Biostatistics (2018), **0**, 0, *pp.* 1–12
doi:10.1093/biostatistics/SuppMat

Supplementary material for "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 Université d'Évry Val d'Essonne, 91037 Évry, France*

[5] *Caisse Nationale de l'Assurance Maladie, 75986 Paris Cedex 20, France*

maryan.morel@polytechnique.edu

1. Likelihood in SCCS models

From (Daley D.J., 2003), the Poisson likelihood of a single patient i can then be written as

$$L_i(n_i; t_i|X_i) = e^{-\int_{a_i}^{b_i} \lambda_i(s,X_i)ds} \prod_{k=1}^{n_i} \lambda_i(t_{ik}, X_i),$$

and the total number of events $n_i = N_i([a_i, b_i])$ follows a Poisson distribution

$$\mathbb{P}(n_i|X_i) = \frac{(\int_{a_i}^{b_i} \lambda_i(s, X_i)ds)^{n_i}}{n_i!} e^{-\int_{a_i}^{b_i} \lambda_i(s,X_i)ds}.$$

*To whom correspondence should be addressed.

Conditioning the likelihood by the total number of events and on the covariates histories leads to the SCCS likelihood of a patient history

$$\begin{aligned} L_i(t_i|n_i, X_i) &= \frac{L_i(n_i; t_i|X_i)}{\mathbb{P}(n_i|X_i)} \\ &= \frac{e^{-\int_{a_i}^{b_i} \lambda_i(s,X_i)ds} \prod_{k=1}^{n_i} \lambda_i(t_{ik}, X_i)}{e^{-\int_{a_i}^{b_i} \lambda_i(s,X_i)ds} \frac{(\int_{a_i}^{b_i} \lambda_i(s,X_i)ds)^{n_i}}{n_i!}} \\ &= n_i! \prod_{k=1}^{n_i} \frac{\lambda_i(t_{ik}, X_i)}{\int_{a_i}^{b_i} \lambda_i(s, X_i)ds}, \end{aligned}$$

where we used the convention $\prod_{k=1}^{0} \ldots = 1$ (i.e., the likelihood is equal to 1 if a patient does not have any event, namely $n_i = 0$). The likelihood of m patients can therefore be expressed, up to constants independent on the intensities, as

$$L \propto \prod_{i=1}^{m} \prod_{k=1}^{n_i} \frac{\lambda_i(t_{ik}, X_i)}{\int_{a_i}^{b_i} \lambda_i(s, X_i)ds}.$$

2. Discrete time SCCS

We assume that, for $i = 1, \ldots, m$, the intensity $\lambda(t, X_i(t))$ is constant over time intervals $I_k = (t_k, t_{k+1}]$, $k = 1, \ldots, K$ that form a partition of the observation interval $(a, b]$. We choose I_k to be of constant length, chosen without loss of generality equal to 1. In practice, we use the smallest granularity allowed by data. We therefore can assume that $(a_i, b_i] \cap I_k$ is either $\emptyset$ of I_k for all $i = 1, \ldots, m$, and $k = 1, \ldots, K$, which means that the observation period of each individual is a union of intervals I_k. The discrete-time likelihood writes

$$\begin{aligned} L(t_i; n_i|X_i) &= \exp\Big(\sum_{k=1}^{K} \int_{I_k} \log(\lambda(s, X_i(s)))\mathrm{d}N_i(s) - \sum_{k=1}^{K} \int_{I_k} \lambda(s, X_i(s))\mathrm{d}s\Big) \\ &= \exp\Big(\sum_{k=1}^{K} \log(\lambda_{ik}) N_i(I_k) - \sum_{k=1}^{K} \lambda_{ik}\Big), \end{aligned}$$

where $\lambda_{i,k}$ is the value of $\lambda(t, X_i(t))$ for $t \in I_k$, where $N_i(I_k) = \int_{I_k} dN_i(t)$ and where we used $\int_{I_k} dt = 1$ and the fact that $N_i(I_k) = 0$ and $\lambda_{ik} = 0$ if $I_k \cap (a_i, b_i] = \emptyset$. The distribution of the total number of events for patient i is given by

