Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm

Marta Avalos, Hélène Pouyes, Yves Grandvalet, Ludivine Orriols, Emmanuel Lagarde

Avalos et al. BMC Bioinformatics 2015, 16(Suppl 6):S1
http://www.biomedcentral.com/1471-2105/16/S6/S1

BMC Bioinformatics

RESEARCH Open Access

Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm

Marta Avalos[1,2,3*], Hélène Pouyes[2,4], Yves Grandvalet[5], Ludivine Orriols[1,2], Emmanuel Lagarde[1,2]

From 10th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics Nice, France. 20-22 June 2013

Abstract

This paper considers the problem of estimation and variable selection for large high-dimensional data (high number of predictors p and large sample size N, without excluding the possibility that $N < p$) resulting from an individually matched case-control study. We develop a simple algorithm for the adaptation of the Lasso and related methods to the conditional logistic regression model. Our proposal relies on the simplification of the calculations involved in the likelihood function. Then, the proposed algorithm iteratively solves reweighted Lasso problems using cyclical coordinate descent, computed along a regularization path. This method can handle large problems and deal with sparse features efficiently. We discuss benefits and drawbacks with respect to the existing available implementations. We also illustrate the interest and use of these techniques on a pharmacoepidemiological study of medication use and traffic safety.

Background

Epidemiological case-control studies are used to identify factors that may contribute to a health event by comparing a group of cases, that is, people with the health event under investigation, with a group of controls who do not have the health event but who are believed to be similar in other respects. Logistic regression is the most important statistical method in epidemiology to analyze data arising from a case-control study. It allows to account for the potential confounders (factors independently associated with both the health outcome and the risk factors of main interest) and, if the logistic model is correct, to eliminate their effect.

Cases and controls are sometimes matched: every case is matched with a preset number of controls who share a similar exposure to these matching factors, to ensure that controls and cases are similar in variables that are related to the variable under study but are not of interest by themselves [1]. Matching is useful when the distributions of the confounders differs radically between the unmatched comparison groups. In these situations, the weight of confounding factors is so important that a simple adjustment does not guarantee a straightforward interpretation of results. The case-crossover design, in which each subject serves as his own control, is a particular matched case-control design [2,3]. The association between event onset and risk factors is estimated by comparing exposure during the period of time just prior to the event onset (case period) to the same subject's exposure during one or more control periods. This design inherently removes the confounding effects of time-invariant factors while it is still sensitive to the effects of time-varying risk factors [4,5]. The conditional logistic regression model is the standard tool for the analysis of matched case-control and case-crossover studies.

Big and/or high-dimensional data arise nowadays in many diverse fields of epidemiologic research such as registry-based epidemiology. An advantage of having a large sample size from registry data is the ability to study rare exposures, outcomes, or subgroups in a population large enough to provide sufficient precision [6]. However, the analysis of these studies has to be addressed using

* Correspondence: marta.avalos@isped.u-bordeaux2.fr
[1]Univ. Bordeaux, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France
Full list of author information is available at the end of the article

methods that appropriately account for their statistical and computational complexity. In what concerns matched case-control studies, matching by a high-dimensional propensity or stratification score are increasingly popular approaches to deal with high-dimensional confounding in epidemiological studies investigating effects of a treatment or exposure [7-10]. In these studies, a score is built from independent observations and then used to match data. Standard estimation methods accounting for data dependence due to matching, such as maximum-conditional likelihood are then applied and variable selection is performed by conventional selection procedures. Finding optimal subsets becomes an essential problem [11,12]. When high-dimensionality is related to the risk factors of main interest instead the potential confounders, regularization methods, such as the Lasso (*least absolute shrinkage and selection operator*) [13], have emerged as a convenient approach, but remain unfamiliar to most epidemiologists [14].

Our first implementation of the Lasso to conditional logistic regression was based on the correspondence between the conditional likelihood of conditional logistic regression and the partial likelihood of stratified, discrete-time Cox proportional hazards model (where cases are defined as events and controls are censored) [15,16].

This allowed the analysis of the pharmacoepidemiological case-crossover data of prescription drugs and driving described in Orriols and colleagues [17] for the older driver population. The same algorithm was independently proposed in two other high-dimensional matched case-control studies to identify association between

- Crohn's disease and genetic markers in family-based designs (such as case-sibling and case-parent) [18];
- specific brain regions of acute infarction and hospital acquired pneumonia in stroke patients [19].

Here, we describe a more efficient algorithm, directly targeting the optimization of the conditional likelihood. This algorithm is based on the IRLS (*iteratively reweighted least squares*) algorithm [20], which is widely used for estimating generalized linear models, and which can be applied with Lasso-type penalties. This line of approach was used in

- a matched case-control study to identify association between DNA methylation levels and hepatocellular carcinoma in tumor-adjacent non-tumor tissues [21];
- the case-crossover study of prescription drugs and driving for the whole population [22].

In these two studies, as well as in [23], the algorithms are based on the penalized IRLS, solved as a weighted Lasso problem *via* cyclical coordinate descent [24,25]. However, they differ in their actual implementations, a major benefit of our proposal relying in the simplification of the calculations involved in the likelihood function. The resulting gain in efficiency allows for the processing of large-scale datasets [22]. We detail these calculations here and illustrate their interest on the pharmacoepidemiological study that originally motivated these algorithmical developments.

Methods

Conditional likelihood

We are interested in the relationship between a binary outcome Y and several risk factors $U = (U_1, \ldots, U_p)$. We assume that subjects are grouped into N strata (corresponding to matched sets), consisting in one case ($Y_{in} = 1$) and M controls ($Y_{ln} = 0$, $l \neq i$) each one having a U value: For subject i of stratum n, the vector of observations is $\mathbf{u}_{in} = (u_{in}1, \ldots, u_{inp})$, $i = 0, 1, \ldots, M$, $n = 1, \ldots, N$.

