accuracy in distinguishing negative from positive controls, irrespective of the threshold. As yet another alternative, we could use empirical p-value calibration to restore type I error to nominal and select the method with the lowest type II error. The SCCS method adjusting for all other drugs consistently demonstrates the highest AUC in all four databases across all controls, with an AUC ranging from 0.95 to 0.98. This method also has the lowest type II error after calibration in all four databases. Across the eight subsets of controls, these findings are highly consistent, where this method has either the highest AUC or is very close to the best performance. These findings are in line with those from the second OMOP experiment, which also showed the highest AUCs for self-controlled methods (Ryan & Schuemie, 2013).

Even though one specific design might demonstrate the best performance on one or several metrics in our evaluation, it may very well be advantageous to use more than one method when estimating an effect. We observe that the correlation between estimates produced by the various methods can be quite low and even negative as shown in the Supplementary Materials, suggesting some orthogonality in the information provided by each method.

One interesting finding is that IPTW consistently presents considerably lower coverage than PS stratification and matching across scenarios and is in fact on par with using no PS adjustment. We hypothesize this may be due to extreme weights; using those weights may bias an estimate, but trimming extreme weights may introduce a different type of bias. This is in line with earlier findings comparing various PS adjustment strategies (Elze et al., 2017).

Mostly, the findings in the other databases agree with those for the CCAE database discussed above. One striking difference is that many methods appear strongly positively biased in the PanTher database. Upon investigation, it was revealed that the cause might be the definition of the observation period in this database. The observation period is defined to start at the first observed activity for a person, but for many subjects the first diagnosis and drug exposures are not observed until many years after observation start, suggesting incomplete data capture. Methods such as the SCCS will include these blind spots as time without exposure and without outcome, thus estimating a much lower rate of the outcome when not exposed, resulting in high incidence rate ratio estimates. It is interesting to note that the empirical calibration appears to largely correct for this systematic error.

4.2 Limitations

The results reported here may be specific to the databases that were used and may not generalize to others. It is encouraging, however, to see consistency in the main findings across the four databases, even though they represent very different types of observational healthcare data.

Whereas our negative controls reflect real confounding, both measured and unmeasured, our positive controls retain just some of that confounding. The additional synthetic outcomes only reflect measured confounding. Furthermore, these additional outcomes do not represent some effects of measurement error; positive controls imply constant positive predictive value and sensitivity, which may not be true in reality. Our performance metrics based on these positive controls may therefore be somewhat optimistic.

Our process for synthesizing positive controls aims to ensure that the true effect size holds for the various effect statistics used by the different methods, such as the incidence rate ratio, hazard ratio, and odds ratio, either marginal or conditional. This allows us to compare all methods on equal footing, but in real world situations these different statistics may deviate. Many will be willing to make the assumption that these differences are inconsequential, as evidenced for example by the fact that many meta-analyses simply combine these different estimates. Under this assumption, our evaluation may inform on what method to select overall. For those that are unwilling to make this assumption, but are able to specify exactly what effect statistic suits their needs (e.g., a conditional hazard ratio estimate of the effect in the treated), our results may still inform on the optimal choice under this constraint. Our negative controls do precisely reflect real world situations, and the results on these controls should therefore be informative to all.

A limitation of our negative controls is that by definition there is no causal association between exposure and outcome, including beneficial ones. This means that the outcome can never be the indication for a drug, which would be the case if we would like to estimate the effectiveness of a drug. Our findings are therefore most relevant for the estimation of (previously unknown) adverse effects, but less so for effectiveness studies.

By necessity, we evaluate only a selection of population-level effect estimation methods. Others, such as those using manually selected covariates instead of our large-scale PS, using a non-exposure group in a new-user cohort method when performing direct effect estimation, G-estimation (Robins, Blevins, Ritter, & Wulfsohn, 1992) and estimation through use of instrumental variables (Ertefaie, Small, Flory, &Hennessy, 2017) (with or without G-estimation) should also be evaluated in due time, although implementations that allow systematics application of these methods on large numbers of exposure-outcome pairs will need to be developed first.

4.3 Open science and transparency

One important aspect of the work presented here is that of Open Science. Although we are unable to share the patient-level data from CCAE, PanTher, JMDC, and MDCR, all other study artifacts such as study code, code implementing the various methods discussed here, and result sets are made publicly available in our GitHub repositories. Our results are

furthermore shared through an interactive app to allow readers to explore the results independently of our interpretation as discussed in this paper. We strongly believe this open approach to science will be of great benefit to the field of data science and beyond.

5 Conclusions

Large observational healthcare databases allow us to answer many important questions, including questions about causal effects. We provide a benchmark for evaluating effect estimation methods on real data and apply this to a large set of methods currently used to inform medical decision making. Our results show most methods display operating characteristics that are far from nominal, having the true effect size outside of the 95% confidence interval most of the time, and incorrectly rejecting the null when the null is true most of the time. Empirical calibration can largely restore these nominal characteristics. Empirically calibrated self-controlled methods such as the SCCS yield the highest precision, as well as AUC, and perhaps provide a reasonable default approach for future analyses.

Although our results inform on how methods perform in a wide range of scenarios, we strongly recommend including negative and positive controls in each observational study both to measure operating characteristics in a specific research setting, and to facilitate empirical calibration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported through the National Institutes of Health F31 LM012636 and R01 LM006910.

References

Anglemyer A, Horvath HT, & Bero L (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev(4), Mr000034. doi:10.1002/14651858.MR000034.pub2 [PubMed: 24782322]

Armstrong B (1987). A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. Am J Epidemiol, 126(2), 356–358. doi:10.1093/aje/126.2.356 [PubMed: 3605062]

Austin PC (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat, 10(2), 150–161. doi:10.1002/pst.433 [PubMed: 20925139]

Brookhart MA, Wyss R, Layton JB, & Sturmer T (2013). Propensity score methods for confounding control in nonexperimental research. Circ Cardiovascular Qual Outcomes, 6(5), 604–611. doi:10.1161/CIRCOUTCOMES.113.000359

Chang W, Cheng J, Allaire J, Xie Y, & McPherson J (2018). shiny: Web Application Frame-work for R (Version R package version 1.2.0). Retrieved from https://CRAN.R-project.org/package=shiny

Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, … Trifiro G (2013). A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf, 36(1), 13–23. doi:10.1007/s40264-012-0002-x [PubMed: 23315292]

Deaton A, & Cartwright N (2018). Understanding and misunderstanding randomized controlled trials. Soc Sci Med, 210, 2–21. doi:10.1016/j.socscimed.2017.12.005 [PubMed: 29331519]

Duke JD, Ryan PB, Suchard MA, Hripcsak G, Jin P, Reich C, … M JS (2017). Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network. Epilepsia. doi:10.1111/epi.13828

DuMouchel W, Ryan PB, Schuemie MJ, & Madigan D (2013). Evaluation of disproportionality safety signaling applied to healthcare databases. Drug Saf, 36 Suppl 1, S123–132. doi:10.1007/s40264-013-0106-y [PubMed: 24166229]

Dusetzina SB, Brookhart MA, & Maciejewski ML (2015). Control outcomes and exposures for improving internal validity of nonrandomized studies. Health Serv Res, 50(5), 1432–1451. doi:10.1111/1475-6773.12279 [PubMed: 25598384]

