Biostatistics (2018), **0**, 0, *pp.* 1–28
doi:10.1093/biostatistics/convsccs

ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection

MARYAN MOREL*[,1], EMMANUEL BACRY[1,2], STÉPHANE GAÏFFAS[1,3], AGATHE GUILLOUX[1,4], FANNY LEROY[5],

[1] *CMAP Ecole polytechnique 91128 Palaiseau Cedex, France*
[2] *CEREMADE Université Paris-Dauphine, PSL, 75765 Paris Cedex 16, France*
[3] *LPMA Université Paris-Diderot, 75013 Paris, France*

[4] *LAMME, Univ. Evry, CNRS, Université Paris-Saclay, 91025, Evry, France* [5] *Caisse Nationale de l'Assurance Maladie, 75986 Paris Cedex 20, France*

maryan.morel@polytechnique.edu

Summary

With the increased availability of large electronic health records (EHRs) databases comes the chance of enhancing health risks screening. Most post-marketing detection of adverse drug reaction (ADR) relies on physicians' spontaneous reports, leading to under-reporting. To take up this challenge, we develop a scalable model to estimate the effect of multiple longitudinal features (drug exposures) on a rare longitudinal outcome. Our procedure is based on a conditional Poisson regression model also known as self-controlled case series (SCCS). To overcome the need of precise risk periods specification, we model the intensity of outcomes using a convolution between exposures and step functions, which are penalised using a combination of group-Lasso and total-variation. Up to our knowledge, this is the first SCCS model with flexible intensity able to handle multiple longitudinal features in a single model. We show that this approach improves the state-

*To whom correspondence should be addressed.

of-the-art in terms of mean absolute error and computation time for the estimation of relative risks on simulated data. We apply this method on an ADR detection problem, using a cohort of diabetic patients extracted from the large French national health insurance database (SNIIRAM), a claims database containing medical reimbursements of more than 53 million people. This work has been done in the context of a research partnership between Ecole Polytechnique and CNAMTS (in charge of SNIIRAM).

Key words: Conditional Poisson Model; Self-Controlled Case Series; Risk screening; Penalisation; Scalability; Total Variation

1. Introduction

In recent years, there has been a rapid increase in health data volume and availability. Large observational databases (LODs) such as claims databases contain electronic health records (EHRs) of millions of patients. One way to leverage this data is adverse drug reaction (ADR) detection. ADRs are adverse outcomes caused by drugs which might not have been detected during pre-licensing studies. ADRs can be related to multiple factors such as dose or time effects or even to patients' susceptibility due to genetic variation, gender, age, etc. (Aronson and Ferner, 2003). This paper focuses on time effects, i.e. on the relationship between ADR occurrences and occurrences of other past events (e.g. drug purchases), since it is known that some ADRs can be identified years after commercialisation (Downing *and others*, 2017).

While LODs have been used to investigate ADRs after spontaneous reports, a more extensive use could improve ADR detection by generating hypotheses directly from the data using screening strategies (Trifiro *and others*, 2009). In recent years, this perspective led to an increased research effort involving the use of LODs (Hripcsak *and others*, 2015).

However, using LODs for ADR screening is not a trivial task. This kind of data can be quite

heterogeneous, in terms of data types, structure, granularity and quality, due to fragmentation across multiple institutions for example. Several research projects are focusing on mitigating these issues. The Observational Medical Outcomes Partnership (OMOP, Overhage *and others* (2011)), and later the Observational Health Data Sciences and Informatics (OHDSI, Hripcsak *and others* (2015)) produced data models standards and methodologies allowing to improve EHR homogeneity across several institutions across several countries. In this work, we focus on the large French national claims database SNIIRAM (Tuppin *and others*, 2010). Its data is collected, harmonised and curated from multiple institutions across the country by CNAMTS, resulting in a country-wide claims database containing information on 83% of the French population. This database might be less biased than many LODs due to its large population coverage and quite accurate thanks to the automation of large parts of the data recording and cleaning processes.

A first challenge comes from the scale of the data. Indeed, LODs allows to study millions of patients across several years, hence it requires the use of scalable algorithms. The scalability must also be thought in terms of the number of drugs the patients are exposed to. When using LODs for risk screening, prior knowledge on the potentially problematic drugs might be scarce, consequently, the number of combination of drugs and outcomes to consider is potentially very large.

Many other challenges comes from the fact that EHR data tends to reflect the healthcare system rather than the patients' physiology. Indeed, EHR data are likely to contain non-random errors, record gaps, misleading timestamps and uncontrolled confounding (Hripcsak and Albers, 2013). For example, as the diagnoses result from clinical findings, raw timestamps could suggest that diseases follow their effects (Hripcsak *and others*, 2011). As a result, mapping complex, raw EHR data to clinical conditions is a very hard task, and is a research field by itself. While our work does not solve this problem, we hope to alleviate some it by using LODs to perform ADR screening.

There does not seem to be a clear consensus about which methods should be preferred when working with EHRs. However, models based on a self-control strategy, such as univariate self-control case series model (Farrington, 1995) or temporal pattern discovery algorithms (Norén *and others*, 2010) seems to perform better empirically than cohort and case-control methods (Ryan *and others*, 2013). The poor performance of case-control methods can be explained by the lack of proper "metadata" about patients (smoker status, wealth, etc.) in LODs, which are used to find proper controls in case-control studies. Besides, self-control methods might be more robust to unobserved confounders than cohort methods as they ignore non-longitudinal confounders (Farrington, 1995).

We focus on Self-Controlled Case Series (SCCS) models, originally developed for vaccine safety studies (Farrington, 1995), since then applied in post-marketing studies using LODs (Gault *and others*, 2017). The SCCS model scales quite well since it is fitted on cases only. Moreover, as explained below, its goodness-of-fit function cancels out non-longitudinal confounders, which reduces potential non-longitudinal biases. Thus, an SCCS model helps with the scalability and unobserved confounding issues described earlier. However, an SCCS model relies heavily on the definition of a time-at-risk period, which makes it hard to use in multivariate settings.

Previous attempts to solve this problem relied on the use of splines to provide a more flexible modelling of drug effects (Schuemie *and others*, 2016; Ghebremichael-Weldeselassie *and others*, 2016, 2017). However, the use of splines makes the estimation of the model more complicated, resulting in models able to fit the effect of a single drug in addition to a temporal baseline. This can be problematic when performing ADRs screening, as SCCS is sensitive to temporal confounders, and thus, to the omission of longitudinal features.

This paper introduces a new approach in the framework of SCCS models that addresses the three challenges mentioned previously:

- it considers several longitudinal features at the same time (longitudinal drug exposures),

- it cancels out non-significant drug effects automatically

- it learns automatically and in a flexible way the significant drug effects, with no precise knowledge on a time-at-risk period,

- it runs faster than comparable algorithms when studying many drugs at a time.