Denote by $\pi_{in} = \pi(\mathbf{u}_{in})$ the probability of event for the i-th subject of the n-th stratum. We model the dependence of the probability of disease on the risk factors values via the logistic model, supposing that each stratum has its own baseline odds of disease, which may differ across strata:

$$\text{logit}(\pi_{in}) = \alpha_n + \mathbf{u}_{in}\beta, \tag{1}$$

where α_n are coefficients representing the global effect of matching factors on the response; and coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ express the log odds ratios corresponding to the risk factors. When the differences among strata are not relevant (in the sense that matching factors are potential confounders, but are not the potential risk factors of interest), we just need to estimate $\boldsymbol{\beta}$. Therefore, strata-specific parameters α_n are eliminated from the likelihood by conditioning on the fact that exactly one subject in every matched case-control set is a case. Consider the n-th stratum, the unconditional probability of observing the occurrence of the event only in the i-th subject is:

$$\pi_{in} \times \prod_{l \neq i} (1 - \pi_{\ln}) = \frac{\pi_{in}}{1 - \pi_{in}} \prod_{l=0}^{M} (1 - \pi_{ln}). \tag{2}$$

Under the logistic model, the conditional probability that within a matched set, the assignment of the $M + 1$ values is given by:

$$\frac{\frac{\pi_{in}}{1-\pi_{in}} \prod_{l=0}^{M} (1 - \pi_{ln})}{\sum_{l=0}^{M} \frac{\pi_{ln}}{1-\pi_{ln}} \prod_{i=0}^{M} (1 - \pi_{in})} = \frac{\frac{\pi_{in}}{1-\pi_{in}}}{\sum_{l=0}^{M} \frac{\pi_{ln}}{1-\pi_{ln}}} = \frac{e^{\mathbf{u}_{in}\beta}}{\sum_{l=0}^{M} e^{\mathbf{u}_{ln}\beta}}. \tag{3}$$

We use the convention that all cases are indexed by $i = 0$ and all controls are indexed by $i \in \{1, \ldots, M\}$

($Y_{0n} = 1$ and $Y_{in} = 0$, $i \neq 0$). When $M = 1$ (that is 1:1 matching), the likelihood evaluated at $\boldsymbol{\beta}$ simplifies to:

$$L(\beta, D) = \prod_{n=1}^{N} \frac{e^{\mathbf{u}_{0n}\beta}}{e^{\mathbf{u}_{0n}\beta} + e^{\mathbf{u}_{1n}\beta}} = \prod_{n=1}^{N} \frac{1}{1 + e^{-(\mathbf{u}_{0n} - \mathbf{u}_{1n})\beta}} = \prod_{n=1}^{N} \frac{1}{1 + e^{-\mathbf{x}_n\beta}}, \quad (4)$$

where $\mathbf{x}_n = \mathbf{u}_{0n} - \mathbf{u}_{1n}$, $D = \{(\mathbf{x}_n, 1)\}_{n = 1,\ldots,N}$. Thus, in a 1:1 matched case-control design, the conditional likelihood is identical to the unconditional likelihood of the binary logistic model with $\mathbf{x}_n$ as covariates, no intercept, and a constant response equal to 1. When $M > 1$ (that is 1:M matching), it will be useful in the algorithms introduced below to rewrite the likelihood function:

$$L(\beta, D) = \prod_{n=1}^{N} \frac{e^{\mathbf{u}_{0n}\beta}}{e^{\mathbf{u}_{0n}\beta} + \sum_{l=1}^{M} e^{\mathbf{u}_{ln}\beta}}, \quad (5)$$

also in terms of differences:

$$L(\beta, D) = \prod_{n=1}^{N} \frac{1}{1 + \sum_{l=1}^{M} e^{-(\mathrm{u}_{0n} - \mathrm{u}_{ln})\beta}} = \prod_{n=1}^{N} \frac{1}{1 + \sum_{l=1}^{M} e^{-\mathrm{x}_{ln}\beta}}, \quad (6)$$

where $\mathbf{x}_{in} = \mathbf{u}_{0n} - \mathbf{u}_{in}$, $D = \{(\mathbf{x}_{in}, 1)\}_{n = 1,\ldots,N;\, i = 1,\ldots,M}$.

Usually, the parameters $\boldsymbol{\beta}$ are estimated by maximizing the conditional log-likelihood function $\log(L(\boldsymbol{\beta}, D))$. However, maximum likelihood analysis may lead to an inflated variance and/or a biased estimation of log odds ratios in studies with a small number of strata or several unbalanced risk factors, especially when the number of covariates is large or the proportions are close to zero. Different methods have been developed, generally in low-dimensional settings, to correct bias while controlling for variance [26-31]. In moderate to high-dimensional settings, penalized methods such as the Lasso [13] have been proposed to reduce variance (to improve prediction accuracy) and to identify the subset of exposures that exhibit the strongest associations with the response [15,18,16]. The Lasso applied to conditional logistic regression consists in maximizing the conditional log-likelihood function penalized by the L^1 norm of the unknown coefficient vector, or equivalently, minimizing the negative objective function:

$$\min_{\beta}(-\log(L(\beta, D)) + \lambda||\beta||_1), \quad (7)$$

where λ is a regularization parameter, and $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ is the L^1-norm of coefficients. Then, the parameter estimator is:

$$\hat{\beta}_D^{\mathrm{L}}(\lambda) = \arg\min_{\beta}\Big(\sum_{n=1}^{N} \log\Big(1 + \sum_{l=1}^{M} e^{-\mathrm{x}_{ln}\beta}\Big) + \lambda||\beta||_1\Big) \quad (8)$$

Algorithms

The conditional likelihood in (6) derived for the conditional logistic model corresponds to the partial likelihood of the stratified, discrete-time Cox proportional hazards model. Standard survival data analysis software can be used for the analysis of a 1:M matched case-control study. Analogously, algorithms proposed for solving the Lasso for the Cox model allowing for stratification can be used.

