Drug Saf (2013) 36 (Suppl 1):S95–S106
DOI 10.1007/s40264-013-0101-3

ORIGINAL RESEARCH ARTICLE

Empirical Performance of a Self-Controlled Cohort Method: Lessons for Developing a Risk Identification and Analysis System

Patrick B. Ryan · Martijn J. Schuemie · David Madigan

Abstract

Background Observational healthcare data offer the potential to enable identification of risks of medical products, but appropriate methodology has not yet been defined. The self-controlled cohort method, which compares the post-exposure outcome rate with the pre-exposure rate among an exposed cohort, has been proposed as a potential approach for risk identification but its performance has not been fully assessed.

Objectives To evaluate the performance of the self-controlled cohort method as a tool for risk identification in observational healthcare data.

Research Design The method was applied to 399 drug-outcome scenarios (165 positive controls and 234 negative controls across 4 health outcomes of interest) in 5 real observational databases (4 administrative claims and 1 electronic health record) and in 6 simulated datasets with no effect and injected relative risks of 1.25, 1.5, 2, 4, and 10, respectively.

Measures Method performance was evaluated through area under ROC curve (AUC), bias, and coverage probability.

Results The self-controlled cohort design achieved strong predictive accuracy across the outcomes and databases under study, with the top-performing settings exceeding AUC >0.76 in all scenarios. However, the estimates generated were observed to be highly biased with low coverage probability.

Conclusions If the objective for a risk identification system is one of discrimination, the self-controlled cohort method shows promise as a potential tool for risk identification. However, if a system is intended to generate effect estimates to quantify the magnitude of potential risks, the self-controlled cohort method may not be suitable, and requires substantial calibration to be properly interpreted under nominal properties.

The OMOP research used data from Truven Health Analytics (formerly the Health Business of Thomson Reuters), and includes MarketScan® Research Databases, represented with MarketScan Lab Supplemental (MSLR, 1.2 m persons), MarketScan Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan Multi-State Medicaid (MDCD, 10.8 m persons), MarketScan Commercial Claims and Encounters (CCAE, 46.5 m persons). Data also provided by Quintiles® Practice Research Database (formerly General Electric's Electronic Health Record, 11.2 m persons) database. GE is an electronic health record database while the other four databases contain administrative claims data.

P. B. Ryan (✉)
Janssen Research and Development LLC,
1125 Trenton-Harbourton Road, Room K30205,
PO Box 200, Titusville, NJ 08560, USA
e-mail: ryan@omop.org

M. J. Schuemie
Department of Medical Informatics, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands

D. Madigan
Department of Statistics, Columbia University, New York, NY, USA

P. B. Ryan · M. J. Schuemie · D. Madigan
Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD, USA

1 Background

Observational healthcare data, such as administrative claims and electronic health records, offer the potential to greatly enhance the understanding of the effects of medical products. These data sources have already been actively

used to conduct pharmacoepidemiology studies of specific hypotheses of the effects of particular medical product exposure and subsequent health outcomes of interest. In recent years, there has been interest in exploring whether these same data resources could also be used to support the identification of potential risks on a more rapid, proactive and systematic basis. In 2007, Congress passed the Food and Drug Administration (FDA) Amendment Act, which called for the establishment of an "active postmarket risk identification and analysis system" with access to patient-level observational data from 100 million lives by 2012 [1]. It is envisioned that such a system would "use sophisticated statistical methods to actively search for patterns in prescription, outpatient, and inpatient data systems that might suggest the occurrence of an adverse event, or safety signal, related to drug therapy" [2]. Software systems have already been developed within industry to support these types of pharmacovigilance analysis [3]on observational data, but the performance of the methods within these solutions have not been fully evaluated.

In order to appropriately use in a risk identification system, the necessary empirical evidence base around the operating characteristics of the analysis approach needs to be first established. One analysis approach currently used in practice is the self-controlled cohort method, which estimates the strength of association by comparing the post-exposure incidence rate with the pre-exposure incidence rate among the patients exposed to the target drug of interest [4]. The self-controlled cohort method represents a unique approach in that is does not use an external active comparator group like the new user cohort design, and it also does not restrict its analysis to within-person comparisons among the subset of patients with both exposure and outcome like the self-controlled case series design.

In this study, we evaluated the performance of a self-controlled cohort method as a potential analytical method for a risk identification system. We tested the self-controlled cohort method in 5 real observational healthcare databases and 6 simulated datasets, retrospectively studying the predictive accuracy of the method when applied to a collection of 165 positive controls and 234 negative controls across 4 outcomes: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding. We estimate how well the method can be expected to identify true effects and discriminate from false findings and explore the statistical properties of the estimates the design generates. With this empirical basis in place, the self-controlled cohort method can be evaluated to determine whether it represents a potential alternative tool to be considered in establishing a risk identification and analysis system to study the effects of medical products.

2 Methods

2.1 Overview of self-controlled cohort method

In the self-controlled cohort design, the only patients used in the analysis are those patients with at least one exposure to the target drug. The design is self-controlled in that the cohort serves as its own control, by comparing unexposed time among the exposed cohort with the exposed time. Within this exposed cohort, post-exposure and pre-exposure incidence rates are calculated. The post-exposure incidence rate is estimated as the number of outcomes observed during the post-exposure time-at-risk, divided by the total length of time-at-risk across the exposed cohort. The pre-exposure incidence rate is estimated as the number of outcomes observed during the pre-exposure control period, divided by the total length of control period across the exposed cohort. An incidence rate ratio (IRR) is then calculated as the post-exposure incidence rate, divided by the pre-exposure incidence rate. This is expressed simply as:

$$IRR = \frac{(x_0/t_0)}{(x_1/t_t)}$$

where t_0 is post-exposure person time-at-risk, t_1 is pre-exposure person-time in the control period, x_0 is the number of outcomes observed during the cohort post-exposure time-at-risk, and x_1 is the number of outcomes observed during the cohort pre-exposure control period. We assume the incidence rate ratio is a ratio of two Poisson distributed rates, and use the closed form solution by Graham et al. [5] to calculate the associated confidence interval (CI), where Z_α is the Z statistic at the desired α is the type I error rate:

$$IRRCI = \left(\frac{t_1}{t_0}\right) * \left(\frac{2x_1x_0 + z_{\frac{\alpha}{2}}^2 (x_1 + x_0) \pm \sqrt{z_{\frac{\alpha}{2}}^2 (x_1 + x_0) * \left(4x_1x_0 + z_{\frac{\alpha}{2}}^2 (x_1 + x_0)\right)}}{2x_1^2}\right)$$