Hence, it provides an important extension to the usage of SCCS models, allowing to *study multiple exposures at the same time, while requiring much less attention to the definition of time-at-risk periods*. An application of this methodology is described in Section 4.2 below, and leads to a scalable approach with respect to the number of drugs. On the one hand, it does not require a high precision work when preparing the dataset (as done in (Neumann *and others*, 2012)). On the other hand, it is not thought as a replacement of such approaches but rather as a screening method to identify potential problematic drugs that might require specific subsequent investigations (using (Neumann *and others*, 2012) types of approach).

The paper is organised as follows. We first describe SCCS models in Section 2 and construct our method in Section 3. Numerical experiments are given in Section 4. It includes in Section 4.1 experiments on simulated data, with a comparison to state-of-the-art methods from the SCCS literature. In particular, these simulations as designed to reproduce some of the problems met with the data used in Section 4.2, in order to test the robustness of our algorithm compared to the state-of-the-art. Section 4.2 gives an application of our method on a LOD from the French national health insurance information system (SNIIRAM, a database built around medical reimbursements of more than 53 million people). Our model produces consistent results with a population-based cohort study (Neumann *and others*, 2012) when estimating the effect of pioglitazone (a hypoglycemic agent) on the risk of bladder cancer. A conclusion is given in Section 5, and mathematical and numerical details are provided in Supplementary Material.

2. Self-controlled case series models

SCCS models allow to estimate the impact of longitudinal features (such as time-varying exposures to drugs) on the occurrence intensity of events of interest (such as dates of adverse events), see (Farrington, 1995). An interesting particularity with this family of methods is that individuals form their own controls: individuals who do not experience the event of interest are not used to fit the model. This construction relies on the property of order statistics of the Poisson process and the statistical output of such models is an estimation of the *relative incidence* of the longitudinal features, i.e. the relative increase of the outcomes intensity.

2.1 *Conditional Poisson regression and SCCS models*

Data is available from a global observation period $(a, b]$, where the time can be either calendar or measured by the age of individuals. Each patient $i = 1, \dots, m$ has an observation period $(a_i, b_i] \subset (a, b]$, in which we observe:

- the time occurrences $t_{i,1} < t_{i,2} < \cdots$ of the event of interest (also called *outcome* in what follows), or, equivalently a counting process N_i, defined as $N_i(t) = \sum_{k \geqslant 1} \mathbf{1}_{t_{i,k} \leqslant t}$ and $n_i = \int_{(a_i, b_i]} dN_i(t)$ the total number of outcomes of patient i, ;
- a vector of d longitudinal features

$$X_i = \big(X_i(t) = (X_i^1(t) \ \cdots \ X_i^d(t)) : t \in (a_i, b_i]\big),$$

  where in the context of drug safety studies, $X_i^j(t)$ gives us information about the exposure of patient i to drug j at time $t \in (a, b]$.

The model developed in this paper relies on the usual SCCS model key assumptions (Farrington and Whitaker, 2006). Namely, we assume that

(1.) The features are exogenous, meaning that the counting process N_i does not have any

influence on the features X_i;

(2.) The interval of observation $(a_i, b_i]$ is independent of N_i;

(3.) The process N_i is a Poisson process conditionally to $(X_i(t) \ : \ t \in (a_i, b_i])$.

Assumption (1.) allows to condition on the full trajectory of the longitudinal features X_i in (2.1). In addition, thanks to (2.), the following derivations have to be understood conditionally to $(a_i, b_i]$. We may then define the conditional intensity of process N_i as

$$\lambda_i(t, X_i) = \mathbb{P}(dN_i(t) = 1 \mid X_i) \tag{2.1}$$

for $t \in (a_i, b_i]$. This model can be, therefore, understood as a regression model, allowing to regress the outcomes in N_i on the longitudinal features X_i.

In order to study acute vaccine adverse effects, (Farrington and Whitaker, 2006) considers the following model for the intensity:

$$\lambda(t, X_i) = \exp\big(\psi_i + \gamma_i + \phi(t) + X_i(t)^\top \beta\big),$$

where ψ_i is the baseline incidence of patient i and γ_i is a sum of non-temporal fixed and random individual effects. The parameter $\phi(t)$ is a time-dependent baseline which is common to all individuals. If age is used as the time scale, this term can help to capture age effects. The vector of parameters $\beta \in \mathbb{R}^d$ quantifies the effect of the longitudinal features $X_i(t)$ on the intensity. The idea of the SCCS method is to condition on both X_i and n_i. Usual arguments (see Section 1 in Supplementary Material) imply that the likelihood of $N_i|(X_i, n_i)$ of $i = 1, \dots, m$ independent patients is proportional to

$$\prod_{i=1}^{m} \prod_{k=1}^{n_i} \frac{\lambda_i(t_{i,k}, X_i)}{\int_{a_i}^{b_i} \lambda_i(s, X_i) ds} = \prod_{i=1}^{m} \prod_{k=1}^{n_i} \frac{\exp\big(\phi(t_{i,k}) + X_i(t_{i,k})^\top \beta\big)}{\int_{a_i}^{b_i} \exp\big(\phi(s) + X_i(s)^\top \beta\big) ds}. \tag{2.2}$$

Note that the conditioning with respect to n_i induced two notable properties of (2.2):

- *Improved scalability:* the likelihood only depends on patients i such that $n_i \geqslant 1$ (while the "full" likelihood of $N_i|X_i$ does depend on patients i for whom $n_i = 0$). This is beneficial when studying rare adverse effects in large LODs.

- *Robustness to non-longitudinal confounders:* the non-longitudinal effects ψ_i and γ_i cancel out in the likelihood (2.2). This makes SCCS models particularly robust to the patient's susceptibility.

These two properties are appealing when working with LODs such as claims databases, as it helps to mitigate issues related to missing variables and the data scale. However, only relative incidences can be computed by taking the exponential of the corresponding coefficient, such as $\exp(\phi(t))$ for the baseline relative incidence.

SCCS models were initially designed for vaccine safety studies (Farrington, 1995), using the suspected ADR as the outcome. In this context, estimating the relative incidence of drug use requires defining related time-at-risk periods in which the suspected ADR might occur. The longitudinal features $X_i(t)$ are then used to express the fact of being at risk or not at time t for a particular drug. One must then determine for how long patients are at risk after each exposure to a drug, and if this risk occurs either immediately or after some amount of time. Defining proper time-at-risk windows is a hard problem when studying a single (drug, ADR) pair, which worsens even further when considering a set $(\text{drug}_1, \text{ADR}), \ldots, (\text{drug}_d, \text{ADR})$ of such pairs. In the case of ADR screening over multiple drugs, such a methodology might even become inappropriate.

