# The Self-Controlled Case Series: Recent Developments

David Madigan
*Columbia University, U.S.A.*
madigan@stat.columbia.edu

Shawn Simpson
*Columbia University, U.S.A.*
shawn@stat.columbia.edu

Wei Hua
*Food and Drug Administration, U.S.A.*
wei.hua@fda.hhs.gov

Antonio Paredes
*Food and Drug Administration, U.S.A.*
antonio.paredes@fda.hhs.gov

Bruce Fireman
*Kaiser Permanente*
bruce.fireman@kp.org

Malcolm Maclure
*University of Bristish Columbia, Canada*
malcolm.maclure@gov.bc.ca

Summary

Purpose: To describe the self-controlled cases series and review

David Madigan is Professor of Statistics and Shawn Simpson is a PhD student at the Department of Statistics, 1255 Amsterdam Avenue, Columbia University, New York, NY 10027. Wei Hua and Antonio Paredes are with the Division of Epidemiology in the Center for Biologics Evaluation and Research at the FDA. Bruce Fireman is a Biostatistician and Research Scientist at the Division of Research, Kaiser Permanente Northern California. Malcolm Maclure is BC Academic Chair in Patient Safety and Professor in the Department of Anesthesiology, Pharmacology and Therapeutics at the University of British Columbia.

recent related methodological developments.

Methods: Literature review.

Results: The self-controlled case series offers several advantages for active surveillance for drug safety but we also outline some key limitations. We describe approaches for addressed some of these limitations.

Conclusions: The self-controlled case series model and its extensions may prove to be a useful tool for active surveillance.

Conflicts of Interest: None

Word Count: 3912

Key Points

- The self-controlled case series (SCCS) represents one particular methodology that may be useful for active surveillance of drug safety.
- SCCS has strengths and weaknesses.
- Modifications of the basic model can address some but not all of the weaknesses.
- Further research is required to establish the operating characteristics of SCCS-based active surveillance.

*Keywords and Phrases:* Drug safety; Shrinkage; Poisson regression; Case series;

## 1. INTRODUCTION

Increasing scientific, regulatory and public scrutiny focuses on the obligation of the medical community, pharmaceutical industry and health authorities to ensure that marketed drugs have acceptable benefit-risk profiles. This is an intricate and ongoing process that begins with carefully designed randomized clinical trials prior to approval but continues after regulatory market authorization when the drug is in widespread clinical use. In the post-approval environment, surveillance schemes based on spontaneous reporting systems (SRS) represent a cornerstone for the early detection of novel drug hazards. Key limitations of SRS-based pharmacovigilance include under-reporting, duplicate reporting, and the absence of a denominator or control group to provide a comparison.

Newer data sources have emerged that overcome some of the SRS limitations but present methodological and logistical challenges of their own.

Longitudinal observational databases (LODs) provide time-stamped patient-level medical information, such as periods of drug exposure and dates of diagnoses. Typical examples include medical claims databases and electronic health record systems. The scale of some of these databases presents interesting computational challenges – the larger claims databases contain upwards of 50 million lives with up to 10 years of data per life. A nascent literature on risk identification and refinement in LODs now exists including adaptations of some of the Bayesian methods developed in the SRS context.

In this paper we consider one particular approach, the self-controlled case series. We present a Bayesian analysis of this method and provide an overview of some recent related developments.

## 2. LONGITUDINAL OBSERVATIONAL DATABASES

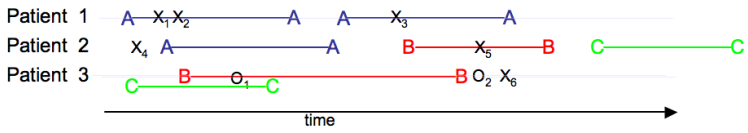


**Figure** 1: *A longitudinal observational dataset with three patients, three distinct drugs (A, B, and C) and two distinct outcome events (X and O)*

Figure 1 provides a schematic of LOD data for coverage periods for three patients. Patient 1 was exposed to drug A during two separate exposure periods. While on drug A, patient 1 experienced outcome event X on three different occasions. Patient 2 was exposed to drugs A, B, and C during successive non-overlapping eras. Patient 2 experienced outcome event X before consuming any drugs and also experienced outcome event X while consuming drug C. Patient 3 was exposed to drug C and later starting taking drug B in addition to drug C. This patient experienced outcome event O while taking both B and C and later experienced outcome events O and X after the drug B and C eras had ended. We note that LODs generally provide drug prescription dates so that construction of drug "eras" involves subtle decisions concerning gaps between successive prescriptions as well as off-drug risk periods. With outcome events, we think of outcomes as occurring at points in time whereas in truth outcomes are processes spread out in time.