$$\mathbb{P}(n_i|X_i) = \frac{\big(\int_{a_i}^{b_i} \lambda(s, X_i(s))\mathrm{d}s\big)^{n_i}}{n_i!} e^{-\int_{a_i}^{b_i} \lambda(s,X(s))\mathrm{d}s} = \frac{\big(\sum_{k=1}^{K} \lambda_{ik}\big)^{n_i}}{n_i!} e^{-\sum_{k=1}^{K} \lambda_{ik}},$$

which leads to

$$\begin{aligned} L(t_i|n_i, X_i) = \frac{L(n_i; t_i|X_i)}{\mathbb{P}(n_i|X_i)} &= \frac{\exp\big(\sum_{k=1}^{K} \log(\lambda_{ik}) N_i(I_k) - \sum_{k=1}^{K} \lambda_{ik}\big)}{\frac{(\sum_{k=1}^{K} \lambda_{ik})^{n_i}}{n_i!} e^{-\sum_{k=1}^{K} \lambda_{ik}}} \\ &= n_i! \prod_{k=1}^{K} \Big(\frac{\lambda_{ik}}{\sum_{k'=1}^{K} \lambda_{ik'}}\Big)^{N_i(I_k)}, \end{aligned}$$

where we use the convention $0^0 = 1$, i.e. only the exposition period $(a_i, b_i]$ contributes to the likelihood, and since once again $N_i(I_k) = \lambda_{ik} = 0$ whenever $I_k \cap (a_i, b_i] = \emptyset$. Then, defining $y_{ik} := N_i(I_k)$, the previous equation can be rewritten as

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

3. Numerical implementation

We use the state-of-the-art SVRG algorithm from (Xiao and Zhang, 2014) for the minimization of our penalized negative log-likelihoof. It is known to typically lead to faster convergence than quasi-newton algorithms, such as L-BFGS-B, see (Liu and Nocedal, 1989), while allowing to deal with non-smooth objectives. Solving (3.7) requires to compute the proximal operator (see (Bach *and others*, 2012) for a definition) of $\mathrm{pen}(\theta)$. This can be done very fast numerically: $\mathrm{pen}(\theta)$ can be separated into two separate proximal operators for total-variation and group-Lasso, see (Zhou *and others*, 2012). The proximal operator of group-Lasso is explicit and given by group soft-thresholding, see (Bach *and others*, 2012), while the prox of total-variation is not, but can be computed very efficiently with the fast algorithm from (Condat, 2013).

4. Software

Our model is implemented in the `Tick` library, see (Bacry *and others*, 2017), which is a `Python` library focused on statistical learning for time dependent systems. It is open-source and available at `https://github.com/X-DataInitiative/tick`. The implementation is done in `C++`, with a Python API, and is thoroughly documented at `https://x-datainitiative.github.io/tick/`. The code used to run the experiments described in this paper is available in GitHub, at `https://github.com/MaryanMorel/ConvSCCS`.

5. Simulations details

About the simulation of longitudinal features. Let us give some details on the way we simulated longitudinal features using Hawkes processes.

Namely, we simulate dates of purchases $\{t_i^j\}_{i \geqslant 1}$, of drugs $j = 1, \ldots, d$ using a multivariate Hawkes process $N_t = [N_t^1 \cdots N_t^d]$, for $t \geqslant 0$, where $N_t^j = \sum_{k \geqslant 1} \mathbf{1}_{t_k^j \leqslant t}$ for any $t \geqslant 0$. The process N_t is a multivariate counting process, whose components N^j have intensities

$$\lambda_t^j = \mu_j + \sum_{j'=1}^{d} \sum_{k \geqslant 1} A_{j,j'} \alpha \exp(-\alpha(t - t_k^{j'})) \tag{5.1}$$

for $j = 1, \ldots, d$. This corresponds to a Hawkes process with so-called *exponential kernels.* The $\mu_j \geqslant 0$ are called *baselines* intensities, and correspond to the exogenous probability of being exposed to drug j. In the matrix $A = [A_{j,j'}]_{1 \leqslant j,j' \leqslant d}$, called the *adjacency matrix*, the entry $A_{j,j'} \geqslant 0$ quantifies the impact of past exposures to drug j' on the exposition intensity to drug j and $\alpha > 0$ is a memory parameter. A single matrix A is simulated for the whole population, but a new one is generated in each round of simulation. Recalling that the simulated events t_i^j correspond to the purchase date of drugs (this is the only information available in the LOD described in Section 4.2 below), we consider that a patient is exposed to a drug j at time t_1^j.