For the particular design consisting in one case and one control, we may apply a penalized unconditional logistic regression. Indeed, as showed above, the conditional likelihood function simplifies dramatically resulting in the likelihood function for the unconditional logistic regression without intercept term applied to the differences of the predictors. The L^1 penalized logistic regression problem (7) is convex but not differentiable. This characteristic leads to greater difficulty in solving the optimization problem. There has been very active development on numerical algorithms. An extensive, although not exhaustive, review and comparison of existing methods can be found in [32,33].

The IRLS algorithm [20] uses a quadratic approximation for the average logistic loss function, which is consequently solved by a L^1 penalized least squares solver. This method is particularly easy to implement since it takes advantage of existing algorithms for the Lasso linear regression. We revisit here the IRLS algorithm proposed for solving the L^1 logistic model by [13,34,35]. These methods differ basically in the algorithm applied to resolve the Lasso linear step. The last authors applied the Lars-Lasso algorithm [36] to find a Newton direction at each step and then used a backtracking line search to minimize the objective value. They also provided convergence results. Essentially, we applied this proposal to the particular objective function arisen in 1:1 matching, but replacing the Lars algorithm by the cyclic coordinate descent algorithm [24,25]. We generalize then this approach to estimate the conditional logistic likelihood coefficients in 1:M matching (6). Sparsity-related works of other research areas have also explored the use and properties of IRLS [37-39].

IRLS-cyclic coordinate descent for the 1:1 matching

Denote by $f(\beta, D, \lambda) = -\log(L(\beta, D)) + \lambda||\beta||_1 = \sum_{n=1}^{N} \log(1 + e^{-\mathrm{x}_n\beta}) + \lambda||\beta||_1$ the objective function in (7) and (8) with $M = 1$. In particular, the unpenalized objective function $f(\boldsymbol{\beta}, D, 0)$ is noted $f(\boldsymbol{\beta})$. Let $\mathbf{X}$ be the $N \times p$ matrix of the observed differences $x_{nj} = u_{0nj} - u_{1nj}$, $n = 1, \ldots, N$, $j = 1, \ldots, p$; let $\mathbf{g}(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ be the gradient and Hessian of the unpenalized objective function:

$$\mathbf{g}(\beta) = -\mathbf{X}^t\left(\frac{e^{-\mathbf{x}_1\beta}}{1 + e^{-\mathbf{x}_1\beta}}, \ \ldots, \ \frac{e^{-\mathbf{x}_N\beta}}{1 + e^{-\mathbf{x}_N\beta}}\right)^t, \quad (9)$$
$$\mathbf{H}(\beta) = \mathbf{X}^t\mathbf{W}(\beta)\mathbf{X},$$

where

$$\mathbf{W}(\beta) = \text{diag}\left(\frac{e^{-\mathbf{x}_1\beta}}{(1+e^{-\mathbf{x}_1\beta})^2}, \ldots, \frac{e^{-\mathbf{x}_N\beta}}{(1+e^{-\mathbf{x}_N\beta})^2}\right), \quad (10)$$

is the matrix of weights. The Newton method consists in finding a step direction by computing the optimum $\boldsymbol{\gamma}^{[k]}$ of the quadratic approximation at $\boldsymbol{\beta}^{[k]}$ (the current point in the k-th iteration) as:

$$\gamma^{[k]} = \beta^{[k]} - \mathbf{H}^{-1}(\beta^{[k]})\mathbf{g}(\beta^{[k]}) \quad (11)$$

The next iterate is then computed using the step direction by a line search over the step size parameter t:

$$\beta^{[k+1]} = (1-t)\beta^{[k]} + t\gamma^{[k]}, \quad (12)$$

Algorithm 1 IRLS-cyclic coordinate descent algorithm for the 1:1 matching

1: Fix $\mathbf{X}_{N\times p}$, $\lambda \geq 0$, $\boldsymbol{\beta}$, $0 \leq \alpha_1 \leq 0.5$, $0 < \alpha_2 < 1$ and $\tau > 0$.
2: **while** the stopping criterion is not satisfied **do**
3: Compute $\mathbf{W}^{[k]}$ and $\mathbf{z}^{[k]}$ using (10) and (13).
4: Resolve (16) applying cyclic coordinate descent. Let $\boldsymbol{\gamma}^{[k]}$ be the solution.
5: Backtracking line-search:
Initialize $t = 1$, set $\Delta\boldsymbol{\beta}^{[k]} = \boldsymbol{\gamma}^{[k]} - \boldsymbol{\beta}^{[k]}$.
6: **while** the stopping criterion is not satisfied **do**
7: $\boldsymbol{\gamma}^{[k]} = \boldsymbol{\beta}^{[k]} + t\Delta\boldsymbol{\beta}^{[k]}$
8: Check the stopping criterion: $f(\boldsymbol{\gamma}^{[k]}) \leq f(\boldsymbol{\beta}^{[k]}) + \alpha_1 t\, \mathbf{g}(\boldsymbol{\beta}^{[k]})^t\Delta\boldsymbol{\beta}^{[k]}$
9: **if** The stopping criterion not satified **then**
10: $t \leftarrow \alpha_2 t$
11: **end if**
12: **end while**
13: Compute $\boldsymbol{\beta}^{[k+1]} = (1-t)\boldsymbol{\beta}^{[k]} + t\boldsymbol{\gamma}^{[k]}$.
14: Check the stopping criterion: $\frac{|f(\beta^{[k+1]}) - f(\beta^{[k]})|}{|f(\beta^{[k+1]})|} \leq \tau$
15: **end while**

