Drug Saf (2013) 36 (Suppl 1):S83–S93
DOI 10.1007/s40264-013-0100-4

ORIGINAL RESEARCH ARTICLE

Empirical Performance of the Self-Controlled Case Series Design: Lessons for Developing a Risk Identification and Analysis System

Marc A. Suchard · Ivan Zorych · Shawn E. Simpson · Martijn J. Schuemie · Patrick B. Ryan · David Madigan

Abstract

Background The self-controlled case series (SCCS) offers potential as an statistical method for risk identification involving medical products from large-scale observational healthcare data. However, analytic design choices remain in encoding the longitudinal health records into the SCCS framework and its risk identification performance across real-world databases is unknown.

Objectives To evaluate the performance of SCCS and its design choices as a tool for risk identification in observational healthcare data.

Research Design We examined the risk identification performance of SCCS across five design choices using 399 drug-health outcome pairs in five real observational databases (four administrative claims and one electronic health records). In these databases, the pairs involve 165 positive controls and 234 negative controls. We also consider several synthetic databases with known relative risks between drug-outcome pairs.

Measures We evaluate risk identification performance through estimating the area under the receiver-operator characteristics curve (AUC) and bias and coverage probability in the synthetic examples.

Results The SCCS achieves strong predictive performance. Twelve of the twenty health outcome-database scenarios return AUCs >0.75 across all drugs. Including all adverse events instead of just the first per patient and applying a multivariate adjustment for concomitant drug use are the most important design choices. However, the SCCS as applied here returns relative risk point-estimates

This work was partially supported by National Science Foundation grant IIS 1251151. 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.

M. A. Suchard (✉)
Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California,
Los Angeles, CA, USA
e-mail: msuchard@ucla.edu

M. A. Suchard
Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California,
Los Angeles, CA, USA

M. A. Suchard
Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA

I. Zorych · S. E. Simpson · D. Madigan
Department of Statistics, Columbia University,
New York, NY, USA

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

P. B. Ryan
Janssen Research and Development, Titusville, NJ, USA

biased towards the null value of 1 with low coverage probability.

Conclusions The SCCS recently extended to apply a multivariate adjustment for concomitant drug use offers promise as a statistical tool for risk identification in large-scale observational healthcare databases. Poor estimator calibration dampens enthusiasm, but on-going work should correct this short-coming.

1 Introduction

In 2007, Congress enacted the Food and Drug Administration (FDA) Amendment Act, calling for the establishment of an “active postmarket risk identification and analysis system” utilizing patient-level observational data from 100 million lives by 2012 [1]. This system should “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]. One-off pharmacoepidemiology studies examining specific hypotheses about the effects of particular medical product exposure and subsequent health outcomes of interest have previously used observational healthcare data, including administrative claims and electronic health records. However, advancing health informatics raises the potential to exploit these same data to provide a systematic process for monitoring all regulated products for a wide range of health outcomes of interest. Before a risk identification and analysis system can be properly embedded within safety decision-making processes, several outstanding questions require resolution, including ‘which analytical methods to apply?’, ‘which data should be used?’, and ‘how credible is the evidence from observational analyses?’.

Researchers have proposed several methods for use in a risk identification system [3], but little empirical research exists to inform the expected operating characteristics of these approaches. One such method is the self-controlled case series (SCCS) [4]. The SCCS is a model for recurrent adverse events (AEs) that derives from a conditional Poisson regression. The method is “self-controlled” in that it does a within-person comparison of the event rate during exposure to the subject-specific baseline event rate while unexposed [5]. The conditioning produces two beneficial properties. First, it controls for fixed baseline covariates; this aspect is attractive when the data do not measure baseline covariates with sufficient precision to effectively adjust for confounding. Second, the SCCS needs the inclusion of only exposed and affected cases, greatly reducing the computational demands in the massive observational database setting. Grosso et al. [6] and Douglas and Smeeth [7] highlight the growing interest in the application of SCCS to postmarketing surveillance for prescription drug safety.

In this study, we evaluate the performance of a SCCS as a potential analytical method for a risk identification system. We test the SCCS in five real observational healthcare databases and 6 simulated data sets, retrospectively studying the predictive accuracy of the method when applied to a collection of 165 positive controls and 234 negative controls across four 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, stake-holders can evaluate the SCCS 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 the Self-Controlled Case Series

SCCS models AEs as arising from a non-homogeneous Poisson process, where drug exposure modulates the time-varying event rate. Each patient $i = 1, \ldots, N$ carries an unknown individual baseline event rate of e^{ϕ_i} and periods of exposure to drug $j = 1, \ldots, J$ measured each day result in a multiplicative effect of e^{β_j} to this baseline rate. Hence, e^{β_j} quantifies the relative rate of an AE during exposure to drug j. Put together, we observe patient i for a total of τ_i days in the database and the event rate for patient i on day d becomes $\lambda_{id} = e^{\phi_i + \mathbf{x}_{id}^t \boldsymbol{\beta}}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^t$, $\mathbf{x}_{id} = (x_{id1}, \ldots, x_{idJ})^t$, and drug indicator $x_{idj} = 1$ if patient i is exposed to drug j on day d and 0 otherwise.