We sample μ_j using a uniform distribution on $[0, 5 \times 10^{-3}]$, which will produce unbalanced

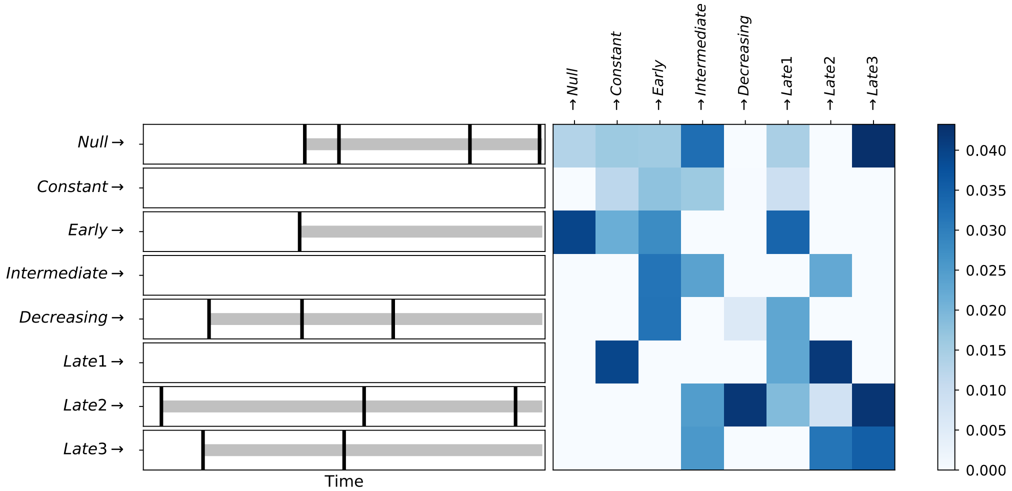

Fig. 1. *Left*: example of simulated dates of drugs purchases (vertical black lines). Exposure starts at the date of the first purchase (gray horizontal lines). *Right*: an example of generated adjacency matrix A for longitudinal feature simulation using the Hawkes process. This matrix encodes the correlation structure of exposures to drugs. To ease the reading, this figure represents the transposed adjacency matrix $A^\top$. For example, a purchase of a 'null' drug increases the probability of purchasing a 'Late3' drug. In *Left* and *Right* we simulate potential exposures to 8 drugs, each of them have a different risk profile (named "null", "constant", "early", etc.). These profiles are described in Section 5 of the supplementary material.

exposures in the simulations, and set $\alpha = 0.5$. The diagonal entries $A_{j,j}$ are equal to μ_j, and we sample q non diagonal entries using a uniform distribution $[0, 5 \times 10^{-3}]$, while setting all other entries to zero. We set $d = 4,\ q = 8$ in the first experiment, $d = 14,\ q = 24$ in the second experiment. We normalize A so that its largest singular value is 0.1, in order to ensure that the process does not generate too much events. Simulation is achieved through the thinning technique, see (Ogata, 1981), and easily achieved using the `tick` library ((`https://x-datainitiative.github.io/tick/`)), see (Bacry *and others*, 2017). An example of simulated matrix A is illustrated in Figure 1. Our simulation setup allows to generate realistic exposures, since it can reproduce the following phenomena that are typically observed in LODs:

- Depending on the drug, a patient using it has a higher probability to use it again in the future: this is quantified by the value of the diagonal entries $A_{j,j}$;

- Some drugs are often purchased at the same time, because of the underlying medical treatment: a patient using drug j' has a higher probability to use drug j, which is quantified by $A_{j,j'}$;

- Most of the patients use only a subset of all available drugs during their observation period, so several entries of A are zeros.

About the risk profiles. We provide below a precise description of the two sets of risks profiles considered in our simulations.