Elze MC, Gregson J, Baber U, Williamson E, Sartori S, Mehran R, … Pocock SJ (2017). Comparison of propensity score methods and covariate adjustment: Evaluation in 4 cardiovascular studies. J Am Coll Cardiol, 69(3), 345–357. doi:10.1016/j.jacc.2016.10.060 [PubMed: 28104076]

Ertefaie A, Small DS, Flory JH, & Hennessy S (2017). A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf, 26(4), 357–367. doi:10.1002/pds.4158 [PubMed: 28239929]

Farrington CP (1995). Relative incidence estimation from case series for vaccine safety evaluation. Biometrics, 51(1), 228–235. [PubMed: 7766778]

Farrington CP, Anaya-Izquierdo K, Whitaker HJ, Hocine MN, Douglas I, & Smeeth L (2011). Self-controlled case series analysis with event-dependent observation periods. J Amer Stat Assoc, 106(494), 417–426. doi:10.1198/jasa.2011.ap10108

Frieden TR (2017). Evidence for health decision making - beyond randomized, controlled trials. N Engl J Med, 377(5), 465–475. doi:10.1056/NEJMra1614394 [PubMed: 28767357]

Hennessy S, & Leonard CE (2015). Comment on: "Desideratum for evidence-based epidemiology". Drug Saf, 38(1), 101–103. doi:10.1007/s40264-014-0252-x [PubMed: 25511911]

Hernan MA, & Robins JM (2016). Using big data to emulate a target trial when a randomized trial Is not available. Am J Epidemiol, 183(8), 758–764. doi:10.1093/aje/kwv254 [PubMed: 26994063]

Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, … Ryan PB (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Stud Health Technol Inform, 216, 574–578. [PubMed: 26262116]

Ioannidis JP (2005). Contradicted and initially stronger effects in highly cited clinical research. Jama, 294(2), 218–228. doi:10.1001/jama.294.2.218 [PubMed: 16014596]

Lehmann EL, Romano Joseph P. (2005). Testing Statistical Hypotheses (third edition). New York: Springer.

Maclure M (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. Am J Epidemiol, 133(2), 144–153. doi:10.1093/oxfordjournals.aje.a115853 [PubMed: 1985444]

Madigan D, Schuemie MJ, & Ryan PB (2013). Empirical performance of the case-control method: Lessons for developing a risk identification and analysis system. Drug Saf, 36 Suppl 1, S73–82. doi:10.1007/s40264-013-0105-z [PubMed: 24166225]

Noren GN, Bergvall T, Ryan PB, Juhlin K, Schuemie MJ, & Madigan D (2013). Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: lessons for developing a risk identification and analysis system. Drug Saf, 36 Suppl 1, S107–121. doi:10.1007/s40264-013-0095-x [PubMed: 24166228]

Noren GN, Caster O, Juhlin K, & Lindquist M (2014). Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Saf, 37(9), 655–659. doi:10.1007/s40264-014-0198-z [PubMed: 25005708]

Overhage JM, Ryan PB, Reich CG, Hartzema AG, & Stang PE (2012). Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc, 19(1), 54–60. doi:10.1136/amiajnl-2011-000376 [PubMed: 22037893]

Overhage JM, Ryan PB, Schuemie MJ, & Stang PE (2013). Desideratum for evidence based epidemiology. Drug Saf, 36 Suppl 1, S5–14. doi:10.1007/s40264-013-0102-2 [PubMed: 24166219]

Overhage JM, Ryan PB, Schuemie MJ, & Stang PE (2015). Authors' reply to Hennessy and Leonard's comment on "Desideratum for evidence-based epidemiology". Drug Saf, 38(1), 105–107. doi:10.1007/s40264-014-0254-8 [PubMed: 25511912]

Prasad V, & Jena AB (2013). Prespecified falsification end points: can they validate true observational associations? Jama, 309(3), 241–242. doi:10.1001/jama.2012.96867 [PubMed: 23321761]

Ramcharran D, Qiu H, Schuemie MJ, & Ryan PB (2017). Atypical antipsychotics and the risk of falls and fractures among older adults: An emulation analysis and an evaluation of additional confounding control strategies. J Clin Psychopharmacol, 37(2), 162–168. doi:10.1097/jcp.0000000000000647 [PubMed: 28225746]

Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, & Schneeweiss S (2012). Oneto-many propensity score matching in cohort studies. Pharmacoepidemiol Drug Saf, 21 Suppl 2, 69–80. doi:10.1002/pds.3263 [PubMed: 22552982]

Reich CG, Ryan PB, & Schuemie MJ (2013). Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Saf, 36 Suppl 1, S181–193. doi:10.1007/s40264-013-0111-1 [PubMed: 24166234]

Reps JM, Schuemie MJ, Suchard MA, Ryan PB, & Rijnbeek PR (2018). Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc, 25(8), 969–975. doi:10.1093/jamia/ocy032 [PubMed: 29718407]

Robins JM, Blevins D, Ritter G, & Wulfsohn M (1992). G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology, 3(4), 319–336. [PubMed: 1637895]

Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55. doi:10.1093/biomet/70.1.41

Rush CJ, Campbell RT, Jhund PS, Petrie MC, & McMurray JJV (2018). Association is not causation: treatment effects cannot be estimated from observational data in heart failure. Europ Heart J, 39(37), 3417–3438. doi:10.1093/eurheartj/ehy407

Ryan PB, Buse JB, Schuemie MJ, DeFalco F, Yuan Z, Stang PE, … Rosenthal N (2018). Comparative effectiveness of canagliflozin, SGLT2 inhibitors and non-SGLT2 inhibitors on the risk of hospitalization for heart failure and amputation in patients with type 2 diabetes mellitus: Areal-world meta-analysis of 4 observational databases (OBSERVE-4D). Diabetes Obes Metab, 20(11), 2585–2597. doi:10.1111/dom.13424 [PubMed: 29938883]

Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, & Hartzema AG (2012). Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med, 31(30), 4401–4415. doi:10.1002/sim.5620 [PubMed: 23015364]

Ryan PB, & Schuemie MJ (2013). Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf, 36 Suppl 1, S171–180. doi:10.1007/s40264-013-0110-2 [PubMed: 24166233]

Ryan PB, Schuemie MJ, Gruber S, Zorych I, & Madigan D (2013). Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug Saf, 36 Suppl 1, S59–72. doi:10.1007/s40264-013-0099-6 [PubMed: 24166224]

Ryan PB, Schuemie MJ, & Madigan D (2013). Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system. Drug Saf, 36 Suppl 1, S95–106. doi:10.1007/s40264-013-0101-3 [PubMed: 24166227]

Ryan PB, Schuemie MJ, Ramcharran D, & Stang PE (2017). Atypical antipsychotics and the risks of acute kidney injury and related outcomes among older adults: A replication analysis and an evaluation of adapted confounding control strategies. Drugs Aging, 34(3), 211–219. doi:10.1007/s40266-016-0430-x [PubMed: 28124262]

Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, & Hartzema AG (2013). Defining a reference set to support methodological research in drug safety. Drug Saf, 36 Suppl 1, S33–47. doi:10.1007/s40264-013-0097-8 [PubMed: 24166222]

Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, … Madigan D (2013). A comparison of the empirical performance of methods for a risk identification system. Drug Saf, 36 Suppl 1, S143–158. doi:10.1007/s40264-013-0108-9 [PubMed: 24166231]

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, & Brookhart MA (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology, 20(4), 512–522. doi:10.1097/EDE.0b013e3181a663cc [PubMed: 19487948]