where $0 < t \leq 1$. Let $\mathbf{z}$, the working response vector, be defined as:

$$(\mathbf{z}^{[k]})_n = \mathbf{x}_n\beta^{[k]} + (\mathbf{W}^{[k]})_n^{-1}\frac{e^{-\mathbf{x}_n\beta^{[k]}}}{1+e^{-\mathbf{x}_n\beta^{[k]}}} = \mathbf{x}_n\beta^{[k]} + (1+e^{-\mathbf{x}_n\beta^{[k]}}). \quad (13)$$

Then the gradient, Hessian and the step direction in (11) can be reformulated as follows:

$$\begin{aligned}\mathbf{g}(\beta^{[k]}) &= -\mathbf{X}^t\mathbf{W}^{[k]}(\mathbf{z}^{[k]} - \mathbf{X}\beta^{[k]}),\\ \mathrm{H}(\beta^{[k]}) &= \mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X},\\ \gamma^{[k]} &= (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{z}^{[k]}.\end{aligned} \quad (14)$$

Thus $\boldsymbol{\gamma}^{[k]}$ is the solution to the weighted least squares problem:

$$\gamma^{[k]} = \arg\min_{\gamma} ||(\mathbf{W}^{[k]})^{1/2}(\mathbf{z}^{[k]} - \mathbf{X}\gamma)||_2^2. \quad (15)$$

Applying the same development to the penalized problem, we obtain that $\boldsymbol{\gamma}^{[k]}$ is the solution to the penalized weighted least squares problem:

$$\gamma^{[k]} = \arg\min_{\gamma} ||(\mathbf{W}^{[k]})^{1/2}(\mathbf{z}^{[k]} - \mathbf{X}\gamma)||_2^2 + \lambda||\gamma||_1.$$

Generalization to the 1:M matching

As previously, the IRLS algorithm is applied to resolve (7)-(8). It iterates the following steps until convergence: first, for the current $\boldsymbol{\beta}$, update the matrix of weights and the working response vector, then, compute the vector that minimizes penalized weighted least squares problem, using cyclic coordinate descent; finally, perform a line search to determine the step size to update $\boldsymbol{\beta}$. The objective function in (7) is now $f(\beta, D, \lambda) = -\log(L(\beta, D)) + \lambda||\beta||_1 = \sum_{n=1}^{N}\log(1+\sum_{l=1}^{M} e^{-\mathbf{x}_{ln}\beta}) + \lambda||\beta||_1$

Algorithm 2 IRLS-cyclic coordinate descent algorithm for the 1:M matching

1: Fix $\mathbf{X}_{N\,M\times p}$, $\lambda \geq 0$, $\boldsymbol{\beta}^0$, $0 \leq \alpha_1 \leq 0.5$, $0 < \alpha_2 < 1$ and $\tau > 0$.
2: **while** the stopping criterion is not satisfied **do**
3: Compute $\mathbf{W}^{[k]}$ and $\mathbf{z}^{[k]}$ using (18).
4: Resolve (19) applying cyclic coordinate descent. Let $\boldsymbol{\gamma}^{[k]}$ be the solution.
5: Backtracking line-search:
Initialize $t = 1$, set $\Delta\boldsymbol{\beta}^{[k]} = \boldsymbol{\gamma}^{[k]} - \boldsymbol{\beta}^{[k]}$.
6: **while** the stopping criterion is not satisfied **do**
7: $\boldsymbol{\gamma}^{[k]} = \boldsymbol{\beta}^{[k]} + t\Delta\boldsymbol{\beta}^{[k]}$
8: Check the stopping criterion: $f(\boldsymbol{\gamma}^{[k]}) \leq f(\boldsymbol{\beta}^{[k]}) + \alpha_1 t\, \mathbf{g}(\boldsymbol{\beta}^{[k]})^t\Delta\boldsymbol{\beta}^{[k]}$
9: **if** The stopping criterion not satified **then**
10: $t \leftarrow \alpha_2 t$
11: **end if**
12: **end while**
13: Compute $\boldsymbol{\beta}^{[k+1]} = (1-t)\boldsymbol{\beta}^{[k]} + t\boldsymbol{\gamma}^{[k]}$.
14: Check the stopping criterion: $\frac{|f(\beta^{[k+1]}) - f(\beta^{[k]})|}{|f(\beta^{[k+1]})|} \leq \tau$
15: **end while**

Let $\mathbf{X}$ be the matrix of the observed differences $x_{inj} = u_{0nj} - u_{inj}$, $i = 1, \ldots, M$, $n = 1, \ldots, N$, $j = 1, \ldots, p$, now with $M\,N$ rows and p columns. The gradient $\mathbf{g}(\boldsymbol{\beta})$ and Hessian $\mathbf{H}(\boldsymbol{\beta})$ of the unpenalized objective function have now the form:

$$\begin{aligned}g(\beta) &= -\mathbf{X}^t\left(\frac{e^{-\mathbf{x}_{11}\beta}}{1+\sum_{l=1}^{M}e^{-\mathbf{x}_{l1}\beta}}, \ldots, \frac{e^{-\mathbf{x}_{M1}\beta}}{1+\sum_{l=1}^{M}e^{-\mathbf{x}_{l1}\beta}}, \ldots, \frac{e^{-\mathbf{x}_{1N}\beta}}{1+\sum_{l=1}^{M}e^{-\mathbf{x}_{lN}\beta}}, \ldots, \frac{e^{-\mathbf{x}_{MN}\beta}}{1+\sum_{l=1}^{M}e^{-\mathbf{x}_{lN}\beta}}\right)^t,\\ \mathbf{H}(\beta) &= \mathbf{X}^t\mathbf{W}(\beta)\mathbf{X},\end{aligned} \quad (17)$$

with the matrix of weights and the working response vector written now as:

$$\mathbf{W}(\beta) = \mathrm{diag}\left(\frac{e^{-\mathbf{x}_{11}\beta}}{\left(1+\sum_{l=1}^{M} e^{-\mathbf{x}_{11}\beta}\right)^2}, \ldots, \frac{e^{-\mathbf{x}_{1N}\beta}}{\left(1+\sum_{l=1}^{M} e^{-\mathbf{x}_{1N}\beta}\right)^2}, \ldots, \frac{e^{-\mathbf{x}_{N1}\beta}}{\left(1+\sum_{l=1}^{M} e^{-\mathbf{x}_{N1}\beta}\right)^2}, \ldots, \frac{e^{-\mathbf{x}_{NN}\beta}}{\left(1+\sum_{l=1}^{M} e^{-\mathbf{x}_{NN}\beta}\right)^2}\right),$$
$$(\mathbf{z}^{[k]})_{in} = \mathbf{x}_{in}\beta^{[k]} + (\mathbf{W}^{[k]})^{-1}_{in}\frac{e^{-\mathbf{x}_{in}\beta^{[k]}}}{1+\sum_{l=1}^{M} e^{-\mathbf{x}_{in}\beta^{[k]}}} = \mathbf{x}_{in}\beta^{[k]} + (1+\sum_{l=1}^{M} e^{-\mathbf{x}_{in}\beta^{[k]}}). \quad (18)$$

With this $N\,M \times N\,M$ matrix of weights and working response vector of length $N\,M$, $\boldsymbol{\gamma}^{[k]}$ is the solution to the penalized weighted least squares problem:

$$\gamma^{[k]} = \underset{\gamma}{\mathrm{argmin}}\, ||(\mathbf{W}^{[k]})^{1/2}(\mathbf{z}^{[k]} - \mathbf{X}\gamma)||_2^2 + \lambda||\gamma||_1. \quad (19)$$

Notice that when using the likelihood function in (6) as a function of $\mathbf{x}_{in}$, $i = 1, \ldots, M$, $n = 1, \ldots, N$, the matrix of weights is diagonal while, when using it as a function of $\mathbf{u}_{in}$, $i = 0, 1, \ldots, M$, $n = 1, \ldots, N$, $\mathbf{W}(\boldsymbol{\beta})$ is nondiagonal, which complicates the matrix inversion problem in terms of computation.

The regularization path

For a given value of λ, a certain number of predictors with non-zero regression coefficients are obtained by minimizing the L^1-penalized negative log-likelihood. In general, the smaller λ, the more the penalty is relaxed, and the more predictors are selected. Inversely, the higher λ, the more predictors are eliminated. The regularization path is the continuous trace of the Lasso estimates of the regression coefficients obtained when varying λ from 0 (the maximum-likelihood solution for the full conditional logistic model) to a certain threshold, which depends on data, beyond which no predictors are retained in the model. In general, the amount of penalization on the L^1-norm of the coefficients is chosen by computing, first, the regularization path of the solution to (7), as the regularization parameter varies. Then, the value of λ is estimated from a grid of values using an appropriate criterion. Unlike L^1-regularization paths for linear models, paths for logistic models are not piecewise linear, approximate regularization paths should then be considered [36,40-42]. To construct the grid of λ-values, $\lambda_{\max} > \ldots > \lambda_{\min}$, firstly, we calculate $\lambda_{\max}$ and $\lambda_{\min}$, the smallest λ for which all coefficients are zero and the smallest λ for which the algorithm converge without numerical problems, respectively.

It can be shown that, if $\lambda > \max_{j\in\{1,\ldots,p\}} |\frac{\partial f}{\partial \beta_j}(0)|$ then the directional derivatives of the $\lambda\|\beta\|_1$ term at $\boldsymbol{\beta} = \mathbf{0}$ dominate and so $\boldsymbol{\beta} = \mathbf{0}$ is the minimizer of $f(\boldsymbol{\beta}, D, \lambda)$ [41,42]. The evaluation of the gradient function $\mathbf{g}(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \mathbf{0}$ leads to $\lambda_{\max} = \max_{j\in\{1,\ldots,p\}} |\frac{1}{M+1}\sum_{n=1}^{n}\sum_{l=1}^{M} x_{lnj}|$. We fix $\lambda_{\min} = \in \lambda_{\max}$.

Next, we generate T values equally spaced (on the linear or log scale) decreasing from $\lambda_{\max}$ to $\lambda_{\min}$. For $\lambda_1 = \lambda_{\max}$, the initial vector of coefficients is set to $\boldsymbol{\beta}0 = \mathbf{0}$. For each λ_t, $1 < t \leq T$, the initial vector of coefficients is set to $\beta^{[0]} = \hat{\beta}(\lambda_{t-1})$, i.e. the coefficient vector at convergence for the precedent λ value.

After this discretization, the optimal regularization parameter can be chosen by a model selection criterion such as cross-validation or the Bayesian Information Criterion (BIC) [43,44].

Publicly available implementation

Several algorithms have been proposed for solving the Lasso for the Cox model [45,42,46-50]. Among those proposing a publicly available code, only the method proposed by Goeman [49] allows for stratification (implementation publicly available through penalized R-package). However, as discussed in [16], this implementation is not applicable to large datasets.