- Set 1 of risk profiles corresponds to the ones used in (Ghebremichael-Weldeselassie *and others*, 2017), and are represented in Figure 2. We use a lower order of magnitude than (Ghebremichael-Weldeselassie *and others*, 2017), resulting in risk profiles with maximum between 1.5 and 2 matching the magnitudes encountered in our application. The first risk profile is unimodal, the second has a constant effect, two others are continuously decreasing. In this set, risk profiles length matches $p = 50$.
- Set 2 of risk profiles represent effects described in (Aronson and Ferner, 2003): rapid, early, intermediate, late and delayed effect, see Figure 3, with magnitudes similar to Set 1. It contains the four shapes from Set 1, and a null risk, a unimodal risk with a sharp drop and three shapes of continuously increasing risks. This set contains risk profiles for which the optimal risk period is smaller than $p = 50$. We generate 7 features with "Null" risk profile, and one feature for each other risk profile, resulting in 14 features.

Following (Ghebremichael-Weldeselassie *and others*, 2017), we use for all patients a baseline relative incidence given by $\phi(t) \propto 8\sin(.01t) + 9$ (see the right-hand side of Figure 2) which can be thought as the effect of age whenever each patient has the same age.

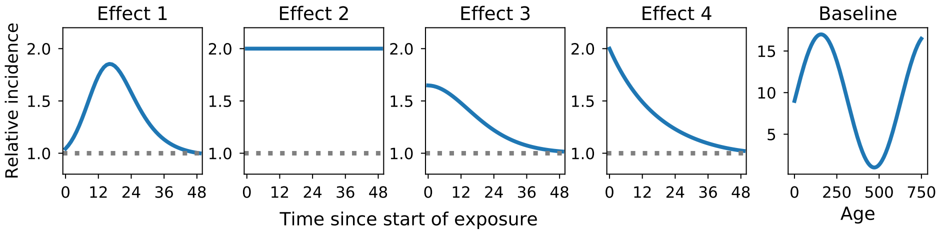

Fig. 2. *Left.* Set 1 of relative risk profiles. The effect of these relative incidences starts with the exposure, and last 50 time periods. The effect on the individual risk is multiplicative. *Right.* Temporal baseline used in all simulations.

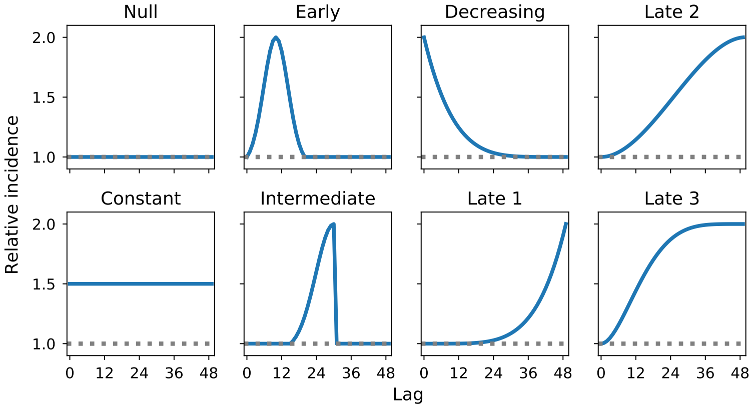

Fig. 3. Set 2 of relative risk profiles. The effect of these relative incidences starts with the exposure, and last at most 50 time periods. The effect on the individual risk is multiplicative. Note that we include 7 features with the "null" risk profile in addition to one feature with each other risk profile in Set 2, to simulate datasets in which there are irrelevant features.

Regarding the sensitivity analysis We consider three scenarii for the perturbations:

1. *Not-at-random missing data.* We simulate a hidden feature correlated to other longitudinal features using a Hawkes process. For each simulated timestamp of this feature, patients' data is masked for a time period of length drawn uniformly in $[0, max\_length]$. Outcomes are simulated using the non-perturbed data, while exposures provided to the model are computed using the censored timestamps. We vary the max_length parameter to assess the sensitivity of ConvSCCS to this perturbation.

2. *Noisy timestamps* We add a random noise draw uniformly in $[0, max\_length]$ to the fea-

tures. We vary the *max_length* parameter to assess the sensitivity of ConvSCCS to this perturbation.