The self-controlled cohort method is notably distinct from the self-controlled series design in two regards: (1) it does not estimate the effect size by conditioning on individual

patients, and (2) it uses exposed and unexposed time from the entire exposed cohort to estimate the incidence rate, rather than using the restriction that the self-controlled case series design imposes which limits its analysis to patients with the event of interest. Whereas the self-controlled case series design identified cases and each case serves as an individual control, the self-controlled cohort design produced estimates for the entire cohort population, with the cohort serving as its own control. The self-controlled cohort method is distinguished from the new user cohort design in its lack of an external comparator group to estimate a baseline rate of events, using the unexposed time within the exposed cohort in its place.

Several analysis choices are required within the self-controlled cohort method to enable a fully specific analysis. Within the open-source software implementation of the design titled 'observational screening', publicly available at http://omop.org/MethodsLibrary, six analysis choices are parameterized and were evaluated in this research. They are illustrated in Fig. 1, and are described below:

1. Should the analysis consider only the first period of persistent exposure, or should all exposures (including prevalence use) be included?
2. Should the analysis consider only the first occurrence of the target outcome of interest (focusing on incident events), or should all occurrences be counted as potential outcomes?
3. Defining post-exposure time-at-risk: what duration of time relative to the start of exposure should be considered as a period for which outcomes are identified? The post-exposure time-at-risk can be a fixed time window, such as 30 days after exposure start, or can be defined relative to the length of exposure, such as using the length of exposure plus an additional 30 day surveillance window after the end of exposure, or can be defined as all available observation time after exposure start.
4. Should the index date (i.e. start of exposure) be included as part of the post-exposure time-at-risk?
5. Defining pre-exposure control period: what duration of time looking back relative to the start of exposure should be considered as a period for which outcomes are identified as background conditions? The pre-exposure control period can be a fixed time window, such as 30 days before exposure start or 180 days before exposure start, or can be defined relative to the length of exposure, such as using the length of

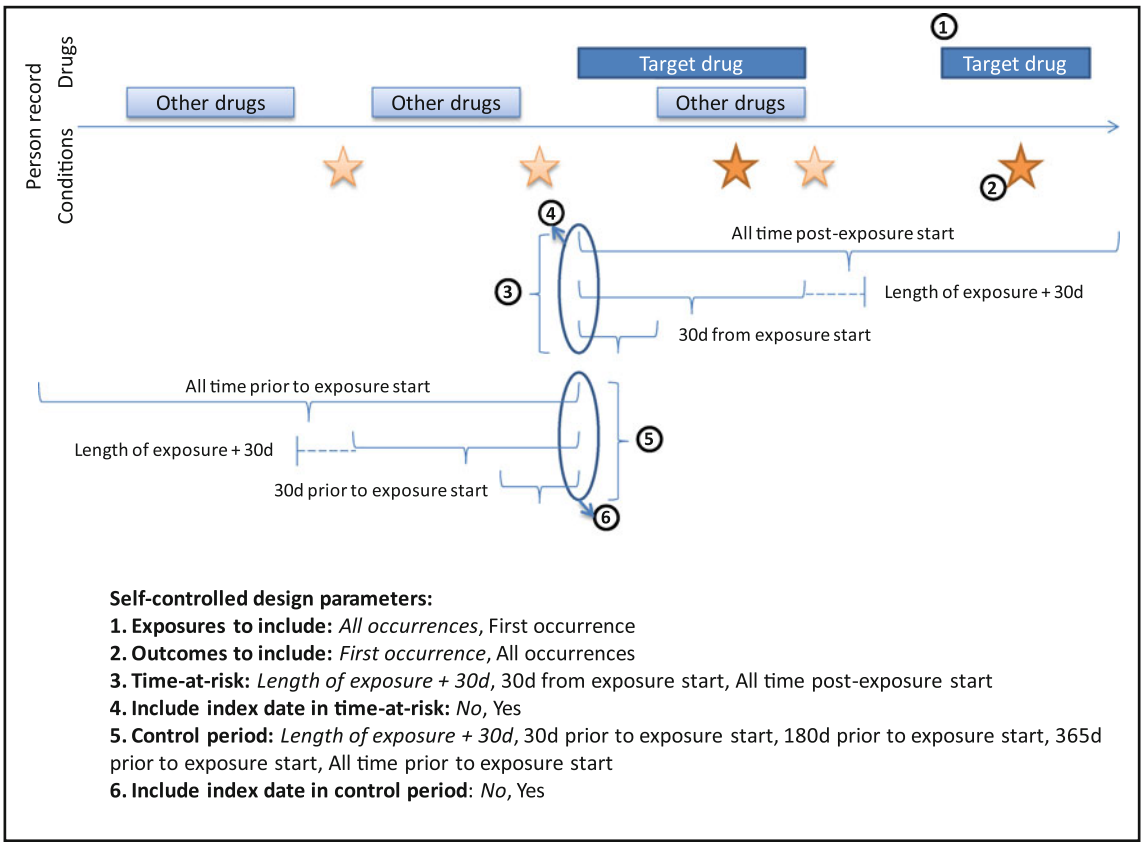


Fig. 1 Parameters within self-controlled cohort design

exposure plus an additional 30 day surveillance window and looking back fur that duration prior to exposure start, or can be defined as all available observation time prior to exposure start.