The methodological challenge is to estimate the strength of the association between each drug and each outcome event, while appropriately accounting for covariates such as other drugs and outcome events, patient demographics, etc.

In this context, several papers have looked at vaccine safety, for example, Lieu *et al.* (2007), McClure *et al.* (2008), and Walker (2009). The Vaccine

Safety Datalink provides an early example of a LOD specifically designed for safety. Papers focusing on drug safety include Curtis *et al.* (2008), Jin *et al.* (2008), Kulldorff *et al.* (2008), Li (2009), Noren *et al.* (2008), and Schneeweiss *et al.* (2009).

## 3. THE SELF-CONTROLLED CASE SERIES METHOD

Farrington (1995) proposed the *self-controlled case series* (SCCS) method in order to estimate the relative incidence of adverse events to assess vaccine safety. The major features of SCCS are that (1) it automatically controls for time-fixed covariates that don't vary within a person during the study period, and (2) only cases (individuals with at least one event) need to be included in the analysis. With SCCS, each individual serves as their own control. In other words, SCCS compares outcome event rates during times when a person is exposed versus outcome event rates during times when the same person is unexposed. In effect, the cases' unexposed time lets us infer expectations about what would have happened during their exposed time had they not been exposed.

SCCS is one of several self-controlled methods that the epidemiology literature describes, many of which are variants on the case-crossover method (Maclure, 1991). However unlike the case-crossover method, which typically requires the choice of a comparator time period to serve as a control, SCCS makes use of all available temporal information without the need for selection.

Epidemiological applications of SCCS tend to focus on situations with small sample sizes and few exposure variables of interest. In contrast, the problem of drug safety surveillance in LODs must contend with millions of individuals and millions of potential drug exposures. The size of the problem presents a major computational challenge – ensuring the availability of an efficient optimization procedure is essential for a feasible implementation.

### 3.1. *One drug, one adverse event*

We will first focus on the case where there is one drug (e.g. Vioxx) and one adverse event (e.g. myocardial infarction, MI) of interest.

To set up the notation, $i$ will index individuals from 1 to $N$. Events and exposures in our databases are recorded with dates, so temporal information is available down to the level of days (indexed by $d$). Let $\tau_i$ be the number of days that person $i$ is observed, with $(i,d)$ being their $d$th day of observation. The number of events on day $(i,d)$ is denoted by $y_{id}$, and drug exposure is indicated by $x_{id}$, where $x_{id} = 1$ if $i$ is exposed to the drug on $(i,d)$, and 0 otherwise.

SCCS assumes that AEs arise according to a non-homogeneous Poisson process, where the underlying event rate is modulated by drug exposure. We

will start with the simple assumption that person $i$ has their own individual baseline event rate $e^{\phi_i}$, which is constant over time. Under the SCCS model, drug exposure yields a multiplicative effect of $e^{\beta}$ on the baseline incidence rate. In other words, the event intensity for person $i$ on day $d$ can be written as a function of drug exposure $x_{id}$.

$$\lambda_{id} = e^{\phi_i + \beta x_{id}}$$

The number of events observed on ($i$,$d$) given the current exposure status is distributed as a Poisson random variable with rate $\lambda_{id}$, which has the following density:

$$P(y_{id} \mid x_{id}) = \frac{e^{-\lambda_{id}} \lambda_{id}^{y_{id}}}{y_{id}!}$$

The SCCS likelihood contribution for person $i$ is the joint probability of the observed sequence of events, conditional on the observed exposures

$$L_i^c = P(y_{i1}, \ldots, y_{i\tau_i} \mid x_{i1}, \ldots, x_{i\tau_i}) = P(\mathbf{y}_i \mid \mathbf{x}_i) = \prod_{d=1}^{\tau_i} P(y_{id} \mid x_{id})$$

There are two assumptions implicit in the Poisson model that allow us to write out this likelihood:

(i) events are conditionally independent given exposures

$$y_{id} \perp\!\!\!\perp y_{id'} \mid \mathbf{x_i} \quad \text{for } d \neq d', \text{ and}$$

(ii) past events are conditionally independent of future exposures given the current exposure

$$y_{id} \perp\!\!\!\perp x_{id'} \mid x_{id} \quad \text{for } d \neq d'.$$

These assumptions are likely to be violated in practice (e.g., one might expect that having an MI increase the future risk of an MI and also impacts future drug usage), however they allow for simplifications in the model.