Let y_{id} count the number of AEs that patient i experiences on day d. Then, by conditioning on the total number $n_i = \sum_{d=1}^{\tau_i} y_{id}$ of AEs experienced by each patient, nuisance quantities ϕ_i fall out of the SCCS likelihood, leaving a log-likelihood of

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} \left[\sum_{d=1}^{\tau_i} y_{id} \mathrm{x}_{id}^t \beta - n_i \log\left(\sum_{\mathrm{d}=1}^{\tau_i} e^{\mathrm{x}_{id}^t \beta} \right) \right]. \tag{1}$$

Examining Eq. 1, only patients with at least one AE and at least one drug exposure affect the curvature of the model likelihood and need inclusion the analysis. Given this formulation, the SCCS assumes that the exposure distribution is independent of the event times; hence bias in estimating $\boldsymbol{\beta}$ arises if an outcome influences future exposures or if an outcome leads to censoring. In practice, such bias may be sizable [8]. A perception lingers that the SCCS only applies for certain types of drug-outcome pairs (e.g., intermittent exposures and transient events); little theory exists to support this assertion and Taylor et al. [9] provide a strong counter-example in exploiting SCCS to rule out an association between the MMR vaccine and autism. For notational convenience, we group all outcomes $\mathbf{Y} = (y_{11}, \ldots, y_{N\tau_N})^t$ and all covariates $\mathbf{X} = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{N\tau_N})^t$.

In the typical SCCS application [4], the number of covariates under consideration $J = 1$ and Eq. 1 entertains a single unknown, β_1, to estimate for the target drug of direct interest, producing a marginal measure of association. However, most patients in longitudinal healthcare databases often take multiple drugs throughout the course of their observation. These exposures are potential confounders [10]. To avoid the spurious effects of uncontrolled confounding, such as the "innocent bystander" effect [11], we exploit recent developments in a multiple SCCS [12]. Under this approach, we can include all possible drug exposures that the patients experience. This leads often to $J = 1000$s and inference via Eq. 1 becomes a high-dimensional regression problem. We avoid the over-fitting and numerical instability of direct maximum likelihood estimation [13] by turning to a regularized regression framework, since we believe *a priori* that most drugs should have no effect on the relative rate of AEs, i.e., $\beta_j = 0$ or is close to 0 for almost all j. We consider a ridge-regression [13] inspired regularization based on the L_2 norm of $\boldsymbol{\beta}$ penalty, $f(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{J} \beta_j^2$, and λ is a shrinkage coefficient. Following Suchard et al. [14], we estimate $\boldsymbol{\beta}$ using a modified cyclic coordinate ascent algorithm to maximize $\mathcal{L}(\boldsymbol{\beta}) - f(\boldsymbol{\beta})$, determine λ through 10-fold cross-validation and employ a non-parametric bootstrap via re-sampling patients to approximate standard errors and 95 % confidence intervals on our point estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Several important analytical design choices still remain to enable a fully specific SCCS analysis for a risk identification system. In general, these choices revolve around the data encoding step in moving from a longitudinal healthcare database to the outcome and covariate variables $(\mathbf{Y}, \mathbf{X})$. Within the open-source software implementation of the design titled 'Self-Controlled Case Series', publicly available at http://omop.org/MethodsLibrary, we parameterize five analytical design choices. Figure 1 grants a cartoon overview of many of these choices, and we describe each choice below:

1. Should the analysis consider all occurrences or only the first occurrence of an adverse event as potential outcomes?
2. Should the analysis adjust for concurrent exposures that vary over the observation period to other drugs (yes/no) through the multiple SCCS extension?
3. What time-at-risk window does the analysis use? The at-risk window defines a time period relative the dates of drug exposure during which the analysis allows for a possible multiplicative drug effect. For example, –30 captures effects that happen within 30 days of initiation, 1 captures effects only on the day of initiation, +30 captures effects anytime within 30 days following the end of exposure and ∞ captures all effects any time following exposure.
4. Should the analysis include drug effects on the same day as exposure (yes/no)?
5. What is the minimum observation length of a subject for inclusion in the analysis (none, 180 days)?

In this study, we evaluate the 64 unique design choice combinations, where each combination represents an analysis.

2.2 Experiment Design

We conduct our study against five observational healthcare databases to allow evaluation of performance across different populations and data capture processes: MarketScan Lab Supplemental (MSLR, 1.2m persons), MarketScan®

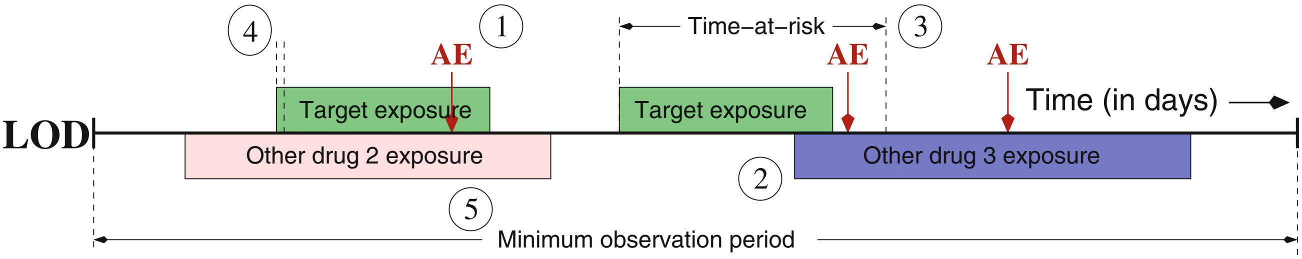


Fig. 1 Design choices within the self-controlled case series experiment. We display the longitudinal observational healthcare database (LOD) record for a single patient who is taking the target drug of interest and two other drugs concomitant and experiences three adverse events. The design choices consider: (1) including the first or all occurrences; (2) adjusting for multiple drugs; (3) the time-at-risk for each exposure; (4) including events that occur on the first day of each exposure; and (5) have a minimum duration of observation in the database