6. Should the index date be included as part of the pre-exposure control period?

In this study, 38 unique analyses comprising combinations of the 6 analysis choices were evaluated.

2.2 Experiment Design

The study was conducted against five observational healthcare databases to allow evaluation of performance across different populations and data capture processes: MarketScan® Lab Supplemental (MSLR, 1.2 m persons), MarketScan® Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan Multi-State Medicaid (MDCD, 10.8 m persons), MarketScan® Commercial Claims and Encounters (CCAE, 46.5 m persons), and the General Electric Centricity™ (GE, 11.2 m persons) database. GE is an electronic health record (EHR) database; the other four databases contain administrative claims data. A 10 m-person simulated dataset was also constructed using the OSIM2 simulator [6] to model the MSLR database, and replicated 6 times to allow for injection of signals of known size (relative risk = 1, 1.25, 1.5, 2, 4, 10). The data used is described in more detail elsewhere [7].

The method was executed using all 38 analyses against 399 drug-outcome pairs to generate an effect estimate and standard effort for each pair and combination of analysis choices. These test cases include 165 'positive controls'—active ingredients with evidence to suspect a positive association with the outcome—and 234 'negative controls'—active ingredients with no evidence to expect a causal effect with the outcome, and were limited to four outcomes: acute liver injury, acute myocardial infarction, acute renal failure, and upper gastrointestinal bleeding. The full set of test cases and its construction is described elsewhere [8]. For every database we restricted the evaluation to those drug-outcome pairs with sufficient power to detect a relative risk of RR $\leq$1.25, based on the age-by-gender-stratified drug and outcome prevalence estimates [9].

2.3 Metrics

The estimates and associated standard errors for all of the analyses are available for download at: http://omop.org/Research.

To gain insight into the ability of a method to distinguish between positive and negative controls the IRR estimates were used to compute the Area Under the receiver operator characteristics Curve (AUC), a measure of predictive accuracy [10]: an AUC of 1 indicates a perfect prediction of which test cases are positive, and which are not. An AUC of 0.5 is equivalent to random guessing.

Often we are not only interested in whether there is an effect or not, but would also like to know the magnitude of the effect. However, in order to evaluate whether a method produces correct relative risk estimates, we must know the true effect size. In real data, this true effect size is never known with great accuracy for positive controls, and we must restrict our analysis to the negative controls where we assume that the true relative risk is 1. Fortunately, in the simulated data sets we do know the true relative risk for all injected signals. Using both the negative controls in real data, and injected signals in the simulated data, we compute the coverage probability: the percentage of confidence intervals that contain the true relative risk. In case of an unbiased estimator with accurate confidence interval estimation we would expect the coverage probability to be 95 %.

Lastly, we are interest in to what extend each parameter can influence the estimated relative risk. For every parameter, we evaluated how much the estimated relative risk changed as a consequence of changing a single parameter while keeping all other parameters constant.

3 Results

3.1 Predictive Accuracy of All Settings

Figure 2 highlights the predictive accuracy, as measured by AUC, of all self-controlled cohort analyses across the 4 outcomes and 5 databases. For each outcome-database scenario we identified analysis choices that yielded the highest AUC, as listed in Table 1. An optimal analysis (OS: 403002) had the highest predictive accuracy for discriminating test cases for acute liver injury in CCAE (AUC = 0.79), acute myocardial infarction in MSLR (AUC = 0.80), and GI bleeding in MSLR (AUC = 0.83). A different analysis (OS: 408013) used only first occurrences of both exposure and outcome and defined the comparison as all-time post-exposure vs. all-time pre-exposure, and yielded the highest AUC in 4 outcome-database scenarios: acute myocardial infarction in CCAE (AUC = 0.85), acute myocardial infarction in GE (AUC = 0.86), GI bleed in GE (AUC = 0.82), and acute renal failure in MDCD (AUC = 0.82).

The optimal analyses in the GE database all involved focus on first exposure and first occurrence of the outcome, but other databases did not follow the same pattern. The