At this point one could maximixe the full log-likelihood over all individuals ($l^c = \sum_i \log L_i^c$) in order to estimate the parameters. However since our primary goal is to assess drug safety, the drug effect $\beta$ is of primary interest and the person-specific $\phi_i$ effects are *nuisance parameters*. A further complication is that claims databases can contain well over 10 million patients. Since the dimension of the vector of person-specific parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)'$ is equal to the number of individuals $N$, estimation of $\boldsymbol{\phi}$ would call for optimization in an ultra high-dimensional space and presumably would be computationally prohibitive.

In order to avoid estimating the nuisance parameter, we can condition on its sufficient statistic and remove the dependence on $\phi_i$. Under the Poisson model this sufficient statistic is the total number of events person $i$ has over their entire observation period, which we denote by $n_i = \sum_d y_{id}$. For a non-homogeneous Poisson process, $n_i$ is a Poisson random variable with rate parameter equal to the cumulative intensity over the observation period:

$$n_i \mid \mathbf{x}_i \sim \text{Poisson}(\sum_{d=1}^{\tau_i} \lambda_{id} = e^{\phi_i} \sum_{d=1}^{\tau_i} e^{\beta x_{id}})$$

In our case the cumulative intensity is a sum (rather than an integral) since we assume a constant intensity over each day. Conditioning on $n_i$ yields the following likelihood for person $i$:

$$L_i^c = P(\,\mathbf{y}_i \mid \mathbf{x_i},\, n_i\,) = \frac{P(\,\mathbf{y}_i \mid \mathbf{x_i}\,)}{P(\,n_i \mid \mathbf{x_i}\,)} \propto \prod_{d=1}^{\tau_i} \left( \frac{e^{\,\beta x_{id}}}{\sum_{d'} e^{\,\beta x_{id'}}} \right)^{y_{id}}$$

Notice that because $n_i$ is sufficient, the individual likelihood in the above expression no longer contains $\phi_i$. This conditional likelihood takes the form of a multinomial, but differs from a typical multinomial regression. Here the number of "bins" (observed days) varies by person, the $\beta$ parameter is constant across days, and the covariates $x_{id}$ vary by day.

Assuming that patients are independent, the full conditional likelihood is simply the product of the individual likelihoods.

$$L^c \propto \prod_{i=1}^{N} \prod_{d=1}^{\tau_i} \left( \frac{e^{\,\beta x_{id}}}{\sum_{d'} e^{\,\beta x_{id'}}} \right)^{y_{id}}$$

Estimation of the drug effect can now proceed by maximizing the conditional log-likelihood to obtain $\hat{\beta}_{CMLE}$. Winkelmann (2008) showed that this estimator is consistent and asymptotically Normal in the Poisson case.

It is clear from the expresson for the likelihood that if person $i$ has no observed events ($\mathbf{y}_i = \mathbf{0}$), they will have a contribution of $L_i^c = 1$. Consequently, person $i$ has no effect on the estimation, and it follows that only *cases* ($n_i \geq 1$) need to be included in the analysis.

SCCS does a *within-person* comparison of the event rate during exposure to the event rate while unexposed, and thus the method is "self-controlled". Intuitively it follows that if $i$ has no events, they cannot provide any information about the relative rate at which they have events. That the SCCS analysis relies solely on data from cases is a substantial computational advantage – since the incidence rate of most AEs is relatively low, typical SCCS analyses will utilize only a modest fraction of the total number of patients .

### 3.2. *Multiple drug exposures and drug interactions*

So far we have discussed the scenario where there is one AE and one drug of interest. However patients generally take multiple drugs throughout the course of their observation period. Additionally, patients may take many different drugs at the same time point, which leads to a potential for drug interaction effects. In order to account for the presence of multiple drugs and interactions, the intensity expression for the SCCS model can be extended in a natural way.

Suppose that there are $p$ different drugs of interest, each with a corresponding exposure indicator $x_{idj} = 1$ if exposed to drug $j$ on day $(i,d)$; 0 otherwise. Let $e^{\beta_j}$ be the multiplicative effect of drug $j$ on the event rate.