Schuemie MJ (2011). Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf, 20(3), 292–299. doi:10.1002/pds.2051 [PubMed: 20945505]

Schuemie MJ, Coloma PM, Straatman H, Herings RM, Trifiro G, Matthews JN, … Sturkenboom MC (2012). Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med Care, 50(10), 890–897. doi:10.1097/MLR.0b013e31825f63bf [PubMed: 22929992]

Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RM, Pedersen L, … Sturkenboom MC (2013). Replication of the OMOP experiment in Europe: Evaluating methods for risk identification in electronic health record databases. Drug Saf, 36 Suppl 1, S159–169. doi:10.1007/s40264-013-0109-8 [PubMed: 24166232]

Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, & Suchard MA (2016). Robust empirical calibration of p-values using observational data. Stat Med, 35(22), 3883–3888. doi:10.1002/sim.6977 [PubMed: 27592566]

Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, & Suchard MA (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. ProcNatl Acad SciUSA, 115(11), 2571–2577. doi:10.1073/pnas.1708282114

Schuemie MJ, Madigan D, & Ryan PB (2013). Empirical performance of LGPS and LEOP-ARD: Lessons for developing a risk identification and analysis system. Drug Saf, 36 Suppl 1, S133–142. doi:10.1007/s40264-013-0107-x [PubMed: 24166230]

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(2), 209–218. doi:10.1002/sim.5925 [PubMed: 23900808]

Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, & Suchard MA (2018). Improving re-producibility by using high-throughput observational studies with empirical calibration. Philos Trans A Math Phys Eng Sci, 376(2128). doi:10.1098/rsta.2017.0356

Simpson SE, Madigan D, Zorych I, Schuemie MJ, Ryan PB, & Suchard MA (2013). Multiple self-controlled case series for large-scale longitudinal observational databases. Biometrics, 69(4), 893–902. doi:10.1111/biom.12078 [PubMed: 24117144]

Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N,… Hripcsak G (2019). Comprehensive comparative effectiveness and safety of first-line antihyperten-sive drug classes: a systematic, multinational, large-scale analysis. Lancet, 394 (10211), 1810–1826. DOI:10.1016/S0140-6736(19)32317-7.

Suchard MA, Simpson SE, Zorych I, Ryan PB, & Madigan D (2013). Massive par-allelization of serial inference algorithms for a complex generalized linear model. ACM Trans. Model. Comput. Simul, 23(1), 1–17. doi:10.1145/2414416.2414791

Suchard MA, Zorych I, Simpson SE, Schuemie MJ, Ryan PB, & Madigan D (2013). Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. Drug Saf, 36 Suppl 1, S83–93. doi:10.1007/s40264-013-0100-4 [PubMed: 24166226]

Suissa S (1995). The case-time-control design. Epidemiology, 6(3), 248–253. [PubMed: 7619931]

Tian Y, Schuemie MJ, & Suchard MA (2018). Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol. doi:10.1093/ije/dyy120

Vandenbroucke JP, & Pearce N (2012). Case-control studies: basic concepts. Int J Epidemiol, 41(5), 1480–1489. doi:10.1093/ije/dys147 [PubMed: 23045208]

Vashisht R, Jung K, Schuler A, Banda JM, Park RW, Jin S,… Shah NH (2018). Association of hemoglobin A1c levels with use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in patients with type 2 diabetes treated with Metformin: Analysis From the

Observational Health Data Sciences and Informatics Initiative. JAMA Netw Open, 1(4), e181755. doi:10.1001/jamanetworkopen.2018.1755 [PubMed: 30646124]

Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, & Schuemie MJ (2017). Accuracy of an automated knowledge base for identifying drug adverse reactions. J Biomed Inform, 66, 72–81. doi:10.1016/j.jbi.2016.12.005 [PubMed: 27993747]

Wang Y, Desai M, Ryan PB, DeFalco FJ, Schuemie MJ, Stang PE,… Yuan Z (2017). Incidence of diabetic ketoacidosis among patients with type 2 diabetes mellitus treated with SGLT2 inhibitors and other antihyperglycemic agents. Diabetes Res Clin Pract, 128, 83–90. doi:10.1016/j.diabres.2017.04.004 [PubMed: 28448895]

Weinstein RB, Ryan P, Berlin JA, Matcho A, Schuemie M, Swerdel J, … Fife D (2017). Channeling in the use of nonprescription Paracetamol and Ibuprofen in an electronic medical records database: Evidence and implications. Drug Saf, 40(12), 1279–1292. doi:10.1007/s40264-017-0581-7 [PubMed: 28780741]

Whitaker HJ, Farrington CP, Spiessens B, & Musonda P (2006). Tutorial in biostatistics: The self-controlled case series method. Stat Med, 25(10), 1768–1797. doi:10.1002/sim.2302 [PubMed: 16220518]

Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, & Smith D (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value Health, 13(2), 273–277. doi:10.1111/j.1524-4733.2009.00671.x [PubMed: 19912596]

Yuan Z, DeFalco FJ, Ryan PB, Schuemie MJ, Stang PE, Berlin JA, … Rosenthal N (2018). Risk of lower extremity amputations in people with type 2 diabetes mellitus treated with sodium-glucose co-transporter-2 inhibitors in the USA: A retrospective cohort study. Diabetes Obes Metab, 20(3), 582–589. doi:10.1111/dom.13115 [PubMed: 28898514]

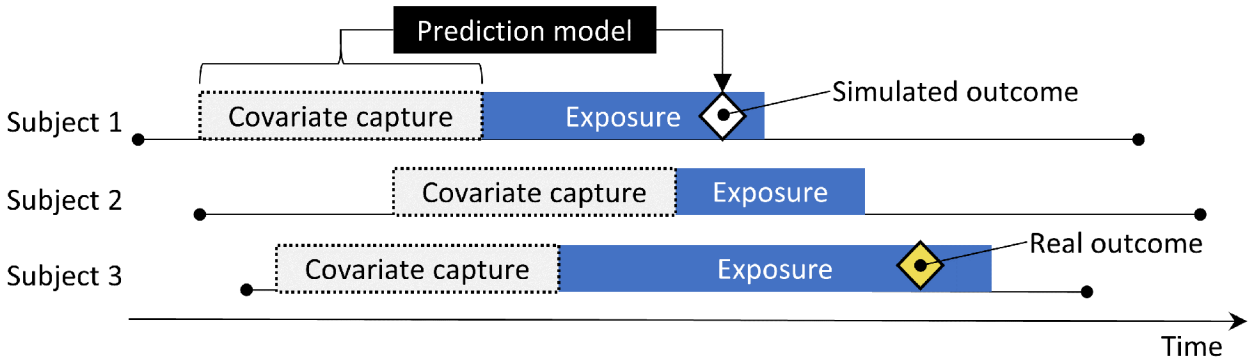

Figure 1:
Synthesizing positive controls from negative controls.