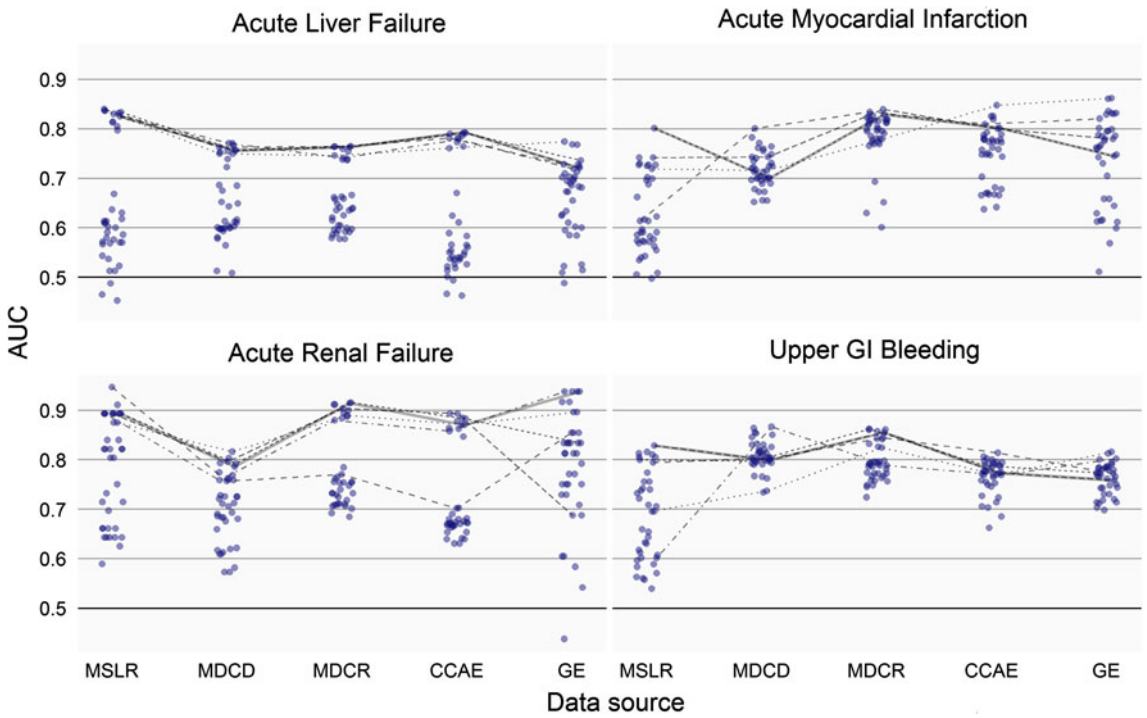


Fig. 2 Area under ROC curve (AUC) for self-controlled cohort parameters, by outcome and database. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity. *Each dot* represents one of the 38 unique parameter combinations of the self-controlled design. The *solid grey line* highlights the parameter that had the highest average area under ROC curve (AUC) across all 20 outcome-database scenarios. The *dashed lines* identify each setting with the highest AUC in at least one database within each outcome

dashed and dotted lines in Fig. 2 indicate the performance of these top-performing setting across databases for the same outcome, showing that the optimal setting for one database can sometimes perform poorly when used in another database for the same outcome.

3.2 Overall Optimal Settings

The analysis with the best average performance across the 20 outcome-database scenarios is highlighted in the shaded grey line, and represents analysis OS:403002. OS:403002 is the unique identifier that reflects the parameter combination which uses all occurrences of drug exposure and all occurrences of the outcome, defined both the time-at-risk and the control period as the length of exposure +30 days, and classified events occurring on the index exposure date as within the time-at-risk. This analysis was observed to have high predictive accuracy across all scenarios, with an average AUC = 0.83 and having AUC $\geq$0.70 in all scenarios. In the remainder of this paper we will use OS:403002 as the representative settings for the self-controlled cohort method.

The Appendix contains the effect estimates for all test cases across the 5 databases using this optimal analysis (OS: 403002). To illustrate patterns in these findings, we discuss four specific test cases for acute liver injury, as shown in Fig. 3. One drug known to be a causative agent is isoniazid [11], which was used as a positive control. The association between isoniazid and acute liver injury was consistently one of the largest effects observed using the self-controlled cohort design, with all 5 databases which generating IRR >2 that was statistically significant at conventional $p < 0.05$. In contrast, erythromycin is another drug thought to be associated with acute liver injury and used as a positive control [12], but the IRR estimates for erythromycin-acute liver injury were near IRR = 1 in all databases except GE. That is, the rate of acute liver injury prior to erythromycin is similar to the rate of acute liver injury during the period of exposure. So while isoniazid illustrates the opportunity for use of this method in a risk identification and would likely be classified as a 'true positive' under most decision thresholds based either on effect size or statistical significance, erythromycin demonstrates the risk of a 'false negative' finding. Erythromycin is an antibiotic used to treat serious infections that frequently require hospitalization and are often associated with hepatic issues, so a potential explanation for this 'false negative' finding is that the background rate of acute livery

Table 1 Optimal parameter settings for self-controlled cohort design, by outcome and database

Data source	Acute liver injury	Acute kidney injury	Acute myocardial infarction	GI bleed
CCAE	AUC = 0.79 (OS: 403002)	AUC = 0.89 (OS: 404002)	AUC = 0.85 (OS: 408013)	AUC = 0.81 (OS: 407002)
	1: All occurrences	1: All occurrences	1: First occurrence	1: All occurrences
	2: All occurrences	2: All occurrences	2: First occurrence after exposure	2: First occurrence
	3: Length of exposure + 30 days	3: Length of exposure + 30 days	3: All time post- exposure start	3: Length of exposure + 30 days
	4: Yes	4: No	4: No	4: No
	5: Length of exposure + 30 days	5: Length of exposure + 30 days	5: All time prior to exposure start	5: Length of exposure + 30 days
	6: No	6: Yes	6: No	6: No
MDCD	AUC = 0.77 (OS: 409013)	AUC = 0.82 (OS: 408013)	AUC = 0.80 (OS: 407004)	AUC = 0.87 (OS: 401004)
	1: First occurrence	1: First occurrence	1: All occurrences	1: All occurrences
	2: First occurrence	2: First occurrence after exposure	2: First occurrence	2: All occurrences
	3: All time post-exposure start	3: All time post-exposure start	3: Length of exposure + 30 days	3: Length of exposure + 30 days
	4: No	4: No	4: No	4: No
	5: All time prior to exposure start	5: All time prior to exposure start	5: 365 days prior to exposure start	5: 365 days prior to exposure start
	6: No	6: No	6: No	6: No
MDCR	AUC = 0.76 (OS: 401002)	AUC = 0.92 (OS: 401002)	AUC = 0.84 (OS: 407002)	AUC = 0.86 (OS: 402002)
	1: All occurrences	1: All occurrences	1: All occurrences	1: All occurrences
	2: All occurrences	2: All occurrences	2: First occurrence	2: All occurrences
	3: Length of exposure + 30 days	3: Length of exposure + 30 days	3: Length of exposure + 30 days	3: Length of exposure + 30 days
	4: No	4: No	4: No	4: Yes
	5: Length of exposure + 30 days	5: Length of exposure + 30 days	5: Length of exposure + 30 days	5: Length of exposure + 30 days
	6: No	6: No	6: No	6: Yes
MSLR	AUC = 0.84 (OS: 406002)	AUC = 0.95 (OS: 405004)	AUC = 0.80 (OS: 403002)	AUC = 0.83 (OS: 403002)
	1: All occurrences	1: First occurrence	1: All occurrences	1: All occurrences
	2: First occurrence after exposure	2: All occurrences	2: All occurrences	2: All occurrences
	3: Length of exposure + 30 days	3: Length of exposure + 30 days	3: Length of exposure + 30 days	3: Length of exposure + 30 days
	4: No	4: No	4: Yes	4: Yes
	5: Length of exposure + 30 days	5: 365 days prior to exposure start	5: Length of exposure + 30 days	5: Length of exposure + 30 days
	6: No	6: No	6: No	6: No
GE	AUC = 0.77 (OS: 409002)	AUC = 0.94 (OS: 409002)	AUC = 0.86 (OS: 408013)	AUC = 0.82 (OS: 408013)
	1: First occurrence	1: First occurrence	1: First occurrence	1: First occurrence
	2: First occurrence	2: First occurrence	2: First occurrence after exposure	2: First occurrence after exposure
	3: Length of exposure + 30 days	3: Length of exposure + 30 days	3: All time post-exposure start	3: All time post-exposure start
	4: No	4: No	4: No	4: No
	5: Length of exposure + 30 days	5: Length of exposure + 30 days	5: All time prior to exposure start	5: All time prior to exposure start
	6: No	6: No	6: No	6: No