A multiplicative model describes the intensity for patient $i$ on day $d$:

$$\lambda_{id} = e^{\phi_i + \boldsymbol{\beta}' \mathbf{x}_{id}} = e^{\phi_i \,+\, \beta_1 \, x_{id1} \,+\, \cdots \,+\, \beta_p \, x_{idp}}$$

where $\mathbf{x}_{id} = (x_{id1}, \ldots, x_{idp})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.

Since $n_i$ is still sufficient for $\phi_i$, person-specific effects will once again drop out of the likelihood upon conditioning. One can derive the expression in a similar manner to the previous case of one AE and one drug case, resulting in:

$$L_i^c = P(\, \mathbf{y}_i \mid n_i, \, \mathbf{X}_i \,) \propto \prod_{d=1}^{\tau_i} \left( \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{id}}}{\sum_{d'} e^{\boldsymbol{\beta}' \mathbf{x}_{id'}}} \right)^{y_{id}} \quad \text{where} \quad \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1}' \\ \vdots \\ \mathbf{x}_{i\tau_i}' \end{bmatrix}$$

To simplify the summation in the denominator, days with the same drug exposures can be grouped together. Suppose that there are $K_i$ distinct combinations of drug exposures for person $i$. Each combination of exposures defines an exposure group, indexed by $k = 1, \ldots, K_i$.

For person $i$ and exposure group $k$, we need to know the number of events $i$ has while exposed to $k$ ($y_{ik}$) along with the length of time $i$ spends in $k$ ($l_{ik}$). For person $i$ we only require information for each of $K_i$ exposure groups, rather than for all $\tau_i$ days. This allows for coarser data and more efficient storage – since patients tend to take drugs over extended periods of time, $K_i$ is typically much smaller than $\tau_i$.

$$L^c \;\propto\; \prod_{i=1}^{N} \prod_{k=1}^{K_i} \left( \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{ik}}}{\sum_{k'} l_{ik'} e^{\boldsymbol{\beta}' \mathbf{x}_{ik'}}} \right)^{y_{ik}} \tag{1}$$

SCCS can be further extended to include interactions and time-varying covariates (e.g. age groups). The intensity on $(i,d)$ including two-way drug

interactions and a vector of time-varying covariates $\mathbf{z}_{id}$ can be written as

$$\lambda_{id} = e^{\,\phi_i + \boldsymbol{\beta}'\mathbf{x}_{id} + \sum_{r \neq s} \gamma_{rs}\, x_{idr}\, x_{ids} + \boldsymbol{\alpha}'\mathbf{z}_{id}}$$

where $\gamma$ denotes a two-way interaction between drugs $r$ and $s$.

*Remark 1.* In practice, many adverse effects can occur at most once in a given day suggesting a binary rather than Poisson model. One can show that adopting a logistic model yields an identical conditional likelihood to (1). This equivalence allows shifting to a logistic model with follow-up truncated at the outcome event, when that event is the onset of an enduring condition that permanently changes exposure propensity (see Discussion below.)

*Remark 2.* It is straightforward to show that the conditional likelihood in (1) is log-concave.

### 3.3. *Bayesian Self-Controlled Case Series*

We have now set up the full conditional likelihood for multiple drugs, so one could proceed by finding conditional maximum likelihood estimates of the drug parameter vector $\boldsymbol{\beta}$. However in the problem of drug safety surveillance in LODs there are millions of potential drug exposure predictors (tens of thousands of drug main effects along with drug interactions). This high dimensionality leads to potential overfitting under the usual maximum likelihood approach, so regularization is necessary.

We take a Bayesian approach by putting a prior over the drug effect parameter vector and performing inference based on posterior mode estimates. There are many choices of prior distributions that shrink the parameter estimates toward zero and address overfitting. In particular, we focus on the (1) Normal prior and (2) Laplacian prior.

(i) *Normal prior.* Here we shrink the estimates toward zero by putting an independent Normal prior on each of the parameter components. Taking the posterior mode estimates would be analogous to a ridge Poisson regression, placing a constraint on the $L_2$-norm of the parameter vector.

(ii) *Laplace prior.* Under this choice of prior a portion of the posterior mode estimates will shrink all the way to zero, and their corresponding predictors will effectively be selected out of the model. This is equivalent to a lasso Poisson regression, where there is a constraint on the $L_1$-norm of the parameter vector estimate.