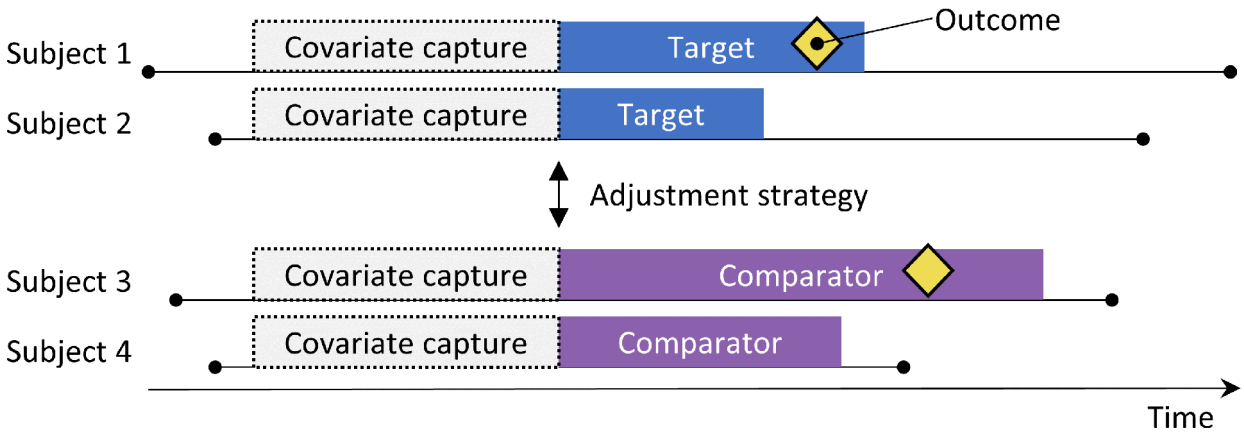

Figure 2:

The new-user cohort design. Subjects observed to initiate the target treatment are compared to those initiating the comparator treatment. To adjust for differences between the two treatment groups several adjustment strategies can be used, such as stratification, matching, or weighting by the propensity score, or by adding baseline characteristics to the outcome model. The characteristics included in the propensity model or outcome model are captured prior to treatment initiation.

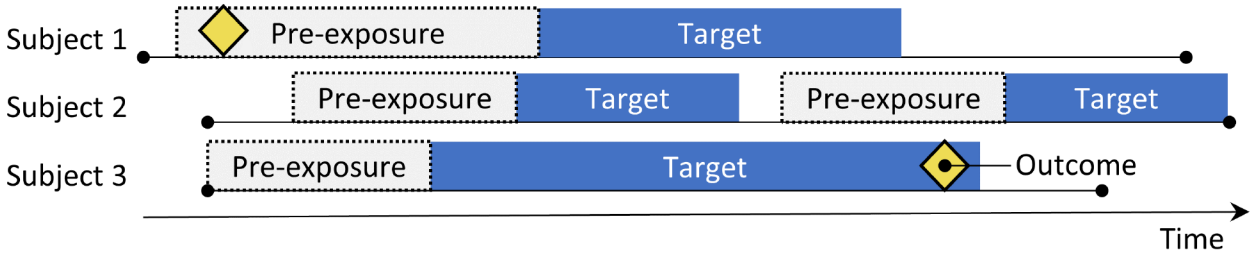

Figure 3:
The self-controlled cohort design. The rate of outcomes during exposure to the target is compared to the rate of outcomes in the time pre-exposure.

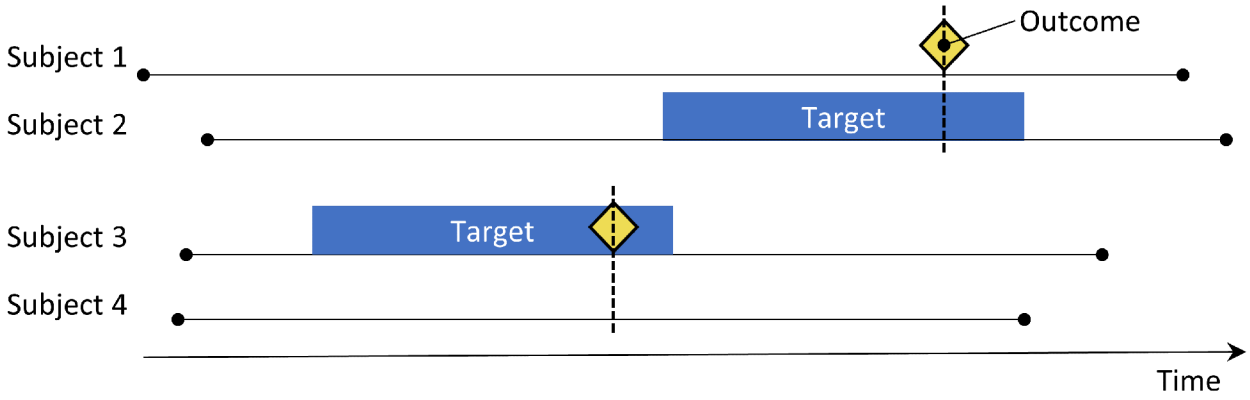

Figure 4:
The case-control design. Subjects with the outcome (“cases”) are compared to subjects without the outcome (“controls”) in terms of their exposure status. Often, cases and controls are matched on various characteristics such as age and sex.

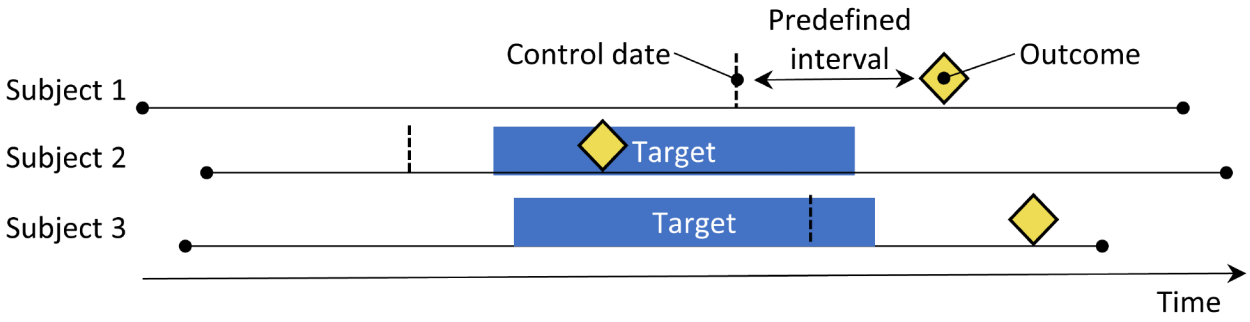

Figure 5:
The case-crossover design. The time around the outcome is compared to a control date set at a predefined interval prior to the outcome date.

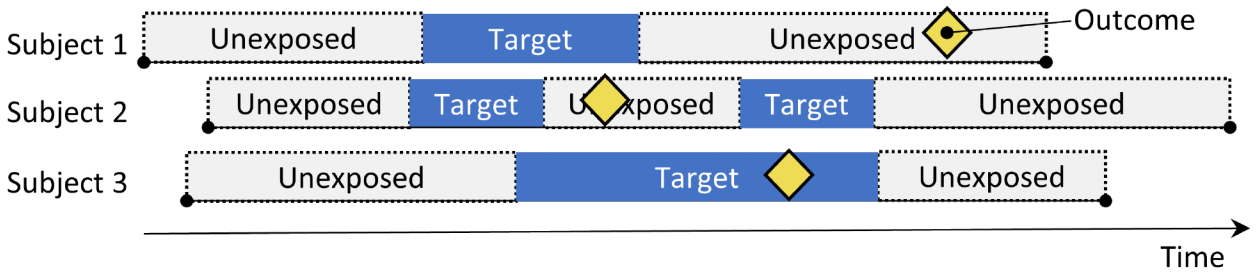

Figure 6:
The Self-Controlled Case Series design. The rate of outcomes during exposure is compared to the rate of outcomes when not exposed.