AUC area under ROC curve; database abbreviations: *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity. Parameter settings: 1. Exposures to include; 2. Outcomes to include; 3. Time-at-risk; 4. Include index date in time-at-risk; 5. Control period; 6. Include index date in control period

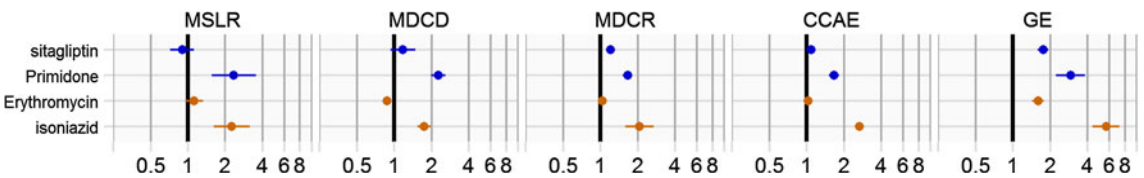


Fig. 3 IRR and 95 % confidence interval for 4 example drugs and acute liver injury, across databases, using the overall optimal settings. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity, *Blue* negative controls, *Orange* positive controls; *each line* represents point estimate and 95 % confidence interval for the drug–outcome pair in a particular database

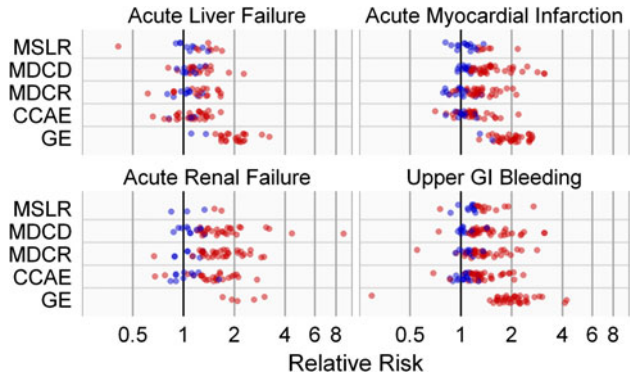


Fig. 4 Bias estimates for the negative control drugs, where the assumed true relative risk is one, using those settings that achieved the highest AUC averaged over all databases and outcomes. *Red* indicates relative risks that are statistically significant different from 1. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

injury pre-exposure is high, thereby masking the potential drug effect. Just as we would anticipate that positive controls should yield large and statistically significant findings, we desire negative controls to produce non-significant findings near the null value of IRR = 1. Sitagliptin is an anti-diabetic medication classified as a negative control due to lack of evidence of any association with acute liver injury; across all 5 databases, the self-controlled cohort design generates non-significant estimates near IRR = 1, thereby likely classifying sitagliptin as a 'true negative'. Another negative control, primidone, is an anticonvulsant that has not previously been associated with acute liver injury. In all 5 database, the estimated association of primidone and acute liver injury was highly significant with IRR >1.5. A plausible explanation for this 'false positive' finding is that primidone is commonly co-prescribed with other anticonvulsant therapies, including carbamazapine and valproate, which are known to be associated with acute liver injury. The self-controlled cohort method does not adjust for confounding by co-therapy, and events observed during concomitant use may be inappropriately attributed to post-exposure time-at-risk, which would elevate the rate ratio.

3.3 Bias

Figure 4 shows the magnitude of bias observed across the estimates for the negative control test cases in the five real databases. We see across all four outcomes and all 5 databases that the self-controlled cohort method is positively biased, that is the expected value for the method when applied to a negative control is greater than 1. The degree of bias and variability around this bias varies considerably by database. Most notably, the most substantial positive bias is observed within the GE database, and is consistent across all four outcomes, where the expected IRR is near 2.0.

3.4 Coverage Probability

Figure 5 shows the coverage probabilities on simulated data. Here, the method had substantially lower coverage probability across all four outcomes, and the degree of coverage decreased as the true effect size increased, with an increasing proportion of true effects falling above the upper bound. In no scenarios did the method achieve a coverage probability >60 %. For acute liver injury, when the true effect size is RR = 1 (that is, no signals injected), the coverage probability = 45 %, with the remaining 55 % of positive controls evenly distributed below and above the estimated intervals. When we injected signals for the acute liver injury positive controls at RR = 2, the coverage probability decreased to 35 %, with half of intervals estimated with upper bounds <2. When the true effect size was increased to RR = 10, the coverage probability was measured at 9 %.