Efficient algorithms exist for finding posterior modes, rendering our approach tractable even in the large-scale setting. In particular, we have adapted the cyclic-coordinate descent algorithm of Genkin *et al.* (2007) to the SCCS context. An open-source implementation is available at `http://omop.fnih.org`.

# 4. EXTENSIONS TO THE BASIC SCCS MODEL

## 4.1. *Relaxing the Independence Assumptions I: Events*

.

Farrington and Hocine (2010) present an approach that extends SCCS to allow for within-individual event dependence. This method treats the vector of observed event times $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})'$ for each individual $i$ as a single point in an $n_i$-dimensional region, where $n_i$ denotes the number of events experienced by individual $i$. This region is restricted to $Q_i(n_i) = \{\mathbf{t}_i \in (a_i, b_i]^{n_i} : t_{i1} < \cdots < t_{in_i}\}$ (where $(a_i, b_i]$ denotes the observation period for individual $i$) since the components of $\mathbf{t}_i$ are ordered by time, and no event times can occur outside of the observation window $(a_i, b_i]$. Standard SCCS assumes that events are realizations of a one-dimensional Poisson process and conditions upon the observed number of events $n_i$. Under Farrington and Hocine's model, however, the event time vector $\mathbf{t}_i$ is treated as a single point arising from an $n_i$-dimensional Poisson process. In this framework, conditioning on $n_i$ is equivalent to conditioning on the occurrence of a single point in the region $Q_i(n_i)$.

If $\lambda_i(t_1, \ldots, t_{n_i} \mid \mathbf{x}_i)$ is the intensity of the $n_i$-dimensional Poisson process on $Q_i(n_i)$, the conditional likelihood of $\mathbf{t}_i$ given the occurrence of one such point in $Q_i(n_i)$ is

$$L_i^{n_i} = \frac{\lambda_i(t_{i1}, \ldots, t_{in_i} \mid \mathbf{x}_i)}{\int\limits_{Q_i(n_i)} \lambda_i(u_1, \ldots, u_{n_i} \mid \mathbf{x}_i) du_1 \cdots du_{n_i}} \tag{2}$$

Farrington and Hocine assume that the $n_i$-dimensional Poisson intensity can be written in the form

$$\lambda_i(t_1, \ldots, t_{n_i} \mid \mathbf{x}_i) = \prod_{j=1}^{n_i} \lambda_i(t_j \mid \mathbf{x}_i) \times H_{n_i}(t_1, \ldots, t_{n_i}) \tag{3}$$

where the product term is made up of independent univariate intensities $\lambda_i(t \mid \mathbf{x}_i)$, and the $H_{n_i}(.)$ function determines the dependence between events. From (2) and (3) we can see that terms of $\lambda_i(t \mid \mathbf{x}_i)$ that are fixed in time will cancel out of the conditional likelihood, as they do in the original SCCS model. Similarly, fixed terms of $H_{n_i}(.)$ will also drop out of the conditional likelihood. Farrington and Hoccine explore different possible choices for $H$.

## 4.2. *Relaxing the Independence Assumptions II: The PD Model*

.

The PD-SCCS model (Simpson, 2011) extends SCCS to allow positive dependence between events, meaning that the occurrence of an event can

increase an individual's future event risk. Let $N_i(t)$ record the number of events that person $i$ has experienced up until time $t$. Assume, as before, that $i$ has $n_i$ total events during their observation period and that these events occur at times $t_{i1} < \cdots < t_{in_i}$. It is convenient to define a counting process, such as $N_i(t)$, in terms of its intensity function $\lambda_i(t \mid \mathbf{x}_i(t))$. This function gives the instantaneous probability that an event occurs at time $t$, given the history of the process and covariates. Under the SCCS model, the Poisson intensity for $i$ at time $t$ is

$$\lambda_i(t \mid \mathbf{x}_i(t)) = e^{\phi_i + \boldsymbol{\beta}' \mathbf{x}_i(t)} \tag{4}$$

as was previously described. PD-SCCS extends this model by incorporating $N_i(t^-)$, the number of events that $i$ has experienced up to but not including time $t$, as an additive effect on the individual baseline $e^{\phi_i}$. The PD-SCCS intensity function takes the form

$$\lambda_i(t \mid \mathbf{x}_i(t)) = (e^{\phi_i} + \delta N_i(t^-))\, e^{\boldsymbol{\beta}' \mathbf{x}_i(t)} \tag{5}$$