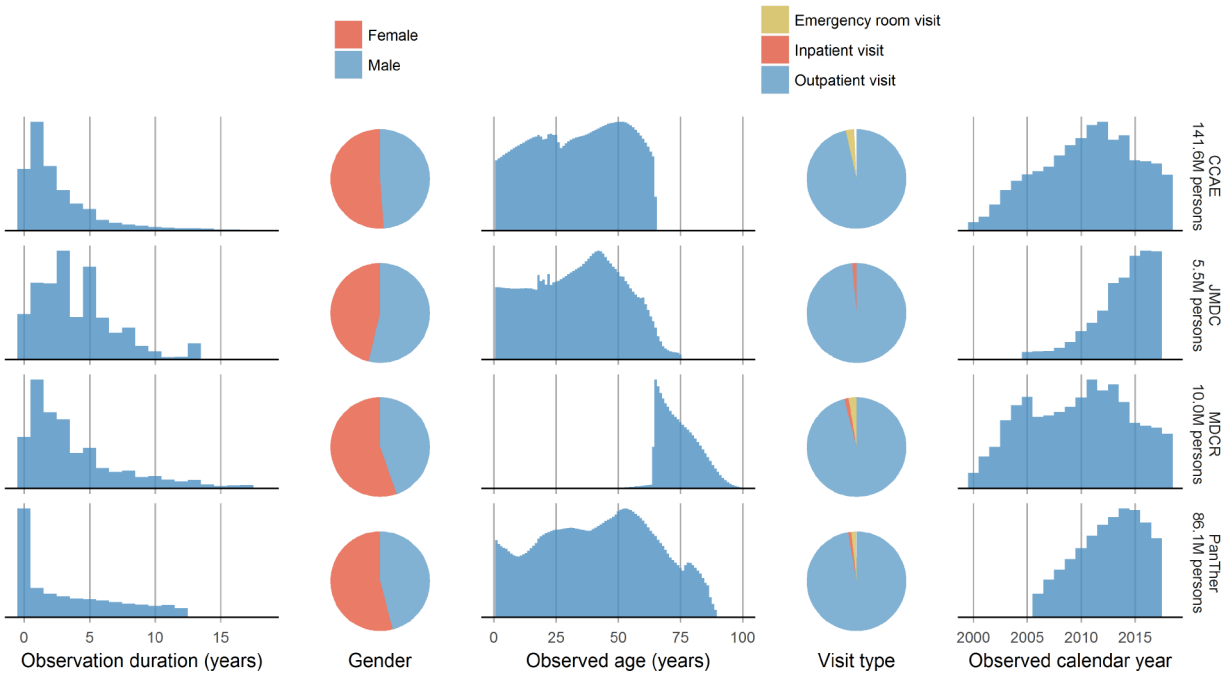

Figure 7:

Summary descriptives of the four databases included in this evaluation. Each row represents a database. Observation duration shows the distribution of the observation time per person. Observed age reflects the number of subjects that were observed for at least one day at that age. The observed calendar year represents the number of subjects that were observed for at least one day in that year. The y-axes of the bar charts show the number of subjects, with scales normalized for each database. The total number of subjects per database is provided on the right.

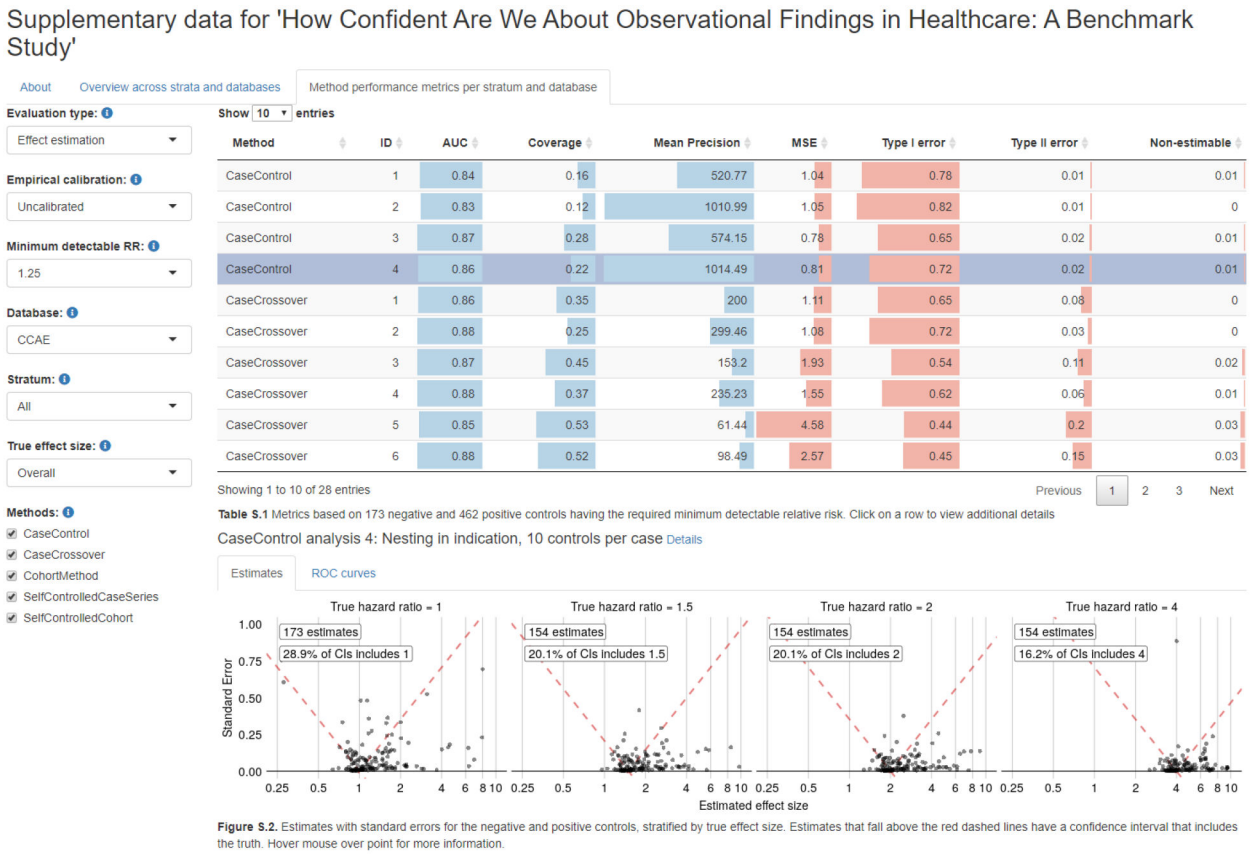

Figure 8:

Screenshot of the Shiny app at http://data.ohdsi.org/MethodEvalViewer/. This app shows the results for all methods on all four databases, and allows filtering the estimates in various ways before computing the performance metrics.