Medicare Supplemental Beneficiaries (MDCR, 4.6m persons), MarketScan® Multi-State Medicaid (MDCD, 10.8m persons), Truven MarketScan® Commercial Claims and Encounters (CCAE, 46.5m persons), and the GE Centricity (GE, 11.2m persons) database. GE is an electronic health record (EHR) database; the other four databases contain administrative claims data. We also employ a 10m-person simulated data set constructed using the OSIM2 simulator [15], modeled after the MSLR database, and replicated 6 times to allow for injection of signals of known size, specification with relative risks of 1, 1.25, 1.5, 2, 4 and 10. Ryan et al. [16] describe these data in more detail.

In the experiment, we execute the SCCS method using all design choice combinations against 399 drug-outcome pairs to generate an effect estimate and standard error for each pair and choice combination. 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, limiting ourselves to four health outcomes of interest: acute liver injury, acute myocardial infarction, acute renal failure, and upper gastrointestinal bleeding. Ryan et al. [17] describe the full set of test cases, their construction and the difficulties involved in considering these cases as gold-standards. For each database, we restrict our analysis to those drug-outcome pairs with sufficient power to detect a relative risk $\geq$1.25, based on the age-by-gender-stratified drug and outcome prevalence estimates [18].

2.3 Metrics

SCCS relative rate estimates $\hat{\beta}_j$ for the target drug 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 our method to distinguish between positive and negative controls, we used these estimates to construct receiver operator characteristics (ROC) and compute the area under the ROC curve (AUC), a measure of predictive accuracy for binary classifiers [19]. 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. To construct these ROC curves, we convert the estimates into a binary classifier by considering that the SCCS labels as + a target drug if the method returns $\hat{\beta}_j > c$, where c is an arbitrary cut-off that characterizes the classifier; otherwise, the method assigns the drug as $-$. Each point on the ROC curve corresponds to a unique choice for the threshold c and determines the shape of the curve.

Often we are not only interested in whether there is a significant effect or not, but would also like to know the magnitude of the effect. However, in order to evaluate whether a method produces accurate relative rate 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 rate is 1. Fortunately, in the simulated data sets we do know the true relative rate 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 at the nominal 5 % Type I Error rate, we would expect the coverage probability to be 95 %. Lastly, we are interested in to what extent each design choice can influence the estimated relative rate. For every design choice, we evaluate how much the estimated relative rates changes as a consequence of changing a single choice while keeping all other choices constant.

3 Results

3.1 Predictive Accuracy of all Settings

Figure 2 highlights the predictive accuracy, as measured by AUC, of all SCCS design choices across the four outcomes and five databases. Performance varies considerably across different health outcomes, with a maximal AUC of only 0.70 for acute liver failure, but well above 0.80 for the three remaining outcomes. For each outcome-database scenario, we identify the design choice settings yielding the highest AUC, as listed in Table 1. For example, an optimal setting (SCCS : 1907010) has the highest predictive accuracy for discriminating test cases for acute kidney injury in MSLR (AUC = 1.00) and in MDCD (AUC = 0.74). Another optimal setting (SCCS : 1949010) has the highest performance for acute kidney injury in GE (AUC = 0.94), acute liver injury in GE (AUC = 0.70) and acute myocardial infarction (MI) in MSLR (AUC = 0.61). Finally, SCCS : 1939010 has the highest performance for gastrointestinal (GI) bleed in GE (AUC = 0.88), acute kidney injury in MDCR (AUC = 0.80) and acute (MI) in MDCD (AUC = 0.67). Across all 20 scenarios, including all occurrences as a design choice produces the highest AUC. In 75 % of the scenarios performing a multivariate adjustment leads to the optimal settings. We reflect this finding in Fig. 2 by shading points in red if they provide a multivariate adjustment. Further, all but 6 optimal settings encode all time post-exposure. However, in spite of these similarities, important variability between settings exists. For example, we return one last time to Fig. 2. In this figure, the dashed lines indicate the performance of the top-