where $\delta$ is the parameter that controls the level of dependence between events. Based on plugging the PD-SCCS intensity (5) into the likelihood expression for a general intensity-based process, one can see that the total number of events $n_i$ is sufficient for the nuisance parameter $\phi_i$. As in the SCCS model, conditioning on $n_i$ removes $\phi_i$ from the likelihood expression. Symmetry arguments yield a closed form for the conditional likelihood, which in the denominator requires integrating over all possible ways for $i$ to have $n_i$ events during their observation period. Inference for $\boldsymbol{\beta}$ and $\delta$ is based on this conditional likelihood. Since the intensity function must be non-negative, the event dependence parameter is restricted to $\delta > 0$. In the case that $\delta = 0$, the PD-SCCS intensity model in (5) reduces to that of the SCCS model in (4).

### 4.3. *Relaxing the Independence Assumptions III: Exposures*

.

As discussed above, the SCCS model assumes that events are conditionally independent of subsequent exposures. Farrington *et al.* (2009) present an ingenious relaxation of this assumption using a counterfactual modeling approach. Their approach applies to the specific situation where the risk returns to its baseline level at the end of each risk period, where the event of interest is non-recurrent, and where the occurence of the event precludes future exposures.

Here we sketch the Farrington *et al.* approach using a simplified version of their running example. Consider a situation in which each individual can have up to two exposures. For individual $i$, again denote by $(a_i, b_i]$ the observation period and denote by $c_{i1}$ and $c_{i2}$ the actual exposure times, should they occur.

For notational simplicity, we consider point exposures followed by some known increased-risk time. The exposures then partition the observation period into up to five periods indexed by $j$: a control period, followed by an increased-risk period, followed by a control period, followed by a second increased-risk period, followed by a final control period. Denote by $n_{ij}$ the number of events occurring in the $j$th period, $n_{ij} \in \{0, 1\}$. Let $\beta_1$ and $\beta_2$ denote the log relative incidences associated with the first and second increased risk periods respectively and denote by $T_i$ the event time.

If $T_i$ occurs after $c_{i2}$ then no further exposures can occur and inference about $\beta_2$ can proceed in the usual fashion. Inference for $\beta_1$ is more complex in the situation where the event occurs after just one exposure becuase the timing of the counterfactual second exposure is then unavailable. Farrington *et al.* then make the following key observation: suppose, counterfactually, that no individual experienced a second exposure. Then it would be possible to estimate $\beta_1$ without bias. For this to work, we would need to know $n^*_{i4}$, the number of events in the fourth period, had no second exposure occurred. This is missing for those individuals that did in fact have a second exposure. However, $n_{i4}e^{-\beta_2}$ is an unbiased estimate of $n^*_{i4}$ for these individuals - this amounts to backing out the actual elevated risk during the second exposure. Using $n^*_{i4}$ in place of $n_{i4}$ then leads to an unbiased estimate of $\beta_1$.

Farrington *et al.* present the general case, an associated sandwich estimator for the variance, and also a computationally efficient equivalent approach based on pseudo likelihood.

We note that Roy *et al.* (2006) present an alternative approach.

### 4.4. *Structured SCCS Models*

. We are currently exploring several extensions to the basic model.

(i) *Hierarchical model: Drugs.* Drugs form drug classes. For example, Vioxx is a Cox-2 inhibitor. Cox-2 inhibitors in turn are non-steroidal anti-inflammatories. A natural extension assumes regression coefficients for drugs from within a single class arise exchangeably from a common prior distribution. This hierarchy could extend to multiple levels.

(ii) *Hierarchical model: AEs.* AEs also form AE classes. For example, an MI is a cardiovascular thrombotic (CVT) event, a class that includes, for example, ischemic stroke and unstable angina. In turn, CVT events belong to a broader class of cardiovascular events. This extension assumes that the regression coefficients for a particular drug but for different AEs within a class arise from a common prior distribution. Again this hierarchy could extend to multiple levels.

## 5. DISCUSSION

We have described self-controlled case series methods for post-approval drug safety risk estimation, some Bayesian and some not. Key advantages of the self-controlled case series approach include:

- SCCS adjusts for all time-invariant multiplicative confounders,
- Estimation requires only cases, and
- A regularized/Bayesian implementation of SCCS scales to large databases with the potential to adjust for large numbers of time-varying covariates.