3. *Missing longitudinal feature.* We simulate more features. Outcomes are simulated taking these features into account, while they are not used when fitting the model. In a first scenario, we vary the number of hidden features at constant relative incidence magnitude. In a second scenario, we vary the the relative incidence magnitude of two hidden features.

All these experiments were performed using 2000 simulated cases.

About the performance measure. As defined in Section 3.1, relative incidence of drug j, k periods after exposure start is defined as $\hat{r}^j_k = \exp(\hat{\theta}^j_k)$, $k = 0, \dots, p$ in our model. In (Ghebremichael-Weldeselassie *and others*, 2016, 2017), the estimated relative incidence is defined as $\hat{r}^j_k = \hat{\theta}^j(k) > 0$ for $k = 0, \dots, p$, see Equation (2.3). Denoting the ground truth relative incidence r^*, the MAE is given by

$$MAE = \frac{1}{dK} \sum_{j=1}^{d} \sum_{k=1}^{K} |r^{j*}_k - \hat{r}^j_k|.$$

Since we assume that all the patients are affected by the baseline in the same way, its order of magnitude cannot be properly estimated by the models. In order to be able to compare baseline relative risks, we constrain their integral to be equal to one as (Ghebremichael-Weldeselassie *and others*, 2017).

Regarding the sensitivity analysis results.

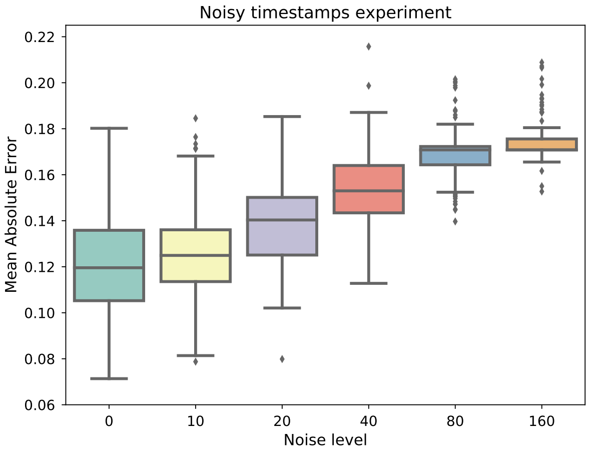

Fig. 4. Sensitivity analysis adding a noise drawn uniformly in $[0, noise\_level]$ to features timestamps using Set 2 of risk profiles (see Figure 3) with $m = 2000$. The boxplots represent the distribution of mean absolute error as defined in Section 4.1, computed over 100 simulated populations.

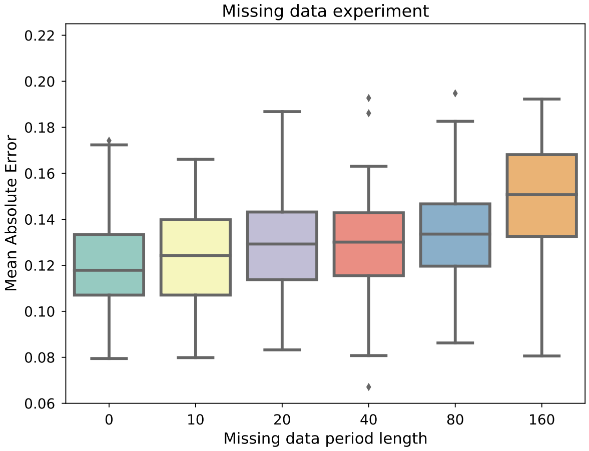

Fig. 5. Sensitivity analysis simulating not-at-random missing data. A hidden feature timestamps are simulated in the same way as regular features. At each time of this feature, other features data is masked for a period of *missing data period length*. Other features are simulated using Set 2 or risk profiles (see Figure 3) with $m = 2000$. The boxplots represent the distribution of mean absolute error as defined in Section 4.1, computed over 100 simulated populations.