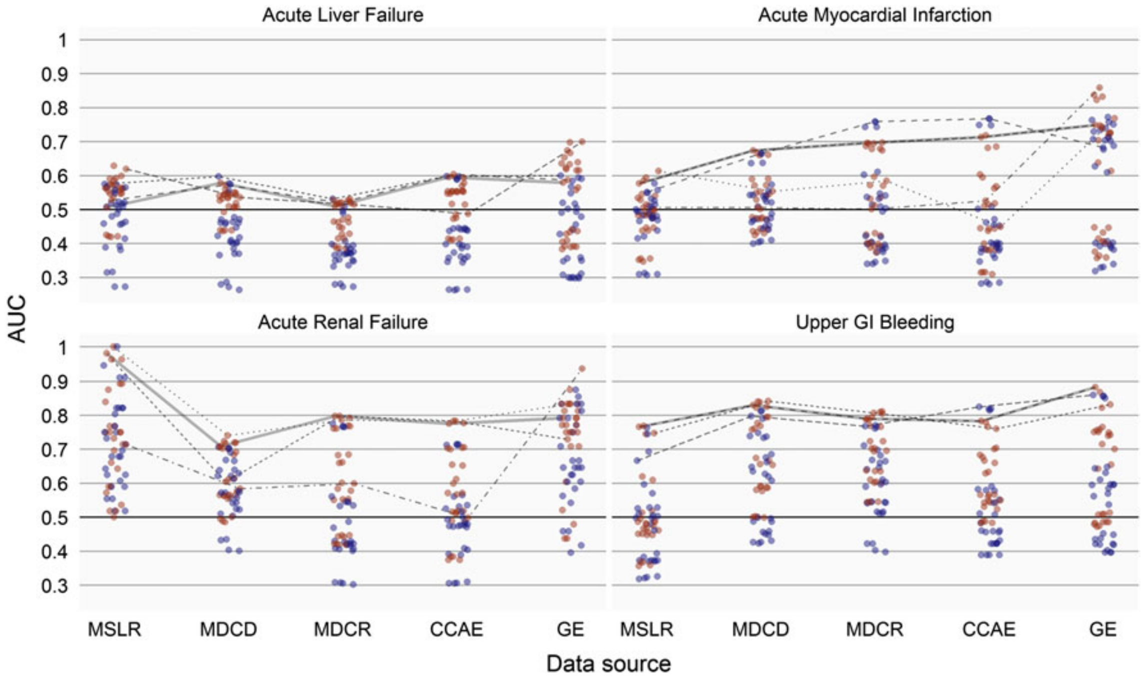


Fig. 2 Area under the receiver operator characteristic curve (AUC) for self-controlled case series (SCCS) design choices stratified by health outcome of interest and data source. Each *dot* represents the 64 unique choice combinations of the SCCS design. The *solid gray line* highlights the choice with the highest AUC across all 20 outcome-database scenarios. The *dashed lines* identify each setting with highest AUC for each database within each outcome. *Red points* identify analyses that provide a multivariate adjustment for concomitant drugs. We define the database acronyms in the main text

performing setting across outcomes within the same database, showing that the optimal setting for one outcome can sometimes perform poorly when used for another outcome in the same database.

3.2 Overall Optimal Settings

The design choice settings with the best average performance across the 20 outcome-database scenarios is SCCS : 1939010. These settings consider all occurrences, provide a multivariate adjustment, examine all time post-exposure, include the index date and place no minimum on the observation period. Across all scenarios, SCCS : 1939010 achieves an average AUC = 0.71. In the remainder of this paper, we use SCCS : 1939010 as the representative settings for the SCCS analyses.

Supplementary Figure 1 plots effect estimates for all test cases across the five databases using these optimal choice settings. To illustrate patterns in these findings, we discuss four specific test cases for acute liver injury in CCAE, as shown in Fig. 3. Used in the treatment of tuberculosis, isoniazid is a causative agent of acute liver injury [20] and a positive control in our experiment. The association between isoniazid and acute liver injury is consistently one of the largest effects we observe using SCCS, with all five data sources returning relative rates >1 and statistically significant at the conventional $p < 0.05$ Type I error rate. Isoniazid illustrates the opportunity for use of this design in a risk identification system and would likely be classified as a 'true positive' under most decision thresholds based either on effect size or statistical significance. In contrast, lamivudine is a nucleoside reverse transcriptase inhibitor employed in the treatment of chronic hepatitis B virus and human immunodeficiency virus infections. Case reports associate lamivudine with acute liver injury [21, 22], and hence we use lamivudine as a positive control. However, in all data sources except MSLR, we find the relative rate estimates for lamivudine-acute liver injury <1 with statistical significance $p < 0.05$. That is, the rate of acute liver injury while taking lamivudine is less than the rate the injury off the drug. In our construction of positive and negative controls [17], this provides an example of a 'false negative' finding. However, when taken to treat severe hepatitis B, lamivudine may improve outcome [23], preventing liver failure coded as acute liver injury.

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 hypothesis relative rate of 1. Salmeterol is a long-acting β_2-adrenergic receptor agonist prescribed to treat asthma and chronic obstructive pulmonary disease that we classify as a negative control due to lack of evidence of any

Table 1 Optimal design choices for the self-controlled case series (SCCS) method stratified by health outcome of interest and data source. For each outcome-database scenario, we report the area under the receiver operator characteristic curve (AUC), the unique OMOP scenario identifier and choice settings: (1) Outcomes to include; (2) Multivariate adjustment; (3) At-risk window; (4) Include index date in window; and (5) Minimum observation length. We define the database acronyms in the main text