The main problems with the SCCS approach concern the underlying independence assumptions, in particular, the assumption that events are conditionally independent, and the assumption that the exposure distribution and the observation period must be independent of event times. We described approaches to circumvent these assumptions and these may be useful in some applications.

Furthermore, since SCCS estimates the exposure-outcome association in cases, it ignores data on individuals in the study population that did not experience the outcome event. For example, there may be seasonality driving both the exposure and the outcome, where season is an important time-varying covariate. To adjust for seasonality, it is helpful to address both (a) the relation between season and the exposure, and (b) the relation between season and the outcome. While SCCS can incorporate time varying covariates, ignoring Sentinel's rich data on the non-cases limits our power to address (a). In another paper in this issue, we discuss how analyses of data from non-cases can supplement case-based analyses

We note that one possible approach to dealing with the exposure independence issue is to truncate observation time after the first event occurrence. This violates other SCCS assumptions but may still be useful in practice. Figure 2 shows estimates for a number of drug-outcome pairs with and without truncation. Clearly the truncation does alter some estimated relative risks substantially and future work will evaluate the empirical performance of this approach.

Real-life LODs are noisy and have the potential to introduce all sorts of artifacts and biases into analyses. For example, conditions and the drugs prescribed to treat the conditions are often recorded simultaneously at a single visit to the doctor, even though the condition actually predated the visit. This can introduce "confounding by indication" - the drug used to treat a condition can appear to be caused by the condition. Many such challenges exist and it remains to be seen whether or not false positives will render risk identification in LODs impractical.

Since all methods rely on dubious assumptions, future research will focus on establishing the operating characteristics of competing approaches. The Observational Medical Outcomes Partnership (OMOP) has empirically compared the predictive performance of SCCS, multivariate SCCS, and a wide variety of competing methods. Initial results suggest that SCCS is competitive with other methods and multivariate SCCS is a top performer. Nonetheless, the performance of all methods in OMOP leaves much room for improvement.

## REFERENCES

Curtis, J. R., Cheng, H., Delzell, E., Fram, D., Kilgore, M., Saag, K., Yun, H., and DuMouchel, W. (2008). Adaptation of Bayesian data mining algorithms to longitudinal claims data. *Medical Care*, **46**, 969-975.

Farrington, P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* **51**, 228–235.

Farrington, C. P., Whitaker, H. J. and Hocine, M. N. (2009). Case series analysis for censored, perturbed or curtailed post-event exposures. *Biostatistics*, **10**, 3–16.

Farrington, P. and Hocine, M.N. (2010). Within-individual dependence in self-controlled case series models for recurrent events. *Appl. Statist.* **59**, 457–475.

Genkin, A., Lewis, D. D., and Madigan, D. (2007) Large-scale Bayesian logistic regression for text categorization, *Technometrics* **49**, 291–304.

Jin, H., Chen, J., He, H., Williams, G.J., Kelman, C., and OḰeefe, C.M. (2008). Mining unexpected temporal associations: Applications in detecting adverse drug reactions. *IEEE Transactions on Information Technology in Biomedicine*, **12**, 488–500.

Kulldorff, M., Davis, R.L., Kolczak, M., Lewis, E., Lieu, T., and Platt, R. (2008). A maximized sequential probability ratio test for drug and vaccine safety surveillance. Preprint.

Li, L. (2009). A conditional sequential sampling procedure for drug safety surveillance. *Statistics in Medicine*. DOI:10.1002/sim.3689

Lieu, T.A., Kulldorff, M., Davis, R.L., Lewis, E.M., Weintraub, E., Yih, K., Yin, R., Brown, J.S., and Platt, R. (2007). Real-time vaccine safety surveillance for the early detection of adverse events. *Medical Care*, **45**, S89–95.

Maclure, M. (1991). The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology*, **133**, 144–153.

McClure, D. L., Glanz, J. M., Xu, S., Hambidge, S. J., Mullooly, J. P., and Baggs, J. (2008). Comparison of epidemiologic methods for active surveillance of vaccine safety. *Vaccine*, doi:10:1016/j.vaccine.2008.03.074.

Noren, G. N., Bate, A., Hopstadius, J., Star, K., and Edwards, I. R. (2008). Temporal pattern discovery for trends and transient effects: its application to patient records. In: *Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining SIGKDD 2008*, 963–971.

Roy, J., Alderson, D., Hogan, J.W., and Tashima, K. T. (2006). Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates, *J. Amer. Statist. Assoc.* **101**, 424–434.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity scoring adjustment in studies of treatment effects using health care claims data. *Epidemiology*, **20**, 512-522.