Among the efficient algorithms solving the Lasso for the logistic model, those proposed by [42] (consisting in a generalization of the Lars-Lasso algorithm described in [36]), [49] (based on a combination of gradient descent and Newton's methods), and [24] (based on a quadratic approximation followed by a cyclic coordinate descent method) have a publicly available R implementation (glmpath, penalized and glmnet packages, respectively). The glmpath package [42] did not accommodate models without intercept. The penalized package [49] allows for several practical options. In particular, a no-intercept Lasso logistic regression model can be fitted using the differences as independent variables and a constant response. However, though the Newton method has fast convergence, forming and solving the underlying Newton systems require excessive amounts of memory for large-scale problems. This package is optimized for situations with many covariates, but does not handle a large amount of observations. Finally, the glmnet package can deal efficiently with very large (sparse) matrices and has been shown to be faster than competing methods [33,24]. However, the logistic function of this package fails to converge when a constant response is used in the logistic model without intercept. A summary is presented in Table 1.

Parallel to our work, Sun et al. [21] and Reid et al. [23] proposed an IRLS-cyclic coordinate descent algorithm to resolve the Lasso for (un)conditional logistic regression. While all these works rely on IRLS-cyclic coordinate descent, the objectives and the strategies implemented to address these objectives are different. The algorithm developed in [21] was implemented into the pclogit R package which can be downloaded at http://www.columbia.edu/~sw2206/. This algorithm is based on the optimization of the original unsimplified likelihood function (5) with a penalty that encourages the grouping encoded by a given network graph. The algorithm developed in [23] was implemented into the clogitL1 R package which can be downloaded from

Table 1. Main publicly available R packages that solves the Lasso and other sparse penalties for the Cox, logistic or conditional logistic models (surveyed October 1st, 2014)

Package	1:1 matching?	1:M matching?	Amenable to processing of with grouping penalties	with large *N*	K:M matching?
Logistic Model					
glmpath [42]	NO	NO	NO	NO	NO
penalized [49]	YES	NO	NO	NO	NO
glmnet [24]	NO	NO	NO	NO	NO
Cox Model					
glmpath [42]	NO	NO	NO	NO	NO
penalized [49]	YES	YES	NO	NO	NO
glmnet [50]	NO	NO	NO	NO	NO
Conditional Logistic					
pclogit [21]	YES	YES	YES	NO	NO
clogitL1 [23]	YES	YES	NO	NO	YES
clogitLasso [43,51]	YES	YES	NO	YES	NO

CRAN. The main concern of the authors is the extension to matched sets consisting in more than one case (denote K the number of cases per stratum) and $M > 1$ controls, with large K and large M . In this situation, the conditional likelihood function has a more complicated form, and the authors apply a recursive formula to compute the likelihood and its derivatives exactly. This scheme results in involving intensive computations that are not amenable to the processing of large datasets.

Our R package clogitLasso is available at our institution's Web page, in the "links and downloads" menu, http://www.isped.u-bordeaux.fr/biostat. Two strategies are implemented. The first one, discussed in [16], is dedicated to small to moderate sample sizes. It is based on the stratified discrete-time Cox proportional hazards model and depending on the penalized package [49]. The second one, discussed in the present paper, is amenable to the processing of large datasets (large N). It directly targets the conditional logistic regression problem, relying on the lassoshooting package [25] for the application of cyclic coordinate descent. The lassoshooting package is particularly well adapted for large-scale problems and provides a no-intercept option. For example, large sparse data matrices (resulting from rare exposures), can be stored in a sparse format as well as the diagonal matrix of weights and working response vector. The Lasso solver lassoshooting proceeds with $\mathbf{X}^t\mathbf{W}\mathbf{X}$ and $\mathbf{X}^t\mathbf{W}\mathbf{z}$ (of dimension $p \times p$ and $p \times 1$, respectively) instead of $\mathbf{W}^{1/2}\mathbf{X}$ and $\mathbf{W}^{1/2}\mathbf{z}$ (of dimension $N\,M \times p$ and $N\,M \times 1$, respectively). Other practical options are available, for example penalized and unpenalized (always included in the model) variables can be specified. The methods, model selection criteria and capabilities of clogitLasso are detailed in [51,44].

Results

Medicinal drugs have a potential effect on the skills needed for driving, a task that involves a wide range of cognitive, perceptual and psychomotor activities. Nevertheless, disentangling their impact on road traffic crashes is a complex issue from a pharmacoepidemiological point of view because, between others, the large variety of pharmaceutical classes. A major approach relies on the use of population-registries data, such as those conducted in UK [52,53], Norway [54], France [17] or Finland [55]. We report the use of the algorithms detailed in this paper for exploratory analysis of the large pharmacoepidemiological data of prescription drugs and driving described in Orriols and colleagues [17].

Data sources and designs

Information on drug prescriptions and road traffic accidents was obtained from the following anonymized population-based registries: the national health care insurance database (which covers the whole French population and includes data on reimbursed prescription drugs), police reports, and the national police database of injurious road traffic crashes. Drivers involved in an injurious crash in France, between July 2005 and May 2008, were included in the study.

Traffic crash data including information about alcohol impairment and drivers' responsibility for the crash were collected. When the breath test is negative (concentration < 0.5 g/L), the driver is recorded as not being under the influence of alcohol. Responsibility was determined by a standardized method that assigns a score to each driver on the basis of factors likely to reduce driver responsibility (such as road, vehicle and driving conditions, type of accident, difficulty of the task involved, and traffic rule obedience, including alcohol consumption).