Data sources	Acute liver injury	Acute kidney injury	Acute myocardial infarction	Gastrointestinal bleed
CCAE	AUC = 0.60 (1955010)	AUC = 0.78 (1955010)	AUC = 0.77 (1947010)	AUC = 0.82 (1931010)
	1: All occurrences	1: All occurrences	1: All occurrences	1: All occurrences
	2: Yes	2: Yes	2: No	2: No
	3: All time post-exposure	3: All time post-exposure	3: All time post-exposure	3: All time post-exposure
	4: Yes	4: Yes	4: Yes	4: No
	5: 180 days	5: 180 days	5: None	5: 180 days
MDCD	AUC = 0.60 (1931010)	AUC = 0.74 (1907010)	AUC = 0.67 (1939010)	AUC = 0.84 (1923010)
	1: All occurrences	1: All occurrences	1: All occurrences	1: All occurrences
	2: No	2: Yes	2: Yes	2: Yes
	3: All time post-exposure	3: All time post-exposure	3: All time post-exposure	3: All time post-exposure
	4: No	4: No	4: Yes	4: No
	5: 180 days	5: None	5: None	5: 180 days
MDCR	AUC = 0.53 (1931010)	AUC = 0.80 (1939010)	AUC = 0.76 (1947010)	AUC = 0.81 (1921010)
	1: All occurrences	1: All occurrences	1: All occurrences	1: All occurrences
	2: No	2: Yes	2: No	2: Yes
	3: All time post-exposure	3: All time post-exposure	3: All time post-exposure	3: Exposure + 30 days
	4: No	4: Yes	4: Yes	4: No
	5: 180 days	5: None	5: None	5: 180 days
MSLR	AUC = 0.63 (1933010)	AUC = 1.00 (1907010)	AUC = 0.61 (1949010)	AUC = 0.77 (1955010)
	1: All occurrences	1: All occurrences	1: All occurrences	1: All occurrences
	2: Yes	2: Yes	2: Yes	2: Yes
	3: Exposure + 30 days	3: All time post-exposure	3: Exposure + 30 days	3: All time post-exposure
	4: Yes	4: No	4: Yes	4: Yes
	5: None	5: None	5: 180 days	5: 180 days
GE	AUC = 0.70 (1949010)	AUC = 0.94 (1949010)	AUC = 0.86 (1951010)	AUC = 0.88 (1939010)
	1: All occurrences	1: All occurrences	1: All occurrences	1: All occurrences
	2: Yes	2: Yes	2: Yes	2: Yes
	3: Exposure + 30 days	3: Exposure + 30 days	3: Exposure only	3: All time post-exposure
	4: Yes	4: Yes	4: Yes	4: Yes
	5: 180 days	5: 180 days	5: 180 days	5: None

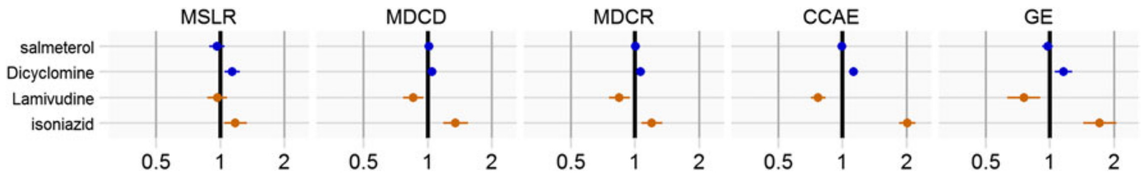


Fig. 3 Estimated relative rates (points) and their 95 % confidence intervals (*whiskers*) for four example drugs and acute liver injury stratified across data sources using the overall optimal settings. *Blue* represents negative controls and orange identifies positive controls

association with acute liver injury. Across all five data sources, SCCS generates non-significant relative rate estimates near 1, thereby likely classifying salmeterol as a ‘true negative’ in a risk identification system. Another negative control, dicyclomine, is an anticholinergic that relieves muscle spasms in the gastrointestinal tract, used in the treatment of intestinal hypermotility and the symptoms of irritable bowel syndrome. Not previously associated with acute liver injury, the SCCS method returns estimated relative rates modestly, but statistically significantly, >1 in the in MSLR, CCAE and GE databases. A plausible explanation for this ‘false positive’ finding centers around the non-specific gastrointestinal complaints for which patients try dicyclomine; these symptoms may reflect disease processes associated directly with liver damage.

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. Across all health outcomes, with the possible exception of acute renal failure, and all five databases, we observe that the SCCS is approximately unbiased. The majority of estimates fall near 1; this means that the expected value for the analysis when applied to a negative control is approximately 1. However, variability of point estimates for the negative controls varies considerably by health outcome. Such variability for acute renal failure is notably larger than for the remaining three outcomes. One simple explanation entertains that the scale of variability trends with outcome incidence; for acute MI is the most common and acute renal failure the least common. However, estimates for negative controls of acute renal failure also demonstrate more positive bias than for the other outcomes, suggesting additional systematic error.

3.4 Coverage Probability

Figure 5 presents estimator coverage probabilities that we estimate through simulated data. Here, the SCCS returns substantially lower than nominal (95 %) coverage probability across all four outcomes, and the degree of coverage decreases as the true effect size increases, with an increasing proportion of true effects falling above the upper bound. In no scenarios does the method achieve a coverage probability >65 %. Coverage is best overall for acute renal failure. When the true effect size is 1, i.e., no injected signal, the coverage probability = 60 %, with the remaining 40 % of positive controls landing unequally below and above the estimated intervals. When we inject signal for acute renal failure at a relative risk of 1.25, coverage falls to 39 %, with all the remaining true values falling above the confidence interval. This tendency continues for stronger signals in all outcomes.