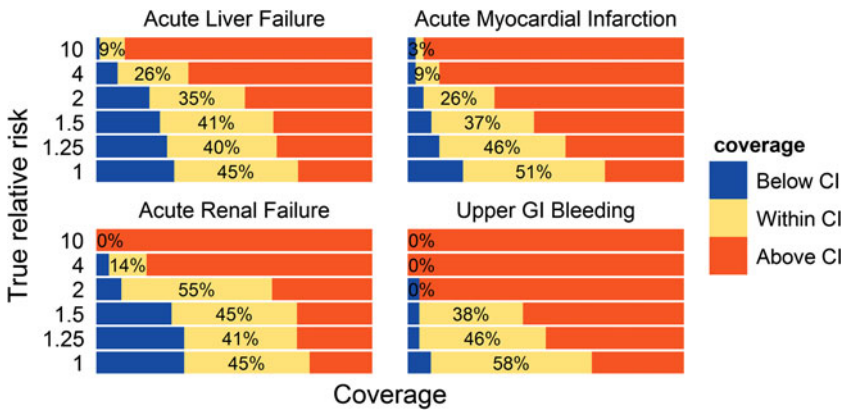


Fig. 5 Coverage probability of self-controlled cohort design at different levels of true effect size, by outcome

Table 2 Parameter sensitivity within the self-controlled cohort design

Parameter	q10 delta	q50 delta	q90 delta
Surveillance window (time-at-risk + control period)	1.04	1.28	2.08
Data source	1.03	1.23	2.19
Exposures to include	1.01	1.09	1.44
Outcomes to include	1.00	1.04	1.25
Include index date in time-at-risk	1.00	1.03	1.71
Include index date in control period	1.00	1.02	1.58

Q10/50/90 delta—10/50/90th percentile on the absolute change in point estimate observed across all outcome/database scenarios by holding all other parameters constant and changing the target parameter to an alternative value

3.5 Analysis Choice Sensitivity

Table 2 shows how sensitive effect estimates were to the various analysis choices. The median change in effect estimates when changing the surveillance windows (the combination of the time-at-risk and control periods) is 28 %. In other words, when holding all other analysis choices constant, there is a 50 % chance that the observed IRR when using one surveillance window will change by ≥28 % either positively or negatively when changing the surveillance window. There is a 10 % chance that the impact of changing the surveillance window will be ≥108 %, or either a doubly or halving of the risk. On expectation, the two most sensitive analysis choices were the choice of surveillance window and the choice of database. However, all analysis choices demonstrated that effect estimates could change by at least 25 % in at least 10 % of situations.

4 Discussion

In this paper, we evaluate the absolute performance of the self-controlled cohort method in terms of predictive accuracy, bias, and coverage probability. The relative performance of the self-controlled cohort method compared with other types of designs, such as new user cohort, case–control and self-controlled case series, is described elsewhere [13]. The self-controlled cohort design borrows key concepts from both the new user cohort approach and the self-controlled case series design to provide a comparison of post-exposure event rate with pre-exposure event rates. The self-controlled cohort design applies a cohort approach for defining the exposed population of interest and defining time-at-risk by looking forward from the index date of first exposure, but it is similar in intent to the self-controlled case series design in that it the same exposed patients' time pre-exposure time as the referent comparator. Whereas the self-controlled case series maximizes a likelihood at an individual level after conditioning on the number of events for each person to estimate the effect of exposure on outcome, the self-controlled cohort method uses population-level rates across the same set of patients to produce its effect estimate. Even without the patient-level conditioning, because the post-exposure and pre-exposure times arise from the same population, the self-controlled cohort design explicitly addresses between-person time-invariant confounding factors, such as gender, age, and genetic attributes. However, it does not address time-varying confounders, such as concomitant treatment or onset of acute illnesses. Further, the design may be limited by unobserved confounders associated with differential healthcare utilization before and after the time of exposure. Moreover, there are scenarios where the self-controlled cohort design would not be appropriate, such as for studying death; since patients would be censored at the time of the event and only patients with exposure are used in the analysis, only events post-exposure could be identified meaning that the pre-exposure time would be subject to immortal time bias [14]. The self-controlled cohort design would be subject to many of the limitations of a self-controlled case series, in that both assume independent Poisson process for occurrence of events, such that the risk of recurrent events are the same as incident events.

In practice, we observed the self-controlled cohort method achieved strong predictive accuracy across the four outcomes and five databases under study. In all 20 outcome-database scenarios, when using IRR as the rank-order statistic, the optimal analysis achieved AUC >0.76. In 16 of those scenarios, the AUC was larger than 0.80, and in 8 scenarios, the AUC exceeded 0.85. When studying acute renal failure, 3 databases achieved AUC >0.90. To put that into context, if the objective of a risk identification system was to achieve 50 % sensitivity—that is, set a threshold such that half of true effects would be identified—then the lowest performing scenario (studying acute liver injury in MDCR, AUC = 0.76) would yield a specificity of 89 % and require a IRR threshold of 1.50. The highest performance scenario (studying acute renal failure in MSLR, AUC = 0.95) would achieve 50 % sensitivity at 95 % specificity using an IRR threshold = 2.80. Alternative thresholds could be set to change the tradeoff between the rate of false positive and false negative findings. Clearly, all stakeholders desire a system that efficiently finds all drug safety issues without raising any false alarms that can require significant resources to mitigate, but such an ideal system would require perfect predictive accuracy.

Instead, we should likely anticipate that a system can provide highly informative, but not definitive, evidence about the effects of products, and decision thresholds for how to act on that evidence will be based on stakeholder preference for the relative compromise between types of errors.