2.2 *Risk screening*

When prior knowledge on time-at-risk windows is not available, a simple method is to use a large window in order to be sure to capture the potential effect. However, this strategy typically "dilutes" the risk over the window, see (Xu *and others*, 2011), leading to a model unable to detect ADRs. Existing works propose to relax the time-at-risk window definition while trying

to overcome this risk dilution. It is proposed in (Xu *and others*, 2011) to select an optimal risk window by testing several window sizes, in a data-driven fashion. However, this method is difficult to adapt for ADR screening when considering d drugs and q risk windows at the same time, since it requires to fit q^d models.

A different approach relies on fitting time-dependent parameters in order to estimate the risk of ADR over large risk windows. The model estimates a time-varying relative incidence function all along the risk window instead of assuming it to be constant. This approach is used in (Schuemie *and others*, 2016), where the drug effect is a function θ of the accumulated exposures. It uses a discrete model with daily granularity, assuming that the integral of $X_i(t)$ over one day is equal to 1 when the patient is exposed to the studied drug. Accumulated exposures up to time t is measured by $\int_{a_i}^{t} X_i(s)\mathrm{d}s$, where $X_i(t)$ is univariate, and expresses the exposure to a single drug at time t, leading to the following model for the intensity:

$$\lambda_i(t, X_i) = \exp\Big(\psi_i + \gamma_i + \phi(t) + \theta\Big(\int_{a_i}^{t} X_i(s)\mathrm{d}s\Big) + X_i(t)\beta\Big),$$

where the function θ is estimated using natural cubic splines. As the splines are not regularised, this model might be prone to overfitting. Alternatively, (Ghebremichael-Weldeselassie *and others*, 2016) use a convolution to model drug effects, writing the intensity as

$$\lambda_i(t, X_i) = \exp\big(\psi_i + \gamma_i + \phi(t)\big) \int_{a_i}^{t} X_i(s)\theta(t-s)\mathrm{d}s.$$

In this model, $X_i(t)$ is either a point exposure $X_i(t) = \delta_{c_i}(t)$ where δ_{c_i} stands for a Dirac mass at date $c_i \in \mathbb{R}^+$, or a continuous exposure to a constant quantity x, namely $X_i(t) = x\mathbf{1}_{(c_i, b_i]}(t)$. In the former case, the intensity can be expressed as

$$\lambda_i(t, X_i) = \exp\left(\psi_i + \gamma_i + \phi(t)\right)\theta(t - c_i). \tag{2.3}$$

The function θ is estimated using M-splines (in order ensure positivity) in (Ghebremichael-Weldeselassie *and others*, 2016, 2017), while the age effect ϕ is estimated by step functions

in (Ghebremichael-Weldeselassie *and others*, 2016) and by splines in (Ghebremichael-Weldeselassie *and others*, 2017). The considered model could deal with multiple point exposures c_i for the drug, given that the maximum time gap between successive exposures is smaller than the support of θ, but the authors have not developed this point.

Both (Schuemie *and others*, 2016) and (Ghebremichael-Weldeselassie *and others*, 2016, 2017) seem restricted to the study of a single (drug, ADR) pair at a time. This can be problematic since SCCS is sensitive to time-varying confounders and benefits from studying multiple drugs at once as shown by both (Simpson *and others*, 2013) and (Moghaddass *and others*, 2016). In order to fit an SCCS model using several drugs at the same time, (Ghebremichael-Weldeselassie *and others*, 2017) propose to extend their work by modelling additional drugs effect with step functions instead of splines. However, such functions are basically not regularised, which can result in overfitting, and are very sensitive to the chosen number of steps.

3. ConvSCCS: an extension of SCCS models

We now introduce our ConvSCCS model. It is an extension of the classical SCCS model in several directions. First, it allows considering exposures to several drugs. More importantly, our model is time-invariant thanks to a convolutional structure. Hence it can learn the potential effects of the drug exposures even without prior definition of precise time-at-risk periods.

More specifically, we construct a model that estimates the effect of longitudinal features using convolutions of low-granularity step functions with point drug exposures. The low-granularity leads to an over-parametrised model with poor estimation accuracy. We solve this issue in Section 3.2 below by using a penalisation technique that combines total-variation and Group-Lasso penalties. The second will perform an automatic variable selection, while the first enforces longitudinal effects to be piece-wise constant over larger steps whenever statistically relevant. As illustrated in Section 4, this leads to improvements over current state-of-the-art methods, and

provides interpretable results on the observational database considered in this paper, see Section 4.2.

3.1 *Discrete convolutional SCCS*

We assume that, for $i = 1, \dots, m$, the intensity λ is constant over time intervals $I_k = (t_k, t_{k+1}]$, $k = 1, \dots, K$ that form a partition of the observation interval $(a, b]$. Without loss of generality, we choose I_k to be of constant length 1. In practice, we use the smallest granularity allowed by data. Hence, we can assume that $(a_i, b_i] \cap I_k$ is either $\emptyset$ or I_k for all $i = 1, \dots, m$, and $k = 1, \dots, K$, which means that the observation period of each individual is a union of intervals I_k. Denoting by $\lambda_{i,k}$ the value of $\lambda(t, X_i(t))$ for $t \in I_k$, and defining $y_{ik} := N_i(I_k)$, the discrete SCCS likelihood can be written as

$$L(y_{i1}, \dots, y_{ik} | n_i, X_i) = n_i! \prod_{k=1}^{K} \left(\frac{\lambda_{ik}}{\sum_{k'=1}^{K} \lambda_{ik'}} \right)^{y_{ik}},$$

where we use the convention $0^0 = 1$, i.e. only the exposition period $(a_i, b_i]$ contributes to the likelihood, and since $N_i(I_k) = \lambda_{ik} = 0$ whenever $I_k \cap (a_i, b_i] = \emptyset$, see Section 2 of Supplementary Material for more details. We consider an intensity given by

$$\lambda_i(t, X_i) = \exp\Big(\psi_i + \gamma_i + \phi(t) + \int_{a_i}^{t} X_i(s)^\top \theta(t - s) \mathrm{d}s\Big).$$

Since the intensity is constant on each I_k, it can be rewritten as

$$\lambda_{ik}(X_i) = \exp\Big(\psi_i + \gamma_i + \phi_k + \sum_{k'=a_i}^{k} X_{ik'}^\top \theta_{k-k'}\Big),$$

where X_{ik} stands for the value of $X_i(t)$ for $t \in I_k$ and $\theta \in \mathbb{R}^{d \times K}$. We observe $l = 1, \dots, L_i^j$ starting dates of exposures c_{il}^j and introduce the features $X_{ik}^j = \sum_{l=1}^{L_i^j} \mathbf{1}_{k = c_{il}^j}$, which leads to the following intensity

$$\lambda_{ik}(X_i) = \exp\Big(\psi_i + \gamma_i + \phi_k + \sum_{j=1}^{d} \sum_{l=1}^{L_i^j} \theta^j_{k - c_{il}^j} \mathbf{1}_{[0,p]}(k - c_{il}^j)\Big). \tag{3.4}$$