Analysis choices	Effect size (95% CI)
Cohort method	
No PS, simple outcome model	1.14 (1.06-1.22)
1-on-1 matching, unstratified outcome model	1.08 (0.99-1.19)
Variable ratio matching, stratified outcome model	1.21 (1.08-1.36)
Stratification	1.09 (1.01-1.19)
IPTW	1.14 (1.06-1.23)
Var ratio matching + full outcome model	1.11 (0.95-1.30)
Self-controlled cohort	
Time exposed, inc. exp. start date	1.20 (1.17-1.24)
30 days, incl. exp. start date	1.32 (1.26-1.38)
Time exposed, inc. exp. start date, require full obs.	1.25 (1.21-1.30)
Time exposed, ex. exp. start date	1.03 (1.00-1.07)
30 days, ex. exp. start date	0.99 (0.94-1.03)
Time exposed, ex. exp. start date, require full obs.	1.05 (1.01-1.09)
Case-control	
2 controls per case	2.45 (2.35-2.56)
10 controls per case	2.45 (2.38-2.53)
Nesting in indication, 2 controls per case	1.84 (1.78-1.91)
Nesting in indication, 10 controls per case	1.82 (1.77-1.87)
Case-crossover	
Simple case-crossover, -30 days	1.65 (1.55-1.75)
Simple case-crossover, -180 days	1.54 (1.46-1.63)
Nested case-crossover, -30 days	1.46 (1.38-1.55)
Nested case-crossover, -180 days	1.39 (1.33-1.46)
Nested case-time-control, -30 days	1.34 (1.23-1.47)
Nested case-time-control, -180 days	1.29 (1.19-1.39)
Self-controlled case series	
Simple SCCS	1.15 (1.07-1.23)
Including day 0	1.44 (1.35-1.54)
Using pre-exposure window	1.17 (1.09-1.26)
Using age and season	1.14 (1.06-1.22)
Using event-dependent observation	0.99 (0.92-1.06)
Using all other exposures	1.12 (0.95-1.32)

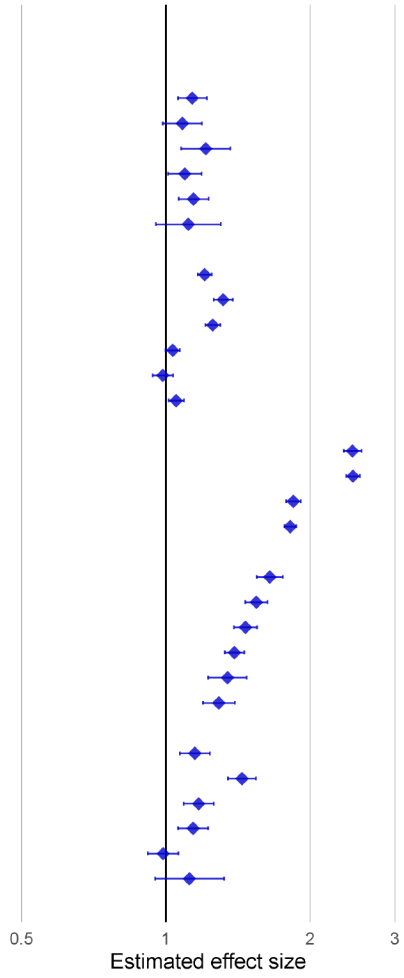

Figure 9:

Effect size estimates (and 95% confidence intervals) for one example control in the CCAE database. We use each analysis variant to estimate the causal effect size of diclofenac on the risk of ingrowing nails. The true effect size is 1. Comparative analyses (i.e., the cohort method) use celecoxib as comparator, and nested analyses restrict to a population with a prior diagnosis of arthralgia. We use the abbreviations “incl.” for “including,” “exp.” for “exposure,” and “ex.” for “excluding.”

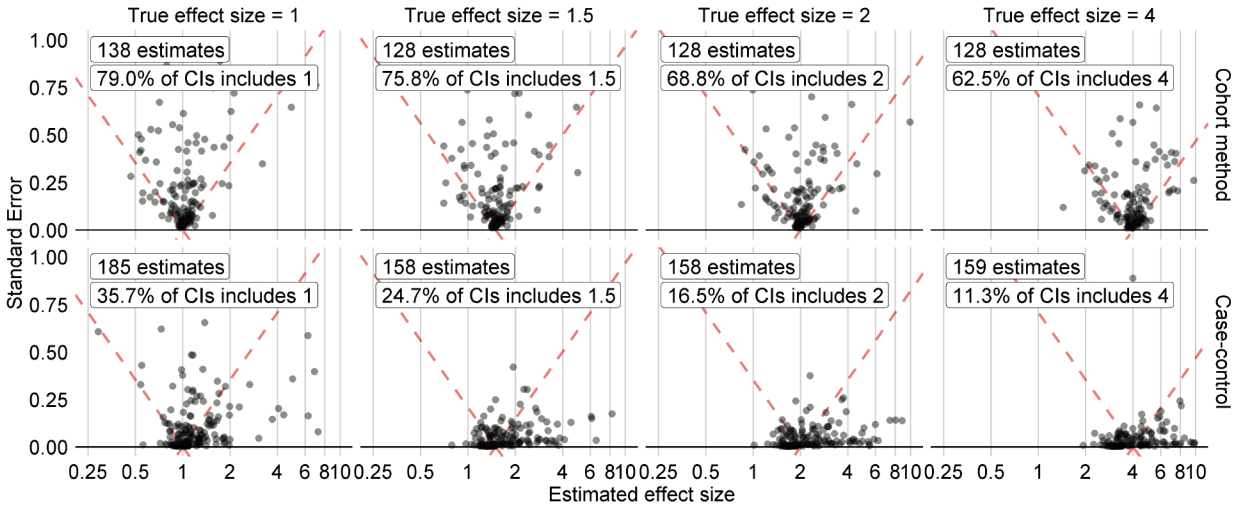

Figure 10:

Effect size estimates and standard errors for all negative and positive controls for two exemplar analyses in the CCAE database. Each dot represents the estimate for a single control. The red dashed line indicates the boundary where the confidence interval no longer contains the truth. The cohort method uses 1-on-1 matching and an unstratified outcome model. The case-control analyses are nested within the relevant indication and select up to 10 controls per case.