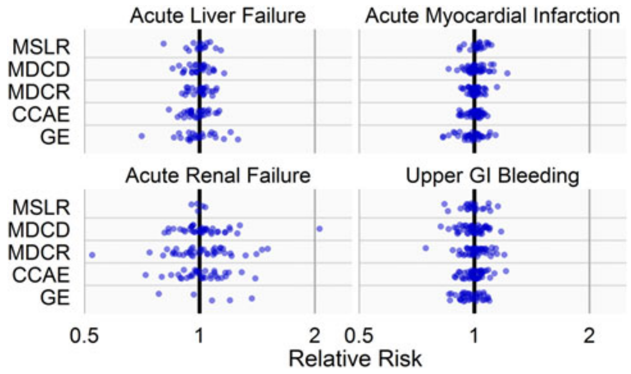


Fig. 4 Bias distributions of relative rate estimates stratified by outcome and data source. *Scatter plot* presents empirically derived point estimates across all negative controls for each outcome and each data source. Bias measures the difference between the point-estimate and the assumed true effect size of 1

3.5 Design Choice Sensitivity

Table 2 shows how sensitive effect estimates are to the design choices. In the table, we list the five design choices in decreasing rank order according to their sensitivity and then append measures of sensitivity to data source for comparison. Effect estimates are most sensitive to the inclusion of all event occurrences or just the first. The median change in effect estimates here is 18 %. In other words, when holding all other design choices constant, there is a 50 % chance that the estimated relative rate will change by ≥18 % either positively or negatively when changing from all occurrences to just the first occurrence. There is a 10 % chance that this impact grows to ≥89 %. Ranked immediately below outcome inclusion is multivariate adjustment, followed then by data source. There remains only a 10 % change that changing the minimum observation length leads to ≥3 % difference in effect estimate.

4 Conclusions

In this paper, we describe and evaluate the SCCS as a new analytical approach to longitudinal observational healthcare data that can be potentially used within a risk identification system. The SCCS compares adverse event rates during times when a person is exposed to a drug of interest to event rates during times when the same person is unexposed, such that each individual acts as their own control. Some key advantages of the SCCS approach include its ability to adjust for all time-invariant

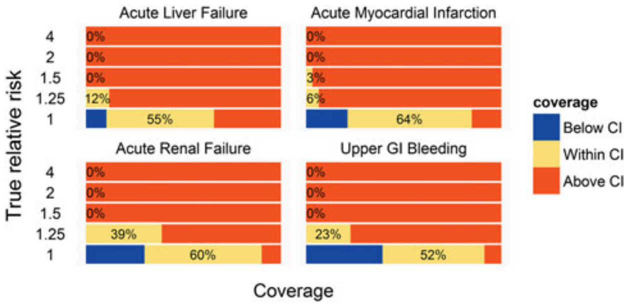


Fig. 5 Estimator coverage probability of the SCCS design at different levels of true effect stratified by health outcome of interest. CI: 95 % confidence interval

Table 2 Design choice sensitivity with the self-controlled case series method. We report percentiles of the absolute change in point-estimates generated across all outcome-database scenarios by holding all other choices constant and changing the target choice to an alternative value

	Percentile		
Choice	10th	50th	90th
Outcomes to include	1.01	1.18	1.89
Multivariate adjustment	1.01	1.13	1.64
At-risk window	1.01	1.07	1.41
Including index date	1.00	1.01	1.09
Minimum observation length	1.00	1.01	1.03
Data source	1.02	1.12	1.57

multiplicative confounders, such as gender, baseline health status and genetic attributes. Estimation under the SCCS also requires a cases-only construction providing massive computational savings in observational healthcare data. Finally, the introduction of regularization into SCCS implementations equips the approach to adjust for large numbers of time-varying covariates, such as concomitant drug use. However, implementing such a method within a risk identification system is not without difficulty.

Notably, real-life longitudinal observational healthcare data are noisy and carry the potential to introduce myriad artifacts and biases into analyses. One obvious example considers a common situation in which one records in the data-base an adverse health condition and the drugs employed to treat this condition simultaneously during a single physician encounter, even though the condition predates the visit. This suggests that the drug used to treat the condition appears to have caused the condition. Design choices in encoding a database for an SCCS analysis help to avoid these biases, but empirical evaluation of how these choices affect performance in a risk identification system has, up to now, remained lacking.

Empirically, we observe that the SCCS achieves strong predictive accuracy across the outcomes and data sources under study. Across all 20 outcome-database scenarios, the optimal designs yield AUC >0.50 when using the estimated relative rate as the rank-order statistic. More impressively, in 12 of those scenarios, AUC >0.75, and in 4 scenarios, AUC >0.85. When studying acute kidney failure, two databases return 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 median performing scenario (studying gastrointestinal bleed in MSLR, AUC = 0.77) would yield a specificity of 83 % and require an estimated relative risk of 1.07. The second highest performance scenario (studying acute kidney failure in GE, AUC = 0.94) would achieve 50 % sensitivity at 100 % specificity using an relative risk of 1.27. Alternative thresholds could be set to change the trade-off 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.