The quantity $\exp(\theta_k^j)$ corresponds to the relative incidence of an exposure to drug j that occurs k time units after an exposure start. Finally, the likelihood is equal to

$$L(y_{i1},\dots,y_{ik}|n_i,X_i)=\prod_{k=1}^{K}\left(\frac{\exp\left(\phi_k+\sum_{j=1}^{d}\sum_{l=1}^{L_i^j}\theta^j_{k-c_{il}^j}\mathbf{1}_{[0,p]}(k-c_{il}^j)\right)}{\sum_{k'=1}^{K}\exp\left(\phi_{k'}+\sum_{j=1}^{d}\sum_{l=1}^{L_i^j}\theta^j_{k'-c_{il}^j}\mathbf{1}_{[0,p]}(k'-c_{il}^j)\right)}\right)^{y_{ik}} \tag{3.5}$$

and depends only on the parameters θ for the exposures and the age effects ϕ.

3.2 *Penalised estimation*

This formulation of intensity (3.4) is flexible since it allows to capture an immediate effect in θ_0^j, or delayed ones using θ_k^j for $k \geqslant 1$. This flexibility comes at a cost: it increases significantly the number of parameters to be estimated, which might lead to inaccurate estimations and to over-fitting of the dataset. To that end, we introduce a penalisation technique which allows handling this issue, and which provides interpretable estimations of the relative risks as a byproduct.

We introduce groups $\theta^j = [\theta_1^j \cdots \theta_p^j] \in \mathbb{R}^p$ of parameters quantifying the impact of exposures to drugs $j = 1,\dots,d$ at different lags $k = 1,\dots p$. To avoid exposure effects overlapping, we assume that exposure starting times are far enough, that is $\min_{l,l'} |c_{il}^j - c_{il'}^j| > p$. We want to induce two properties on the relative risks of drugs exposures: a "smoothness" property along lags $k = 1,\dots,p$, namely we want consecutive relative risks $\exp(\theta_k^j)$ and $\exp(\theta_{k-1}^j)$ to be basically close; and the possibility for a drug to have no effect, namely to induce that θ^j can be the null vector. This can be achieved with the following penalisation that combines total-variation and group-Lasso

$$\text{pen}(\theta) = \gamma_{\text{tv}} \sum_{j=1}^{J}\sum_{k=1}^{p-1} |\theta_{k+1}^j - \theta_k^j| + \gamma_{\text{gl}} \sum_{j=1}^{J} \|\theta^j\|_2 \tag{3.6}$$

over the groups θ^j for $j = 1,\dots,d$, where $\gamma_{\text{tv}} \geqslant 0$ and $\gamma_{\text{gl}} \geqslant 0$ are respectively levels of penalisation for the total-variation and the group-Lasso. The group-Lasso introduced in (Yuan and Lin, 2006) acts like the lasso at the group level: depending on γ_{gl}, it can cancel out a full block θ^j. Total-variation penalisation is known to consistently estimate change points for the estimation of the

intensity of a Poisson process, see (Alaya *and others*, 2015).

We write the penalised negative log-likelihood of our model as follows:

$$-\ell(\phi, \theta) + \text{pen}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log \left(\frac{\lambda_{ik}(\phi, \theta)}{\sum_{k'=1}^{K} \lambda_{ik}(\phi, \theta)} \right) + \text{pen}(\theta), \tag{3.7}$$

where pen is given by (3.6) and where we recall that

$$\lambda_{ik}(\phi, \theta) = \exp\Big(\phi_k + \sum_{j=1}^{d} \sum_{l=1}^{L_i^j} \theta^j_{k - c_{il}^j} \mathbf{1}_{[0,p]}(k - c_{il}^j) \Big).$$

The function (3.7) is convex and $\ell(\phi, \theta)$ is gradient-Lipschitz. However, since the sparsity-inducing penalisation $\text{pen}(\theta)$ is not differentiable, we use a proximal first-order method to minimise efficiently (3.7). Namely, we use the state-of-the-art SVRG algorithm from (Xiao and Zhang, 2014), which is a fast stochastic proximal gradient descent algorithm, using a principle of variance reduction of the stochastic gradients.

Finally, the hyper-parameters γ_{tv} and γ_{gl} are selected using a stratified V-Fold cross-validation on the negative log-likelihood.

4. Experiments

In this section, we compare ConvSCCS with the state-of-the-art, namely SmoothSCCS (Ghebremichael-Weldeselassie *and others*, 2016) and NonparaSCCS (Ghebremichael-Weldeselassie *and others*, 2017), that are described below, see also Section 2.2 for further details.

ConvSCCS is the method introduced in this paper: an extension of SCCS models allowing to fit the effect of *several* drugs on an ADR in a flexible way, see also Table 1 below. ConvSCCS is available in our open-source `tick` library (`https://x-datainitiative.github.io/tick/`), see Section 3 in Supplementary Material for details. The code used to run the experiments described in this paper is available in GitHub, at `https://github.com/MaryanMorel/ConvSCCS`.

SmoothSCCS is introduced in (Ghebremichael-Weldeselassie *and others*, 2016), which uses splines to model the effect of a *single* drug exposure to a disease and step functions to model

the effect of age. We use the `SCCS` R package implementation, available at `http://statistics.open.ac.uk/sccs/r.htm`. We use 12 knots and six groups of age as suggested in (Ghebremichael-Weldeselassie *and others*, 2016). Since this model is designed to fit (drug, ADR) pairs, we fit the model on each drug successively.

NonparaSCCS is introduced in (Ghebremichael-Weldeselassie *and others*, 2017) which uses splines to model both the effect of drug exposure and age. We use the same R package and settings as the ones described for SmoothSCCS.

We did not include (Schuemie *and others*, 2016) as we have not found any open source implementation of this work. We have not tried to use (Simpson *and others*, 2013) since we do not have precise priors on relevant risk periods in the context of ADR screening.

4.1 *Simulations*

The performances of our model against SmoothSCCS and NonParaSCCS are compared in a simulation study. For this purpose, multivariate longitudinal exposures and outcomes are simulated, with a correlation structure between exposures.

Simulation of longitudinal features. The simulation of correlated longitudinal features is a difficult task, for which we use Hawkes processes, see (Hawkes and Oakes, 1974), which is a family of counting process with an autoregressive intensity, see Section 5 of the Supplementary Material for more details. Our simulation setting has been chosen so that it generates correlated exposures, as it is the case with actual exposures from the LOD considered in this paper.