Analysis choices	AUC	95% CI coverage	Mean precision	MSE	Type 1 error	Type 2 error	Non-estimable
Cohort method							
No PS, simple outcome model	0.81	0.37	101.67	0.63	0.55	0.11	0.03
1-on-1 matching, unstratified outcome model	0.84	0.73	35.82	0.21	0.16	0.32	0.20
Variable ratio matching, stratified outcome model	0.82	0.75	25.99	2.47	0.17	0.34	0.19
Stratification	0.87	0.57	77.54	0.48	0.27	0.16	0.03
IPTW	0.80	0.34	95.46	0.68	0.57	0.11	0.03
Var ratio matching + full outcome model	0.82	0.77	13.40	0.29	0.05	0.42	0.37
Self-controlled cohort (SCC)							
Time exposed, inc. exp. start date	0.89	0.21	1048.97	0.30	0.66	0.08	0.01
30 days, incl. exp. start date	0.87	0.41	422.25	0.14	0.55	0.11	0.00
Time exposed, inc. exp. start date, require full obs.	0.90	0.24	829.72	0.29	0.58	0.09	0.01
Time exposed, ex. exp. start date	0.94	0.28	1020.85	0.20	0.54	0.09	0.01
30 days, ex. exp. start date	0.89	0.42	402.36	0.13	0.47	0.12	0.00
Time exposed, ex. exp. start date, require full obs.	0.94	0.31	804.21	0.19	0.49	0.09	0.01
Case-control							
2 controls per case	0.84	0.16	520.77	1.04	0.78	0.01	0.01
10 controls per case	0.83	0.12	1010.99	1.05	0.82	0.01	0.00
Nesting in indication, 2 controls per case	0.87	0.28	574.15	0.78	0.65	0.02	0.01
Nesting in indication, 10 controls per case	0.86	0.22	1014.49	0.81	0.72	0.02	0.01
Case-crossover							
Simple case-crossover, -30 days	0.86	0.35	200.00	1.11	0.65	0.08	0.00
Simple case-crossover, -180 days	0.88	0.25	299.46	1.08	0.72	0.03	0.00
Nested case-crossover, -30 days	0.87	0.45	153.20	1.93	0.54	0.11	0.02
Nested case-crossover, -180 days	0.88	0.37	235.23	1.55	0.62	0.06	0.01
Nested case-time-control, -30 days	0.85	0.53	61.44	4.58	0.44	0.20	0.03
Nested case-time-control, -180 days	0.88	0.52	98.49	2.57	0.45	0.15	0.03
Self-controlled case series (SCCS)							
Simple SCCS	0.94	0.45	416.90	0.37	0.38	0.03	0.01
Including day 0	0.89	0.44	466.58	0.60	0.51	0.03	0.00
Using pre-exposure window	0.93	0.44	402.77	0.39	0.46	0.03	0.01
Using age and season	0.95	0.52	410.81	0.37	0.32	0.03	0.01
Using event-dependent observation	0.91	0.28	396.52	0.40	0.61	0.08	0.01
Using all other exposures	0.97	0.51	302.97	0.26	0.27	0.06	0.01

Figure 11:

Performance metrics on the CCAE database computed using controls with MDRR < 1.25. We use the abbreviations “incl.” for “including,” “exp.” for “exposure,” and “ex.” for “excluding.”

Analysis choices	AUC	95% CI coverage	Mean precision	MSE	Type 1 error	Type 2 error	Non-estimable
Cohort method							
No PS, simple outcome model	0.78	0.92	3.35	0.51	0.08	0.60	0.19
1-on-1 matching, unstratified outcome model	0.78	0.89	9.35	0.33	0.08	0.37	0.31
Variable ratio matching, stratified outcome model	0.78	0.92	7.45	0.34	0.06	0.41	0.30
Stratification	0.79	0.90	8.68	0.41	0.10	0.35	0.26
IPTW	0.77	0.91	2.77	0.55	0.08	0.64	0.19
Var ratio matching + full outcome model	0.78	0.93	7.76	0.31	0.04	0.43	0.45
Self-controlled cohort (SCC)							
Time exposed, inc. exp. start date	0.87	0.92	12.28	0.49	0.09	0.29	0.07
30 days, incl. exp. start date	0.87	0.92	12.25	0.19	0.09	0.34	0.04
Time exposed, inc. exp. start date, require full obs.	0.88	0.92	13.91	0.47	0.09	0.27	0.08
Time exposed, ex. exp. start date	0.87	0.94	11.75	0.22	0.08	0.18	0.23
30 days, ex. exp. start date	0.89	0.93	14.46	0.14	0.08	0.29	0.05
Time exposed, ex. exp. start date, require full obs.	0.87	0.94	13.04	0.20	0.07	0.17	0.24
Case-control							
2 controls per case	0.84	0.92	7.35	0.59	0.08	0.48	0.04
10 controls per case	0.84	0.92	7.37	0.62	0.08	0.48	0.01
Nesting in indication, 2 controls per case	0.87	0.91	12.90	0.54	0.10	0.34	0.01
Nesting in indication, 10 controls per case	0.86	0.92	12.28	0.55	0.10	0.35	0.02
Case-crossover							
Simple case-crossover, -30 days	0.87	0.92	10.17	0.46	0.08	0.43	0.04
Simple case-crossover, -180 days	0.85	0.93	10.79	0.61	0.07	0.44	0.02
Nested case-crossover, -30 days	0.87	0.92	11.78	0.41	0.08	0.34	0.06
Nested case-crossover, -180 days	0.86	0.93	12.26	0.55	0.06	0.36	0.03
Nested case-time-control, -30 days	0.87	0.92	10.76	0.38	0.06	0.38	0.05
Nested case-time-control, -180 days	0.87	0.94	10.80	0.21	0.07	0.35	0.19
Self-controlled case series (SCCS)							
Simple SCCS	0.90	0.95	15.78	0.17	0.07	0.20	0.20
Including day 0	0.87	0.93	11.67	0.53	0.08	0.39	0.09
Using pre-exposure window	0.90	0.95	12.98	0.19	0.08	0.24	0.20
Using age and season	0.91	0.94	22.15	0.16	0.08	0.19	0.20
Using event-dependent observation	0.88	0.95	12.12	0.18	0.08	0.24	0.20
Using all other exposures	0.91	0.95	21.98	0.16	0.05	0.11	0.20

Figure 12:

Performance metrics on the CCAE database after empirical calibration computed using controls with MDRR < 1.25. We use the abbreviations “incl.” for “including,” “exp.” For “exposure,” and “ex.” for “excluding.”