The review of specific effect estimates demonstrate both the promise and challenges faced with using the self-controlled cohort method. While the method successfully discriminated between the acute liver injury positive control of isoniazid and the negative control of sitagliptin, most decision thresholds would likely falsely identify primidone and fail to find erythromycin. Similar examples can be found for the other outcomes. Some of the misclassification can be explained post-hoc by thinking about the clinical situation and positing how the design may fail to address underlying bias, but hypothesizing explanations is not sufficient unless a solution to overcome the misclassification can be implemented and evaluated to demonstrate improved predictive accuracy. Further research in methods enhancement can use the current performance as a benchmark to evaluate how much progress is being made.

While the self-controlled cohort method demonstrated strong predictive accuracy across all 4 outcomes and 5 databases, the actual estimates generated were observed to be biased and require calibration to be properly interpreted under nominal properties [15]. The error distributions generated from the negative controls for each outcome highlight that the self-controlled cohort method is, on expectation, positively biased. While we expect the null distribution for an effect estimator to be centered around IRR = 1, we observed empirically the distribution to be centered closer to IRR = 1.5. As a result, the raw estimate that comes from the method cannot be interpreted directly as the magnitude of increased risk; IRR = 1.5 should not be read as demonstrating a 50 % increased risk, since that 50 % may in fact be attributable to bias. Due to this positive bias, much larger estimates are required to yield a posterior probability that demonstrates strong confidence that there either is or is not an effect. We also observed that the confidence intervals computed within the self-controlled cohort method are overly narrow and increasingly under represent true effects as the effect size increases. These findings suggest that the standard error estimates from the design, which focuses on sampling variability from the Poisson rates, do not account for all potential sources of variability (such as bias) that give rise to increased uncertainty around the point estimate. The residual error in estimating the variance, in conjunction with the positive bias observed in the point estimate, results in 95 % confidence intervals having markedly less coverage than 95 %. Further work is needed to calibrate the self-controlled cohort estimates, shifting the point estimate and increasing the variance, to regain the nominal properties expected from the confidence interval. These data suggest that learning from the self-controlled cohort method requires interpreting the estimates in the context of what has been observed from prior positive and negative controls, and cannot be based on conventional interpretation of relative risk, confidence intervals, and p values as if they represent an unbiased estimator.

These intuitively inconsistent results—that the self-controlled cohort method can have strong predictive accuracy but poor calibration of estimates—underscores a basic challenge in methodological research and highlights the need for evaluating multiple metrics in any methods assessment. We have measured discrimination using AUC, with is a rank-order statistic that does not use the magnitude of the effect estimate in its calculation. In an ideal situation, a method would have perfect predictive accuracy and no error. In practice, it is possible for a method to be severely biased but have perfect discrimination; imagine the situation where a method produced the correct estimate for all drug-outcome pairs but then added a constant value to all relative risks, in this case, the AUC = 1 but the method would be biased by the constant. In contrast, if a method has weak predictive accuracy (e.g. AUC = 0.50), it is not possible for the method to have a small amount of error, because small error among positive and negative controls would dictate that the rank-ordering of estimates would have some degree of discrimination. Neither predictive accuracy nor error are sufficient metrics for evaluation, but instead each provide complementary perspectives regarding a method's potential utility and both yield information required to put study results into proper context. The self-controlled cohort method results suggest potential utility for the task of discrimination but concern for its use in effect estimation unless some calibration is first applied.

This study evaluated the performance of the self-controlled cohort method on four outcomes that have been highlighted to be important events for monitoring in a risk identification system [16]. The study is limited by the integrity of classification of the positive and negative controls, and the applicability of the drugs used in each outcome to represent the distribution of expected scenarios for those outcomes. While the method showed robust predictive accuracy across these four outcomes, we also observed that optimal parameterization and level of performance could differ by outcome and database. As a result, we caution against generalizing these results to other outcomes or other data sources. For example, studying an

outcome that results in a large proportion of fatalities is likely to fail subject to the immortal time bias described earlier, and therefore would not be expected to perform at the levels observed for these four outcomes. Instead, we encourage applying a similar empirical evaluation prior to applying methods to new outcomes, whereby the method is benchmarked against a set of known positive and negative controls to gauge the expected predictive accuracy and estimate bias and coverage probability.