Simulation of relative risks. We assume that all simulated adverse outcomes can take place at most 50 time intervals after the first exposure. We consider two sets of relative risk profiles from (Ghebremichael-Weldeselassie *and others*, 2017) and (Aronson and Ferner, 2003). These sets are precisely described in Section 5 of the Supplementary Material, and contain several types

and shapes of risks profiles.

Simulation of outcomes. We simulate $m = 4000$ patients' exposures over $K = 750$ time intervals. The observation periods are set to $[0, b_i]$, where $b_i = K - e_i$ and e_i are is from an exponential distribution with intensity $1/250$. Intensities λ_{ik} are set to zero for all $k > b_i$. The outcomes are simulated according to a multinomial distribution $\text{Mult}(1; p_{i,0}, \ldots, p_{iK})$ where $p_{ik} = \lambda_{ik} / \sum_{k'=1}^{K} \lambda_{ik'}$.

Sensitivity analysis. We perform extensive simulations to test the robustness of our model to bias sources specific to EHR data using the following scenarios, namely not-at-random missing data, noisy timestamps, missing longitudinal features, see Section 5 of the Supplementary Material for more details.

Performance measure. The performance of the different models is computed using the mean absolute error (MAE) between the estimated relative incidence and the true risk profile, see Section 5 of the Supplementary Material for details. For both sets of relative risk profiles, we simulate $m = 4000$ cases and simulate 100 datasets for each scenario.

Results. Boxplots representing the MAE distribution over the 100 simulated datasets are represented in Figures 1 and 2. In Set 1 of relative exposures, which is an "easy" setting (4 features and 8 non-zero correlations, see Supplementary Material), the gain resulting from studying several drugs at a time seems to be balanced by the bias resulting from using step functions when fitting smooth risk profiles. Indeed, as shown by Figure 1, the estimation errors of drug exposures relative risks are similar across the three considered models. For the baseline estimation, NonParaSCCS performs better than ConvSCCS and SmoothSCCS since the use of splines results in a better approximation than the step functions with six groups of age.

In Set 2 of relative exposures, which is a more difficult setting (14 features, with 24 non-zero correlations, see Supplementary Material), ConvSCCS outperforms both SmoothSCCS and Non-

ParaSCCS. We observe in Figure 2 that fitting the effect of several drugs at the same time and using our penalisation provides a better estimation accuracy than NonParaSCCS and SmoothSCCS, the improvement being larger for the estimation of drugs exposures risks profiles than for the baseline. This illustrates the benefits of fitting several drugs at the same time in the context of an SCCS model. Figure 3 gives the run times of all three procedures. ConvSCCS seems to scale better than both SmoothSCCS and NonParaSCCS when fitting a large number of feature such as $d = 14$ on $m > 2000$ cases. In small studies, however, when $d = 4$ for example, SmoothSCCS is the fastest algorithm, while NonParaSCCS is overall slower than the two other algorithms. According to its improved performance and scalability when studying several drugs, ConvSCCS seems to be a useful model for ADR screening on LODs.

The sensitivity analysis shows that the model is robust to small to moderate noise in timestamps, but its performances degrade with large noise (see Figure 4 in Supplementary Material). With large noise, the model over-penalizes (with the group-Lasso), resulting in a constant relative incidence for each feature. In such situations, reducing the granularity might help to reduce the noise level, but it might also dilute the risk. The model does not seem particularly sensitive to not-at-random missing data or to slightly correlated missing features, see Figures 5, 6 and 7 in Supplementary Material for more details.

4.2 *Application on data from the French national health insurance information system*

We investigate the association between glucose-lowering drugs and the risk of bladder cancer in France with data from the SNIIRAM/PMSI database. Using similar data, a significant association between pioglitazone (glucose-lowering drug) and bladder cancer was reported in (Neumann *and others*, 2012). As a result of this study, the use of pioglitazone was suspended in France in June 2011. Note that other studies, such as Lewis *and others* (2015) did not conclude to a significant effect on this particular association.

The SNIIRAM/PMSI database. The data was extracted from the French national health insurance information system (*Système National d'Information Inter-régimes de l'Assurance Maladie* (SNIIRAM), see (Tuppin *and others*, 2010)) linked with the French hospital discharge database (*Programme de Médicalisation des Systèmes d'Information* (PMSI), see (ATIH) website), in the context of a research partnership between Ecole Polytechnique and CNAMTS. The full SNIIRAM/PMSI database is an SQL database containing hundreds of tables built around medical reimbursements of more than 53 million people (its size is between 150 and 200 TB). Our team set up a 15 nodes Spark cluster and developed an ETL (Extract Transform Load) pipeline to transform the data into a single patient-centric table that can be used to build features that feed various statistical inference algorithms.

Cohort, ADR and expositions definitions. The cohort includes patients covered by the general insurance scheme aged 40 to 79 years on 2006/12/31 who filled at least one prescription for a glucose-lowering drug in 2006. The end of the observation period was set on 2009/12/31. The glucose-lowering drugs investigated are insulin, metformin, sulfonylurea, pioglitazone, rosiglitazone, and other oral hypoglycemic agents.

All patients with any bladder cancer-related events in the six months before follow-up start have not been included. So although the depth of the data was 48 months, the cohort was followed for up to 42 months. The considered outcomes can then be treated as incident cases. We use the same definition for the bladder cancer outcome as in (Neumann *and others*, 2012), which adds particular procedures to a hospital discharge diagnosis (ICD-10-C67). The cohort contains 1699 patients with bladder cancer. Note that we have roughly 400 cases missing in comparison to Neumann *and others* (2012), and less history prior to follow-up to filter prevalent cases, due to French data regulation imposing patients information to be deleted after ten years. More details about cohort structure can be found in Table 1 in supplementary materials.

We consider that patients are exposed to a molecule as soon as they purchase a drug containing

this molecule. Once a patient has been exposed, she is considered as exposed until the end of her follow-up. There is a potential bias concerning drug exposures. Indeed, diabetic patients use hypoglycemic agents continuously. As a result, exposure starting dates might exhibit noisy timestamps.

ConvSCCS. We apply ConvSCCS to the cohort with bladder cancer, and use the smallest available granularity: 30-days time intervals based on calendar time and consider a risk window of 24 months. We do not use age-related features and consider its effect to be part of patients' baseline cancelled out during the model estimation.