We consider all dispensed and reimbursed medicines to the drivers in the study, in the 6 months before the crash, coded by the WHO ATC (World Health Organization Anatomical Therapeutic Chemical) classification fourth level system. For each drug, the exposure period

started one day after dispensing and the length of the exposure period was estimated from median values reported within a survey on drug prescription in France. This leaded to about 400 candidate binary predictors (exposure, coded 1, and unexposure, coded 0, to each medicinal drug).

The objective is to identify the relevant associations between the exposure to medicinal drugs and the risk of being responsible for an injurious non-alcohol related road traffic crash. We considered two study designs, each one addressing a different epidemiological question.

Individually matched case-control study

The epidemiological question is:

"What is different about at-fault drivers, if they are highly comparable to not at-fault drivers on external factors that may influence a road crash such as weather or road conditions?"

The purpose of the analysis is to compare exposure to medicinal drugs probabilities on the day of crash between at-fault drivers (cases) and not at-fault drivers (controls). Thus, we matched each case to one control (1:1 matching) on the basis of the date, hour and location of the crash. We also adjusted for potential confounders: age, sex and long-term chronic diseases, that is, factors that have shown to be associated with the risk of accident and may confound the impact of medicinal drugs on responsibility. These factors were forced in the models (unpenalized). Age was coded by using a discrete qualitative variable with seven categories: 18-24, 25-44, 45-64, 65-70, 70-75, 75-80, ≥80 and then using dummy variables. Long-term diseases were defined by an administrative status in the French health care insurance database allowing full reimbursement of health care expenses related to 30 chronic diseases. Chronic disease was coded by using a binary variable: presence or absence of any fully reimbursed chronic diseases.

Of the 58,700 drivers with a negative alcohol test in the analytic database, 26,568 (45%) were considered responsible for their crash. After matching, 6,857 case-control pairs were highly comparable in terms of external factors, among them, 3,381 matched pairs showed different medicinal drug exposure (for at least one drug) on the day of the crash. Figure 1 shows the flowchart summarizing the selection of the subjects of the database. After eliminating medicines that have been little or not consumed (by less than 10 subjects), we get 189 binary predictors (in addition to the factors forced in the models).

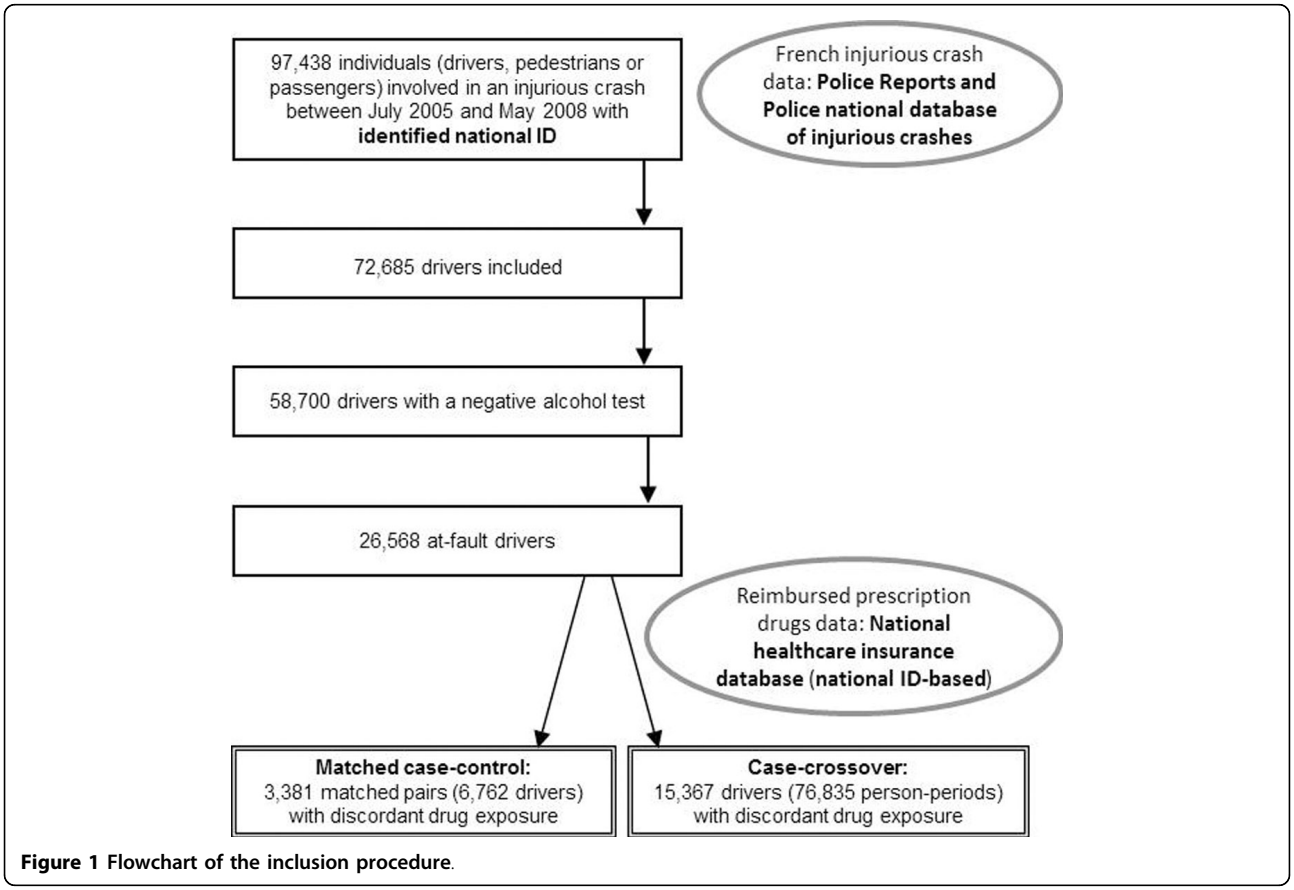

Figure 1 Flowchart of the inclusion procedure.