Simpson, S.E. (2011). The positive-dependence self-controlled case series model. Submitted.

Walker, A. M. (2009). Signal detection for vaccine side effects that have not been specified in advance. Preprint.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer.

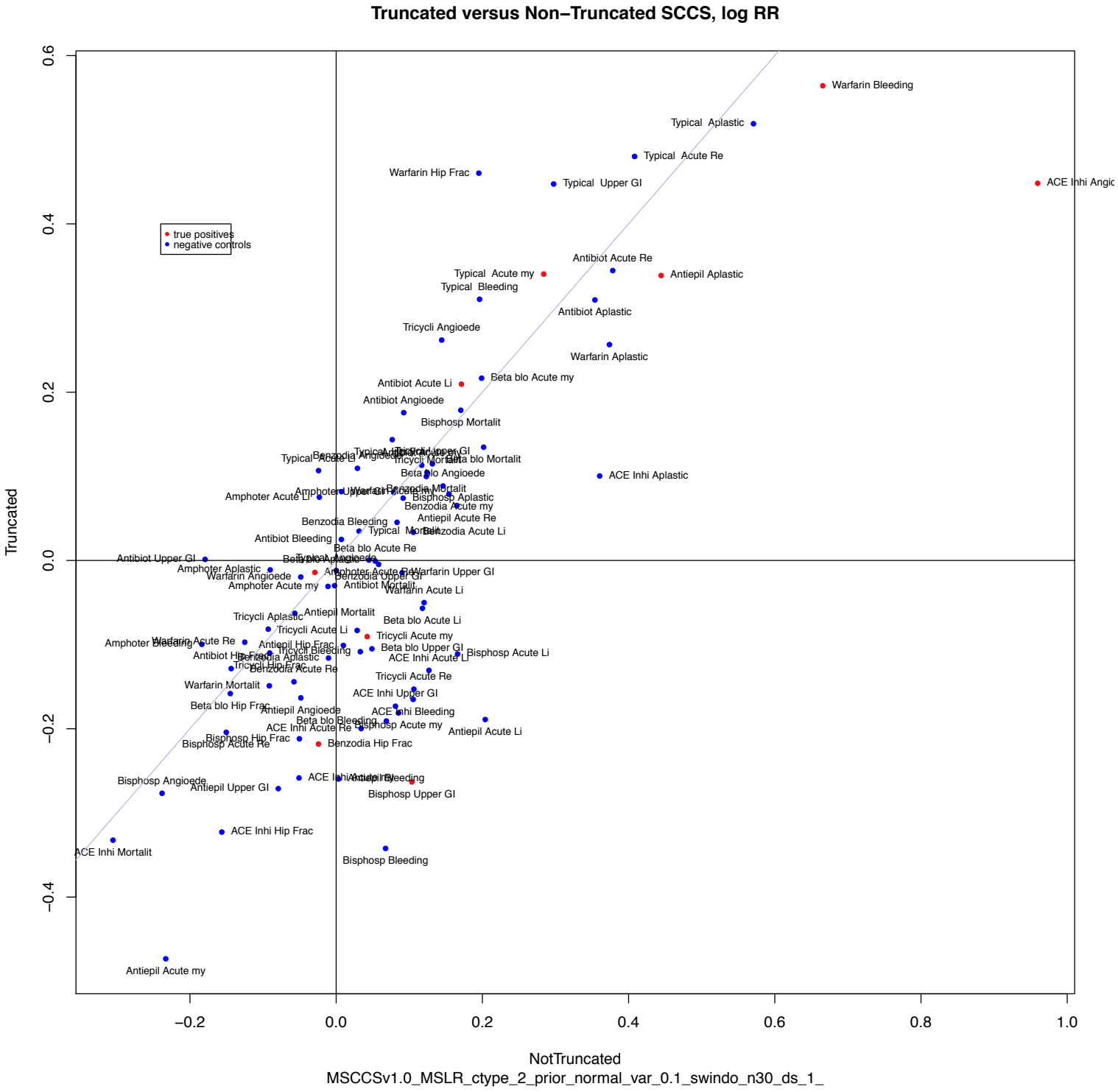


**Figure 2:** *Relative risk estimates from SCCS versus truncated SCCS for 53 drug-outcome pairs in the OMOP experiment.*