ConvSCCS Assumptions (2) and (3) (see Section 2.1) are considered to be unviolated for the following reasons. Bladder cancer times and the observation period do not seem to be correlated: among 1699 cases, we observe only 52 censoring times occurring between 2 and 35 months after outcome times. We thus consider, following Farrington *and others* (2011), that the model performance should not be affected. Assumption (3) is valid when working on rare non-recurrent events (Farrington and Whitaker, 2006). Using the same outcome definition as Neumann *and others* (2012), we find 1699 cases over roughly 1.5 million patients. The construction of this outcome also constrains it to occur only once over the 4 years of observation. It considers successive bladder cancer events as multiple recordings of the same cancer, which is sensible regarding the study length. Hence, it seems reasonable to consider bladder cancer as a rare, non-recurrent event, and thus, Assumption (3), following Farrington and Whitaker (2006).

Concerning Assumption (1), we observe a small shift in the distribution of new exposures to Insulin and Others after the outcome date among the studied cases. If this shift is caused by the outcome time, it would violate the feature exogeneity Assumption (1). However, it is also a characteristic of diabetes care pathways in France: diabetic patients often begin their treatment with metformin, and then switch to another group of molecules later on if it fails to regulate their diabetes, and so on, with Insulin being one of the last options. Timestamps might also be noisy,

since diabetic patients are continuously exposed to hypoglycemic agents. As a result, most of the patients in the cohort are already exposed at the beginning of the follow-up (30% to 70% of the exposures start at beginning of the follow-up depending on the molecule). This might introduce noise in the timestamps, as we do not really know for how long patients have been exposed, and we have shown in our simulation study that ConvSCCS is sensitive to noisy timestamps (see the sensitivity analysis in Section 4). However, this problem met in the data is standard would affect any other method similarly. Despite these unavoidable problems with the data, ConvSCCS is able to detect, as explained below the stronger adverse effect of pioglitazone pointed out in (Neumann *and others*, 2012).

We selected the best hyper-parameters γ_{tv}^* and γ_{gl}^* using stratified 3-fold cross-validation, with random search. Bootstrap confidence intervals are computed with 200 bootstrap samples obtained with the parametric bootstrap on the unpenalised likelihood. We refit the model using the support of the parameters obtained with the penalised procedure before using the bootstrap. Cross-validation and 95% bootstrap confidence intervals computation took 188 seconds using a single thread of an Intel Xeon E5-2623 v3 3.00 GHz CPU.

Study in (Neumann and others*, 2012).* The exposure to pioglitazone is measured in terms of duration from the first purchase, categorised in three intervals. The exposure to other lowering drugs starts when the patient buys the drug two times in a 6-month window, setting the beginning of the exposure in the middle of the 6-month window. A multivariate Cox model to estimate the bladder cancer hazard ratios for glucose-lowering drugs (time-dependent) exposures, adjusted for age using groups of 5 years and gender, was used.

Results. The estimated relative incidences and 95% bootstrap confidence intervals for all investigated glucose-lowering drugs are represented in Figure 4. Thanks to the penalisation used in ConvSCCS, the estimated relative incidences and confidence intervals are piecewise constant on

large steps: this is particularly interesting since it allows to detect only significant variations of the relative risks.

As shown in Figure 4 we recover a strong positive association between pioglitazone and the risk of bladder cancer, which consistently increases over time from 6 to 24 months after exposure start. Since our model estimates longitudinal effects of exposures, we compare ourselves with the duration of pioglitazone use estimates in (Neumann *and others*, 2012) in the paragraph below. Our model estimated a hazard ratio of 0.89 ($[0.56, 1.29]$) for the first 6 months after exposure to pioglitazone, 1.27 ($[0.88, 1.83]$) between 6 and 14 months after pioglitazone exposure start. (Neumann *and others*, 2012) found a hazard ratio of 1.05 ($[0.82, 1.36]$) for pioglitazone exposure of less than 12 months. For exposure greater than 12 months, they estimated a hazard ratio of 1.34 ($[1.02, 1.75]$) and 1.36 ($[1.04, 1.79]$) while our model found 1.39 ($[0.95, 2.2]$) from 14 to 22 months after pioglitazone exposure start and 2.3 ($[1.17, 4.0]$) from 22 to 24 after pioglitazone exposure start. Our results regarding pioglitazone are thus overall consistent with (Neumann *and others*, 2012).

The comparison for other hypoglycemic agents hazard ratios is more difficult since Neumann *and others* (2012) does not estimate longitudinal risks for these molecules. While (Neumann *and others*, 2012) finds the other hypoglycemic agents non statistically significant, our model cancels out the effect of rosiglitazone and find the other molecules non statistically significant during most of the lags after exposure start. However, sulfonylurea and "other" have positive significant estimates from lags 9 to 11, as well as insulin from lag 0 to 5. The shape of these three curves suggests there might be some colinearity issues between these three features, since the magnitude of their relative incidence curves seems to either match or be of opposite signs and magnitude in similar lag values. Metformin seems to be non-significant overall, despite few coefficients suggesting a positive association. While these results are not a perfect match to (Neumann *and others*, 2012), they show that our model might be useful when exploring quickly large sets of molecules

with a reduced amount of data preprocessing, even when the conditions are sub-optimal (noisy timestamps, possible feature endogeneity, and feature colinearity). Indeed, in contrary to (Neumann *and others*, 2012) approach, our methodology is scalable in the number of drugs since it doesn't require the same precise preprocessing work.

5. Conclusion

In this paper, we introduced ConvSCCS, a multivariate SCCS method with a flexible risk formulation. Our approach is based on a discrete-time version of the SCCS model (Farrington, 1995), enjoying its scalability and automatic adjustment for time-independent confounders. Classical SCCS models usually require a precise prior definition of risk windows, which might be unavailable in an adverse drugs reaction screening context. Our model circumvents this problem by modelling exposures-related relative incidences with low-granularity step functions, on which we apply total-variation penalisation. ConvSCCS shows improvements in precision and computational speed compared to the state-of-the-art in moderate to high dimension. It relies on the usual SCCS assumptions: the outcomes are distributed as a Poisson process, conditionally to the longitudinal features that are assumed to be exogenous, and observation periods of the subjects should be independent from outcome times. ConvSCCS exhibits robustness to a departure from the above mentioned SCCS assumptions, as illustrated in extensive numerical experiments, but remains sensitive to a large noise level in timestamps, which can be problematic depending on the data source quality.

An other important advantage of ConvSCCS is its ability to consider exposures to multiple drugs simultaneously in the model. ConvSCCS is, therefore, a flexible tool which could be used for future ADR screening based on LODs. An application of ConvSCCS is provided on a cohort of diabetic patients studied in (Neumann *and others*, 2012), and it is able to recover the ADR detected by the authors.

6. Supplementary Material

Supplementary material for technical details and implementation notes is available online at `http://biostatistics.oxfordjournals.org`.