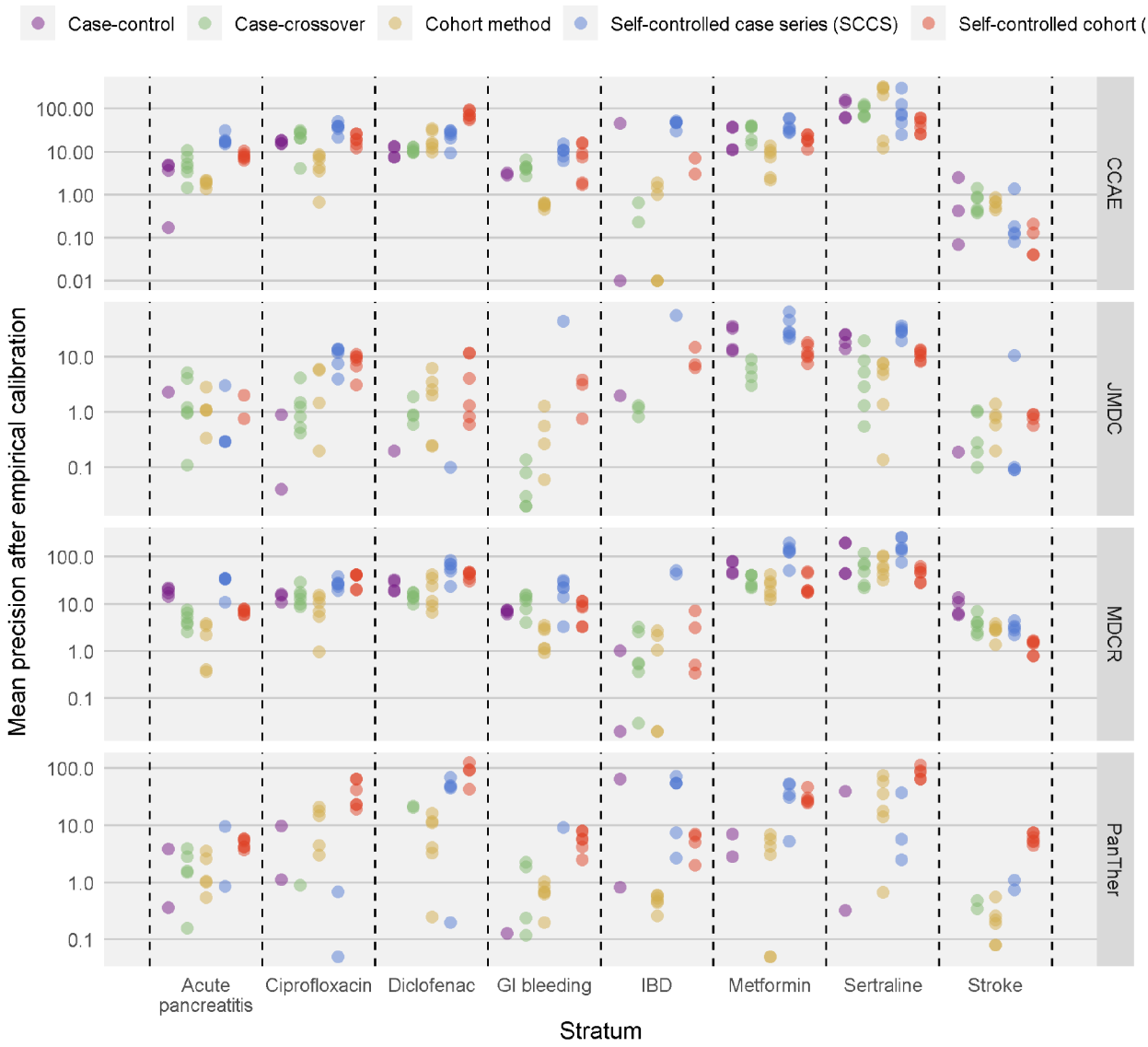

Figure 13:

Performance on the CCAE database after empirical calibration computed using controls with MDRR < 1.25. The dotplot shows mean precision ($1/SE^2$), stratified by main exposure and outcome, and by database. Each dot represents the performance of an analysis variant. Because precision depends, inter alia, on sample size, which differs for the different databases, we used varying scales for the y-axis. Note that some methods did not produce any estimates for some strata-database combinations, and therefore the number of dots is not always the same.

Table 1:

Example entries in the gold standard.

Target	Comparator	Nesting cohort	Outcome	True effect size
Brinzolamide	Levobunolol	Glaucoma	Acute pancreatitis	1.0
Cevimeline	Pilocarpine	Sjogren's syndrome	Acute pancreatitis	1.0
Diclofenac	Celecoxib	Arthralgia	Acute stress disorder	1.0
Diclofenac	Celecoxib	Arthralgia	Ingrowing nail	1.0

Table 2:

Analysis variants of the new-user cohort method included in the evaluation.

Variant description	PS adjustment	Stratified outcome model	Add covariates to the model
No PS, simple outcome model	none	no	no
1-on-1 matching, unstratified model	1-on-1 matching	no	no
Variable ratio matching, stratified model	variable ratio matching	yes	no
Stratification	stratification	yes	no
IPTW	trimming + IPTW	no	no
Var. ratio matching + full model	variable ratio matching	yes	yes

Table 3:

Analysis variants of the self-controlled cohort design included in the evaluation. We use the abbreviations "incl." for "including," "exp." for "exposure," and "ex." for "excluding."

Description	Time at risk	Exposure start date	Require full observation
Time exposed, incl. exp. start date	Time exposed	In exposure window	no
30 days, incl. exp. start date	30 days	In exposure window	no
Time exposed, incl. exp. start date, require full obs.	Time exposed	In exposure window	yes
Time exposed, ex. exp. start date Time exposed	Excluded	no	
30 days, ex. exp. start date	30 days	Excluded	no
Time exposed, ex. exp. start date, require full obs.	Time exposed	Excluded	yes

Table 4:

Analysis variants of the self-controlled cohort design included in the evaluation.

Description	Controls per case	Nesting in indication
2 controls per case	2	no
10 controls per case	10	no
Nesting in indication, 2 controls per case	2	yes
Nesting in indication, 10 controls per case	10	yes

Table 5:

Analysis variants of the self-controlled cohort design included in our evaluation.

Description	Nesting in indication	Control window	Case-time-control
Simple case-crossover, −30 days	FALSE	−30	no
Simple case-crossover, −180 days	FALSE	−180	no
Nested case-crossover, −30 days	TRUE	−30	no
Nested case-crossover, −180 days	TRUE	−180	no
Nested case-time-control, −30 days	TRUE	−30	yes
Nested case-time-control, −180 days	TRUE	−180	yes

Table 6:

Analysis variants of the self-controlled case series design included in the evaluation.

Description	Include start day	Pre-exposure window	Age and season	Event-dependent observation	All other exposures
Simple SCCS	no	no	no	no	no
Including exposure start day	yes	no	no	no	no
Using pre-exposure window	no	yes	no	no	no
Using age and season	no	no	yes	no	no
Using event-dependent observation	no	no	no	yes	no
Using all other exposures	no	no	no	no	yes

Table 7:

Some key median counts for the various analyses variants across all 200 negative controls in the CCAE database. We use the abbreviations “incl.” for “including,” “exp.” for “exposure,” and “ex.” for “excluding.”

Cohort method	**target subjects**	**comparator subjects**	**target outcomes**	**comparator outcomes**
No PS, simple outcome model	926,669	573,353	464	129
1-on-1 matching, unstratified outcome model	85,773	85,773	86	42
Variable ratio matching, stratified outcome model	85,773	433,152	86	95
Stratification	926,669	573,353	464	129
IPTW	880,335	544,685	436	126
Var ratio matching + full outcome model	85,773	433,152	86	95
Self-controlled cohort (SCC)	**subjects**	**exposures**	**exposed outcomes**	**unexposed outcomes**
Time exposed, incl. exp. start date	2,109,605	4,021,441	2,857	1,320
30 days, incl. exp. start date	2,109,605	4,021,441	1,705	756
Time exposed, incl. exp. start date, require full obs.	1,867,062	3,788,069	2,453	1,140
Time exposed, ex. exp. start date	2,109,605	4,021,441	2,362	1,320
30 days, ex. exp. start date	2,109,605	4,021,441	1,428	756
Time exposed, ex. exp. start date, require full obs.	1,864,181	3,783,178	2,061	1,139
Case-control	**cases**	**controls**	**exposed cases**	**unexposed controls**
2 controls per case	151,903	303,805	1,666	1,118
10 controls per case	151,903	1,519,025	1,666	5,745
Nesting in indication, 2 controls per case	61,917	123,834	1,036	811
Nesting in indication, 10 controls per case	61,917	619,170	1,036	4,101
Case-crossover	**cases**	**exposed outcomes**	**unexposed outcomes**	
Simple case-crossover, −30 days	242,160	2,718	1,586	
Simple case-crossover, −180 days	242,160	2,718	1,429	
Nested case-crossover, −30 days	61,917	1,248	744	
Nested case-crossover, −180 days	61,917	1,248	642	
Nested case-time-control, −30 days	61,917	1,248	744	
Nested case-time-control, −180 days	61,917	1,248	642	
Self-controlled case-series (SCCS)	**cases**	**exposed subjects**	**unexposed subjects**	
Simple SCCS	9,088	9,060	854	
Including day 0	9,185	9,176	1,062	
Using pre-exposure window	9,188	9,060	854	
Using age and season	18,601	9,287	866	
Using event-dependent observation	9,088	9,060	854	
Using all other exposures	135,724	9,285	866	