A review of specific SCCS effect estimates demonstrates both the promise and challenges. While the design successfully discriminated between the acute liver injury positive control of isoniazid and the negative control of salmeterol, most decision thresholds would likely falsely identify dicyclomine and fail to find a relative rate >1 for lamivudine. 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 SCCS demonstrates strong predictive accuracy across all four outcomes and five data sources, the actual estimates the method generates require substantial calibration to be properly interpreted under nominal properties. The bias distributions generated from the negative controls for each outcome highlight that the SCCS is, on average, only modestly biased. However, such is no longer the case under the alternative as seen in the simulation studies. Here, the SCCS design returns point-estimates strongly biased towards the null value of 1. While some shrinkage often carries advantages [24], we suspect here that cross-validation is furnishing regularization tuning constants that enforce significant shrinkage. However, preliminary studies within the MSLR and MDCR databases return optimal settings that favor cross-validation over fixing the tuning constant across all outcomes. These studies also support an L_2 over L_1 regularization. Further, [12] find the multiple SCCS returns effect estimates that are on average smaller than without multivariate adjustment. In terms of design effect (see, e.g., Table 2), the importance of selecting cross-validation and the regularization norm falls considerably lower than outcomes to include, multivariate adjustment, time-at-risk definition and data source; hence we do not explore these choices further against all data sources.

To compound bias under the alternative hypothesis, we observe that the confidence intervals computed within our SCCS implementation are overly narrow and increasingly under-represent true effects as the effect size increases. This finding suggests that our SCCS design fails to capture all of the sampling variability in the data sources. We construct our confidence intervals using a non-parametric bootstrapping procedure that re-samples with replacement individual patients. Unfortunately, procedures for generating standard errors for regularized parameter estimates that are both computationally efficient and theoretically well-supported remain out of reach. Naturally, the simple, non-parametric bootstrap approach we pursue here has some short-comings [25]. Further work is needed to calibrate the SCCS 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 SCCS 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. To counter both criticisms, we are actively working towards a full Bayesian implementation to simultaneously estimate the appropriate strength of regularization and furnish well-calibrated measures of uncertainty on the regression coefficients.

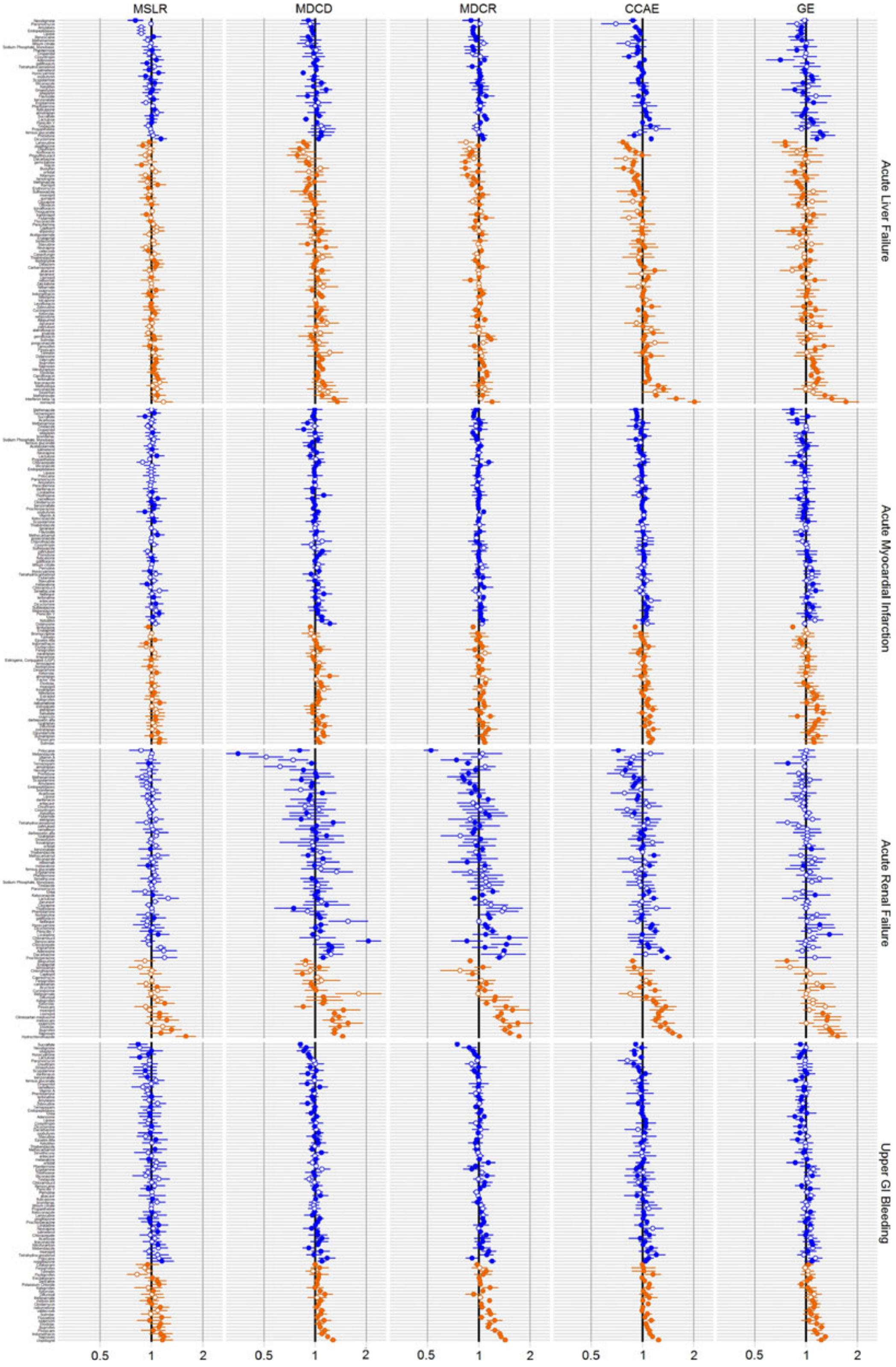
This study evaluates the performance of the SCCS on four outcomes that have been highlighted to be important events for monitoring in a risk identification system [26]. The integrity of the classification of the positive and negative controls, as well as the applicability of the drugs selected in each outcome to represent the distribution of expected scenarios for those outcomes, limits this study. While the SCCS shows robust predictive accuracy across these four outcomes, we also observe that optimal design choice and level of performance differ by outcome and data source. Consequentially, we caution against generalizing the results to other outcomes or other data sources. Instead, we encourage conducting a similar empirical evaluation prior to applying methods to new outcomes, whereby one benchmarks the method against a set of known positive and negative controls to gauge the expected predictive accuracy and estimate bias and coverage probability.