Acknowledgements

This research benefited from the CNAMTS-Polytechnique research partnership, and from the Data Science Initiative of Ecole Polytechnique. We thank Aurélie Bannay, Hélène Caillol, Joël Coste, Claude Gissot, Anke Neumann, Jérémie Rudant, and Alain Weill for their insights and their help with the understanding of the SNIIRAM database. We also would like to thank the research engineers who have been working on this partnership, first of all Youcef Sebiat for the code review and revision work and also Firas Ben Sassi, Prosper Burq, Xristos Giastidis, Sathiya Kumar, Daniel de Paula.

References

Alaya, M. Z., Gaiffas, S. and Guilloux, A. (2015). Learning the intensity of time events with change-points. *IEEE Transactions on Information Theory* **61**(9), 5148–5171.

Aronson, J. K. and Ferner, R. E. (2003). Joining the dots: new approach to classifying adverse drug reactions. *BMJ* **327**(7425), 1222–1225.

ATIH. Website of the Technical Hospitalization Information Agency (ATIH), `http://www.atih.sante.fr`.

Downing, N. S., Shah, N. D., Aminawung, J. A., Pease, A. M., Zeitoun, J.-D., Krumholz, H. M. and Ross, J. S. (2017, may). Postmarket Safety Events Among Novel Therapeutics Approved by the US Food and Drug Administration Between 2001 and 2010. *JAMA* **317**(18), 1854.

Farrington, C. P. (1995). Relative Incidence Estimation from Case Series for Vaccine Safety Evaluation. *Biometrics* **51**(1), 228–235.

Farrington, C. P., Anaya-Izquierdo, K., Whitaker, H. J., Hocine, M. N., Douglas, I. and Smeeth, L. (2011, jun). Self-Controlled Case Series Analysis With Event-Dependent Observation Periods. *Journal of the American Statistical Association* **106**(494), 417–426.

Farrington, C. P. and Whitaker, H. J. (2006, nov). Semiparametric analysis of case series data. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **55**(5), 553–594.

Gault, N., Castañeda-Sanabria, J., Guillo, S., Foulon, S. and Tubach, F. (2017). Self-controlled designs in pharmacoepidemiology involving electronic healthcare databases: a systematic review. *BMC medical research methodology* **17**(1), 25.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J. and Farrington, C. P. (2016). Flexible modelling of vaccine effect in self-controlled case series models. *Biometrical Journal* **58**(3), 607–622.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J. and Farrington, C. P. (2017). Spline-based self-controlled case series method. *Statistics in Medicine* **36**(19), 3022–3038.

Hawkes, A. G and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability* **11**(3), 493–503.

Hripcsak, G. and Albers, D. J. (2013, January). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* **20**(1), 117–121.

Hripcsak, G., Albers, D. J. and Perotte, A. (2011, December). Exploiting time in electronic health record correlations. *Journal of the American Medical Informatics Association* **18**(1), i109–i115.

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., van der Lei, J., Pratt, N., Norén, G. N., Li, Y.-C., Stang, P. E., Madigan, D. *and others*. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in health technology and informatics* **216**, 574–8.

Lewis, J. D, Habel, L. A., Quesenberry, C. P., Strom, B. L., Peng, T., Hedderson, M. M., Ehrlich, S. F., Mamtani, R., Bilker, W., Vaughn, D. J. *and others*. (2015). Pioglitazone use and risk of bladder cancer and other common cancers in persons with diabetes. *Jama* **314**(3), 265–277.

Moghaddass, R., Rudin, C. and Madigan, D. (2016). The Factorized Self-Controlled Case Series Method: An Approach for Estimating the Effects of Many Drugs on Many Outcomes. *Journal of Machine Learning Research* **17**, 1–24.

Neumann, A., Weill, A., Ricordeau, P., Fagot, J. P., Alla, F. and Allemand, H. (2012). Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study. *Diabetologia* **55**(7), 1953–1962.

Norén, G. N., Hopstadius, J., Bate, A., Star, K. and Edwards, I. R. (2010). Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery* **20**(3), 361–387.

Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G. and Stang, P. E. (2011). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* **19**(1), 54–60.

Ryan, P. B., Stang, P. E., Overhage, J. M., Suchard, M. A., Hartzema, A. G., DuMouchel, W., Reich, C. G., Schuemie, M. J. and Madigan, D. (2013, October). A

Comparison of the Empirical Performance of Methods for a Risk Identification System. *Drug Safety* **36**(1), 143–158.

Schuemie, M. J., Trifirò, G., Coloma, P. M., Ryan, P. B. and Madigan, D. (2016). Detecting adverse drug reactions following long-term exposure in longitudinal observational data: The exposure-adjusted self-controlled case series. *Statistical methods in medical research* **25**(6), 2577–2592.

Simpson, S. E., Madigan, D., Zorych, I., Schuemie, M. J., Ryan, P. B. and Suchard, M. A. (2013). Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics* **69**(4), 893–902.

Trifiro, G., Fourrier-Reglat, A., Sturkenboom, M. C. J. M., Díaz Acedo, C., Van Der Lei, J. and EU-ADR Group. (2009). The EU-ADR project: preliminary results and perspective. *Studies in health technology and informatics* **148**, 43–9.

Tuppin, P., de Roquefeuil, L., Weill, A., Ricordeau, P. and Merlière, Y. (2010). French national health insurance information system and the permanent beneficiaries sample. *Revue d'Épidémiologie et de Santé Publique* **58**(4), 286–290.

Xiao, L. and Zhang, T. (2014). A Proximal Stochastic Gradient Method with Progressive Variance Reduction. *SIAM Journal on Optimization* **24**, 2057–2075.

Xu, S., Zhang, L., Nelson, J. C., Zeng, C., Mullooly, J., Mcclure, D. and Glanz, J. (2011). Identifying optimal risk windows for self-controlled case series studies of vaccine safety. *Statistics in Medicine* **30**(7), 742–752.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, series B* **68**, 49–67.

Algorithm	Regularised	Multiple features	Multiple exposures	Flexible effect
MSCCS	yes	yes	yes	no
ESCCS	no	no	accumulated	yes
SmoothSCCS	yes	no	no	yes
NonParaSCCS	yes	no	no	yes
ConvSCCS	yes	yes	yes	yes

Table 1. Comparison of SCCS methods with ConvSCCS. MSCCS is introduced in (Simpson *and others*, 2013), ESCCS in (Schuemie *and others*, 2016), while SmoothSCCS and NonParaSCCS are respectively introduced in (Ghebremichael-Weldeselassie *and others*, 2016, 2017). Regularised models are constrained to avoid overfitting, the constraint being controlled by hyper-parameters. The models can either fit multiple features at a time or be limited to study only one feature at a time. We do not consider SmoothSCCS and NonParaSCCS as able to study multiple features properly since only one feature can be regularised.