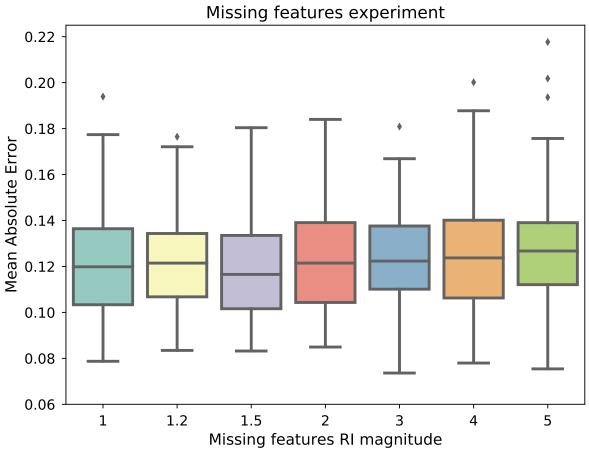

Fig. 6. Sensitivity analysis simulating missing longitudinal features. Simulations results using Set 2 or risk profiles plus two hidden features (see Figure 3) with $m = 2000$. The order of magnitude of hidden features relative incidence vary from 1 to 5. The boxplots represent the distribution of mean absolute error as defined in Section 4.1 computed over 100 simulated populations.

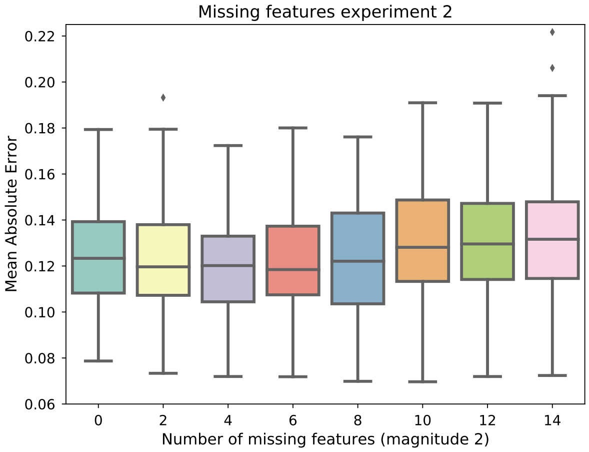

Fig. 7. Sensitivity analysis simulating missing longitudinal features. Simulations results using Set 2 or risk profiles plus hidden features using similar risk profiles (see Figure 3) with $m = 2000$. The number of hidden features vary from 0 to 14. The boxplots represent the distribution of mean absolute error as defined in Section 4.1 computed over 100 simulated populations.

Characteristics	**Overall study population**
N	1,428,637
Men	771,647
Bladder cancers	1,699
Age (years)	
40-44	54,989
45-49	94,986
50-54	160,388
55-59	238,611
60-64	238,394
65-69	223,721
70-74	232,100
75-79	185,448
Number of patients exposed to glucose-lowering drugs *(a patient can appear in several lines)*	
Insulin	343,912
Other OHA	434,352
Rosiglitazone	157,346
Metformin	1,043,967
Pioglitazone	158,619
Sulfonylurea	836,572
Number of patients exposed to a single glucose-lowering drug *(each patient appears at most in a single line)*	
Insulin	102,021
Other OHA	34,927
Rosiglitazone	2,239
Metformin	208,331
Pioglitazone	4,486
Sulfonylurea	145,509

Table 1. Demographics and glucose-lowering drug use of the cohort of French diabetic patients covered by the general insurance scheme (i.e., in the SNIIRAM database), aged 40-79 years and followed from 2006 to 2009.

Regarding the studier cohort structure.

References

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

Bach, F., Jenatton, R., Mairal, J., Obozinski, G. *and others*. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning* **4**(1), 1–106.

Bacry, E., Bompaire, M., Gaïffas, S. and Poulsen, S. (2017, jul). tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling.

Condat, L. (2013, nov). A Direct Algorithm for 1-D Total Variation Denoising. *IEEE Signal Processing Letters* **20**(11), 1054–1057.

Daley D.J., Vere-Jones D. (2003). *An introduction to the theory of Point Processes - Vol. 1: Elementary theory and methods*.

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.

Liu, D. C. and Nocedal, J. (1989, December). On the limited memory bfgs method for large scale optimization. *Math. Program.* **45**(3), 503–528.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory* **27**(1), 23–31.

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

Zhou, J., Liu, J., Narayan, V. A. and Ye, J. (2012). Modeling disease progression via fused sparse group lasso. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. pp. 1095–1103.

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