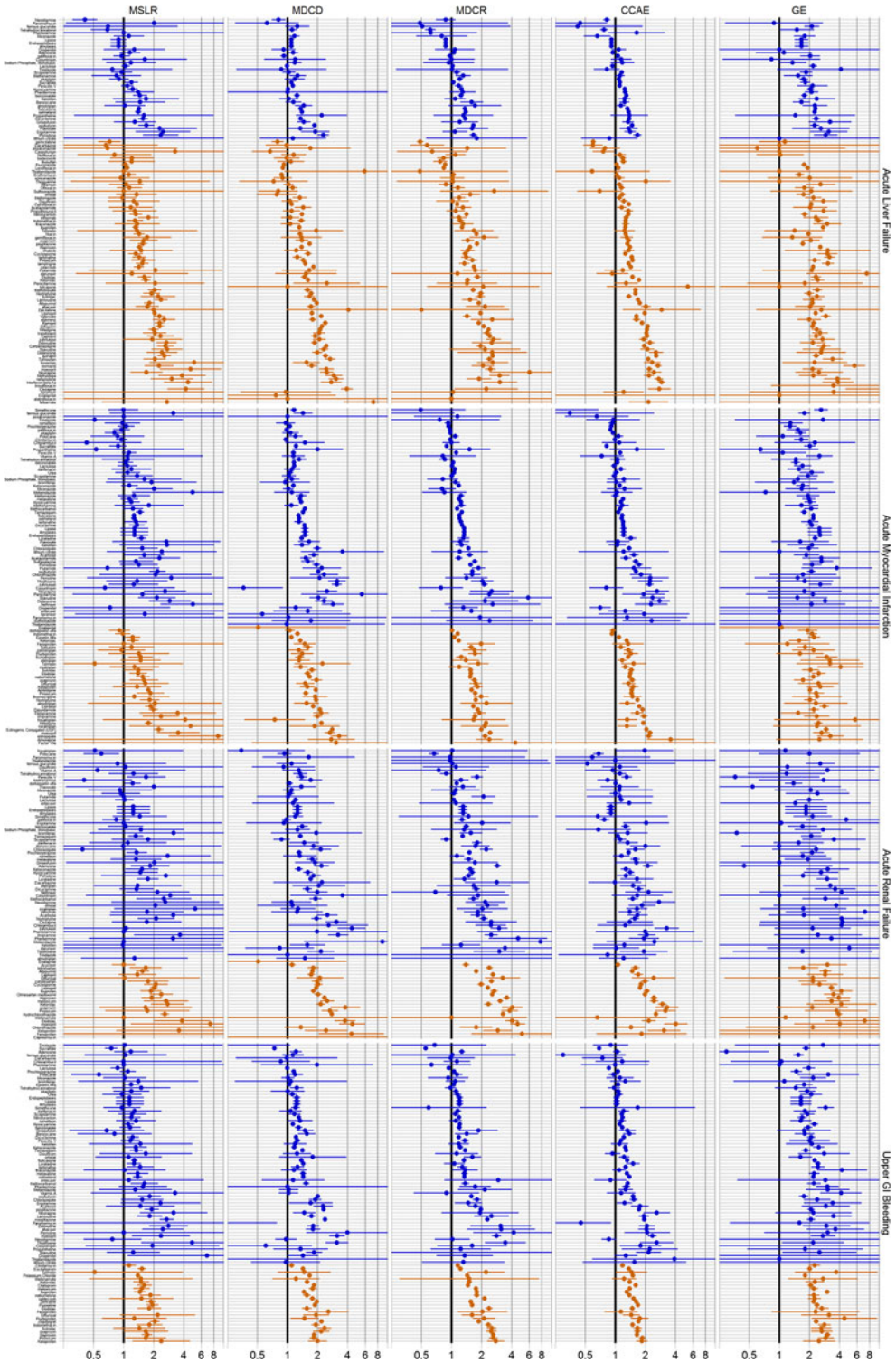
5 Conclusions

The self-controlled cohort design is one of several methods under consideration for use in a risk identification system, but the empirical evidence around its performance suggests that results need to carefully situated for its intended purpose. If the objective for a risk identification system is simply to discriminate between positive effects and negative effects as a means of prioritization, then the self-controlled cohort method shows promise as a potential tool with consistently strong measures of predictive accuracy observed across the four outcomes and five databases studied. However, if a system is intended to generate effect estimates which are to be used quantify the magnitude of potential risks, then the self-controlled cohort method may not be suitable, and requires substantial calibration to be properly interpreted under nominal properties.

Acknowledgments The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Biogen Idec, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, and Takeda. Drs. Ryan and Schuemie are employees of Janssen Research and Development. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration. Drs. Schuemie and Madigan have received funding from FNIH.

This article was published in a supplement sponsored by the Foundation for the National Institutes of Health (FNIH). The supplement was guest edited by Stephen J.W. Evans. It was peer reviewed by Olaf H. Klungel who received a small honorarium to cover out-of-pocket expenses. S.J.W.E has received travel funding from the FNIH to travel to the OMOP symposium and received a fee from FNIH for the review of a protocol for OMOP. O.H.K has received funding for the IMI-PROTECT project from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement no 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

Appendix



Appendix Self-controlled cohort design estimates for all test cases, by database. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity. *Blue* negative controls, *orange* positive controls, *each line* represents point estimate and 95 % confidence interval for the drug–outcome pair in a particular database

References

1. Public Law 110-85: Food and Drug Administration Amendments Act of 2007. 2007.
2. Woodcock J, Behrman RE, Dal Pan GJ. Role of postmarketing surveillance in contemporary medicine. Annu Rev Med. 2011;62:1–10.
3. Poh A. SafetyWorks at GlaxoSmithKline, Best Practice Winner: Translational and Personalized Medicine. Bio-IT World Mag. 2009.
4. Ryan PB, Powell GE, Pattishall EN, Beach KJ. Performance of screening multiple observational databases for active drug safety surveillance. Providence: International Society of Pharmacoepidemiology; 2009.
5. Graham PL, Mengersen K, Morton AP. Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. Stat Med. 2003;22(12):2071–83.
6. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf. (2013) (In this supplement issue). doi:10.1007/s40264-013-0110-2.
7. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf. (2013) (In this supplement issue). doi:10.1007/s40264-013-0102-2.
8. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf. (2013) (In this supplement issue). doi:10.1007/s40264-013-0097-8.
9. Armstrong B. A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. Am J Epidemiol. 1987;126(2):356–8.
10. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Dec Making Int J Soc Med Decis Making. 2000;20(4):468–70.
11. Smith BM, Schwartzman K, Bartlett G, Menzies D. Adverse events associated with treatment of latent tuberculosis in the general population. CMAJ. 2011;183(3):E173–9.
12. Carson JL, Strom BL, Duff A, Gupta A, Shaw M, Lundin FE, et al. Acute liver disease associated with erythromycins, sulfonamides, and tetracyclines. Ann Intern Med. 1993;119(1):576–83.
13. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf. (2013) (In this supplement issue). doi:10.1007/s40264-013-0108-9.
14. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167(4):492–9.
15. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2013. doi:10.1002/sim.5925.
16. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf. 2009;18(12):1176–84.