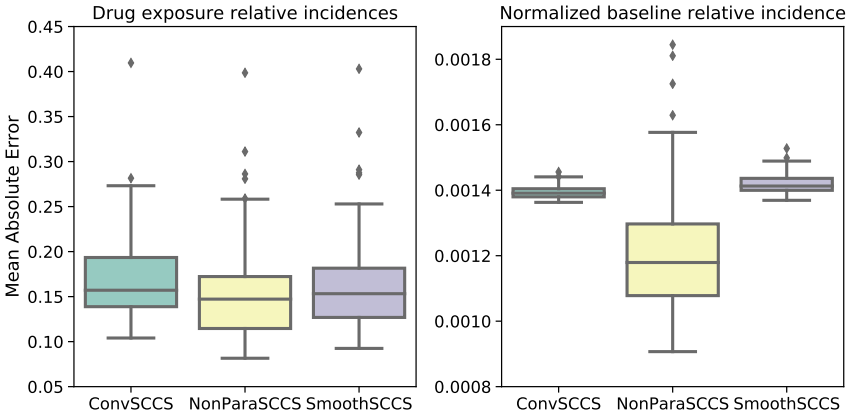

Fig. 1. Simulations results using Set 1 or risk profiles (see Figure 2) with $m = 4000$. The boxplots represent the distribution of mean absolute error as defined in Section 4.1, computed over 100 simulated populations. *Left:* MAE distribution of the drug exposure relative incidences. *Right:* MAE distribution of the baseline relative incidences, constrained so that their integral is equal to one.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

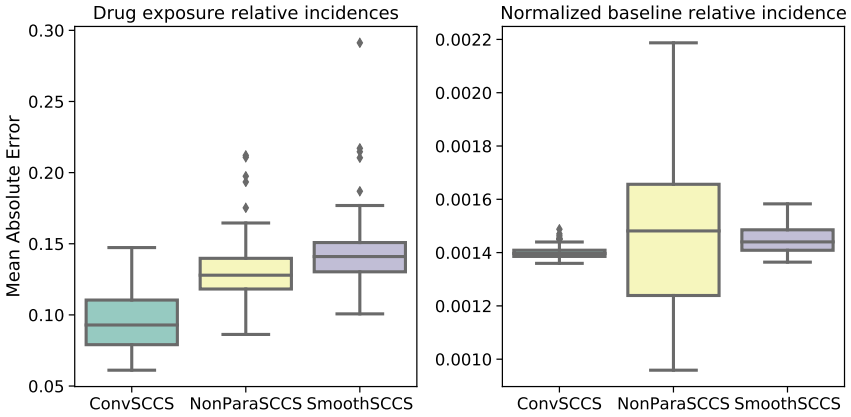

Fig. 2. Simulations results using Set 2 or risk profiles (see Figure 3) with $m = 4000$. The boxplots represent the distribution of mean absolute error as defined in Section 4.1, computed over 100 simulated populations. *Left:* MAE distribution of the drug exposure relative incidences. *Right:* MAE distribution of the baseline relative incidences, constrained so that their integral is equal to one.

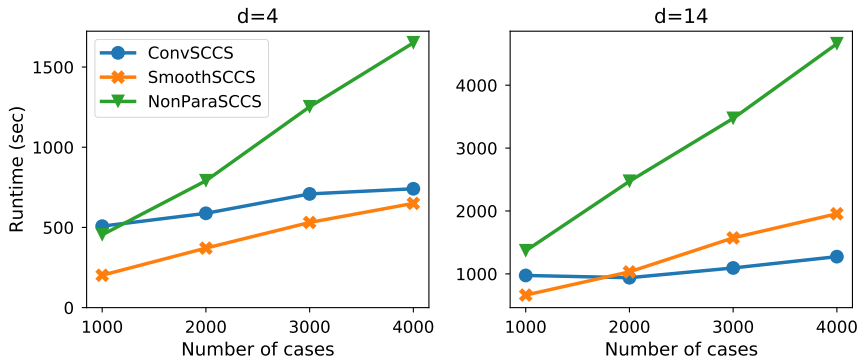

Fig. 3. Run times of ConvSCCS, SmoothSCCS and NonParaSCCS described Section 4 for 1000, 2000, 3000, 4000 cases. *Left:* run times on 4 features. *Left:* run times on 14 features. As SmoothSCCS and NonParaSCCS can only handle one feature at a time, we report the time required to fit them on each studied feature while ConvSCCS is fitted on all the features simultaneously. For each model, a fit includes cross-validation of the hyper-parameters and estimation of confidence bands. Confidence bands of ConvSCCS are estimated using parametric bootstrap, with 200 bootstrap samples.

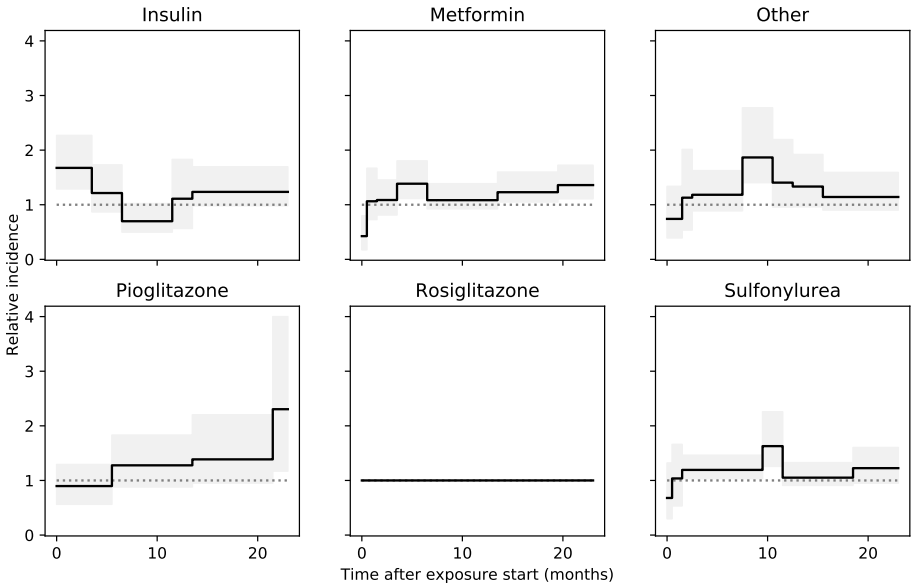

Fig. 4. Estimated relative incidences of glucose lowering drugs on the risk of bladder cancer. Black curves represent the estimated relative incidences $k = 0, \ldots, 23$ months after the beginning of exposure. Gray bands represent 95% confidence intervals estimated by the parametric bootstrap, with 200 bootstrap samples.