The SCCS exhibits promise as a potential tool for use in developing a risk identification system. Naturally, such a system should compare the relative merits of this design with alternative analytical strategies to determine which approaches should be used, independently or in complement. Based on empirical evidence of best practices, the community will be able to confidently move forward in building a risk identification system capable of systematically and efficiently generating reliable evidence from observational data to support the understanding of the effects of medical products.

Acknowledgements The Observational Medical Outcomes Partnership was funded by the Foundation for the National Institutes of Health through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Bristol - Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Johnson & Johnson, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, Takeda and Biogen Idec. Dr. Ryan is a past employee of GlaxoSmithKline, but does not receive compensation for his work with OMOP. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration and is presently an employee of Janssen Research and Development. Drs. Suchard and Madigan received 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



Self-controlled case series design estimates for all test cases stratified by data source

References

1. United States Congress. Food and drug administration amendments act of 2007. Public Law. 2007. p. 115–85.
2. Woodcock J, Behrman RE, Dal Pan GJ. Role of postmarketing surveillance in contemporary medicine. Ann Rev Med. 2011;62:1–10.
3. Stang PE, Ryan PB, Racoosin JA, Overhage JA, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Annal Intern Med. 2010;153:600–6.
4. Farrington CP Relative incidence estimation from case series for vaccine safety evaluation. Biometrics. 1995;51:228–35.
5. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. Stat Med. 2006;25(10):1768–97.
6. Grosso A, Douglas I, Hingorani A, MacAllister R, Smeeth L (2008) Post-marketing assessment of the safety of strontium ranelate; a novel case-only approach to the early detection of adverse drug reactions. British J Cli Pharmacol. 66:689–94.
7. Douglas I, Smeeth L (2008) Exposure to antipsychotics and risk of stroke: self controlled case series study. British Med J. 2008;337:a1227.
8. Nicholas JM, Grieve AP, Gulliford MC (2012) Within-person study designs had lower precision and greater susceptibility to bias because of trends in exposure than cohort and nested case-control designs. J Clin Epidemiol 65:384–93.
9. Taylor B, Miller E, Farrington CP, Petropoulos M-C, Favot-Mayaud I, Li J, Waight PA. Autism and measles, mumps and rubella vaccine: no epidemiological evidence for a causal association. Lancet. 1999;353:2026–9.
10. Hauben M, Madigan D, Gerrits C, Meyboom R. The role of data mining in pharmacovigilance. Expert Opin Drug Saf. 2005;4: 929–48.
11. Fram D, Almenoff J, DuMouchel W. Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. In: Ninth ACM SIGKDD international conference on knowledge discovery and data mining. 2003. p. 359–68.
12. Simpson SE, Madigan D, Zorych I, Schuemie MJ, Ryan PB, Suchard MA Multiple self-controlled case series for large-scale longitudinal observational databases. Biometrics (in press).
13. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55–67.
14. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear models. Trans Model Comput Simul 2013;23(1):10.
15. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf (in submission to this supplement). doi:10.1007/s40264-013-0110-2.
16. 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 (in this supplement issue). doi:10.1007/s40264-013-0108-9.
17. Ryan PB, Schmuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf (in submission to this supplement). doi:10.1007/s40264-013-0097-8.
18. 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.
19. Cantor SB, Kattan MW. Determining the area under the roc curve for a binary diagnostic test. Med Decis Mak. 2000;20(4):468–70.
20. Smith BM, Schwartzman K, Bartlett G, Menzies D (2011) Adverse events associated with treatment of latent tuberculosis in the general population. Can Med Assoc J 183(3):E173–9.
21. Bruno R, Sacchi P, Filice C, Filice G. Acute liver failure during lamivudine treatment in a hepatitis b cirrhotic patient. Am J Gastroenterol. 2001;96(1):265.
22. Clark SJ, Creighton S, Portmann B, Taylor C, Wendon JA, Cramp ME (2002) Acute liver failure associated with antiretroviral treatment for hiv: a report of six cases. J Hepatol. 36(2):295–301.
23. Tillmann HL, Hadem J, Leifeld L, Zachou K, Canbay A, Eisenbach C, Graziadei I, Encke J, Schmidt H, Vogel W, et al. Safety and efficacy of lamivudine in patients with severe acute or fulminant hepatitis b, a multicenter experience. J Viral Hepat. 2006;13(4):256–63.
24. Senn S (2008) Transposed conditionals, shrinkage and direct and indirect unbiasedness. Epidemiology. 19:652–4.
25. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. J Am Stat Assoc. 2011;106(494):608–25.
26. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salamé G, Catania MA, Salvo F, David A, Moore N 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.