Case-crossover study

The epidemiological question is:

"What is different about the day of the crash for at-fault drivers?"

The purpose of the analysis is to compare exposure to medicinal drugs probabilities on the day of crash (case period) and on the day one, two, three and four months prior to the crash date (control periods) for each at-fault driver. Thus, we matched each case period to four control periods (1:4 matching). Each period was separated from the next one by one month, the maximal duration of a treatment dispensed at the pharmacy in France, to avoid any residual effect of an exposure in one period on the following one.

Of the 26,568 at-fault drivers with a negative alcohol test in the analytic database, 15,367 (58%) showed different medicinal drug exposure (for at least one of the five periods; for at least one drug). Thus, 76,835 person-periods contributed to the estimation according to the flowchart in Figure 1, and they were described using the 189 binary predictors already mentioned above.

Pharmacoepidemiological results

Although some drugs are usually prescribed together and correlation problems are possible, we observed only mild correlation, probably because the large sample size, then only the L^1 penalty was applied. We used 10-fold likelihood-based cross- validation to estimate the penalization parameter. Figure 2 shows the Lasso regularization path as a function of λ for the paired case-control study. Since all participants were involved in an accident, positive effects ($\hat{\beta}_j > 0$) have not a direct interpretation as protective factors. Thus, they are not displayed.

Only four medicinal drugs were simultaneously selected as showing relevant associations with the risk of being at-fault for an injurious non-alcohol related traffic crash in both design studies (that is independently of the study design): Carboxamide derivative antiepileptics (N03AF), Benzodiazepine derivatives (N05CD), Other hypnotics and sedatives (N05CX), Antidepressants (N06AX). Odds ratio estimates are presented in table 2. The confounding underlying health conditions are not well controlled in our matched case-control study, however the case-crossover design inherently removes the confounding effects of time-invariant factors such as chronic health conditions. Thus, drug adversities may explain these results instead of chronic health-related complications. The effects of these four drugs on driving are well documented in the literature. In addition these medicines contain warning messages in relation to impairing driving ability. From a prevention perspective, it would be important to identify more precisely which populations are concerned by this at risk behavior. Further analyses should also be necessary to elucidate why these drugs appear to be related to an increased risk of at-fault crashes while other drugs from the same class do not. Such differences can simply be explained by a higher consumption, but other hypotheses are plausible.

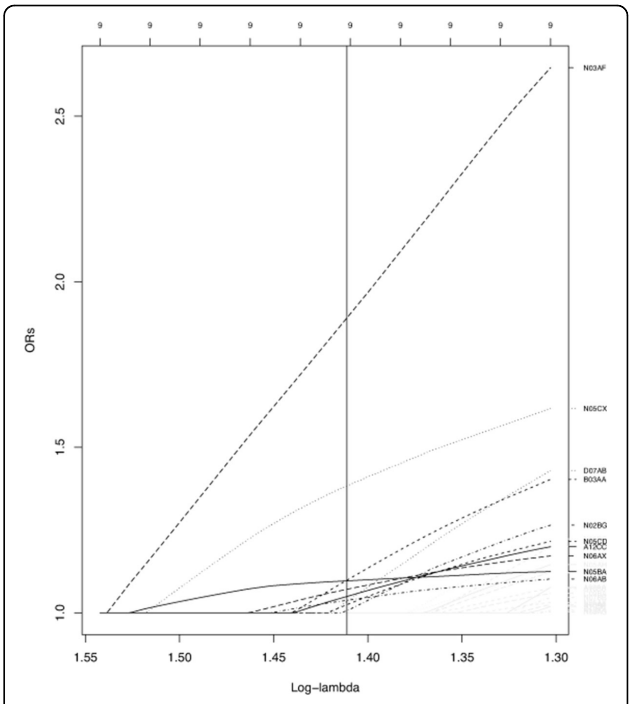

Figure 2 Individually matched case-control study. The Lasso regularization path as a function of λ for the paired case-control design. The black vertical line indicates the λ value optimizing the cross-validation criterion. Coefficient values of drugs selected are indicated in black, the others are in gray. Only drugs estimated to be risk factors (positive $\hat{\beta}_j$) are displayed. Potential confounders (sex, age, chronic disease), forced in the model, are omitted.

Conclusion

We have developed a simple algorithm for the adaptation of the Lasso and related methods to the conditional logistic regression model. Our proposal relies on the simplification of the calculations involved in the likelihood function and the IRLS algorithm, that iteratively solves reweighted Lasso problems using cyclical coordinate descent, computed along a regularization path. As a result, this algorithm can handle large problems and deal with sparse features efficiently.

Problems related to high-dimensionality arise nowadays in many fields of epidemiological research (genetic, environmental or pharmacoepidemiology, for instance). In particular, we illustrate the interest of this methodology on the pharmacoepidemiological study of prescription drugs and driving that originally motivated these algorithmical developments.

The use of Lasso-related techniques is justified in this context as follows. First, regression models, with

Table 2. Odds ratio (OR) by study design

ATC class second level	ATC class fourth level	Case-crossover	Matched case-control
Drugs for acid related disorders	A02BA	1.88	
	A02BX	1.19	
Drugs for functional gastrointestinal disorders	A03FA	1.24	
Laxatives	A06AD	1.37	
Mineral supplements	A12CC		1.10
	A12AX	1.57	
Antianemic preparations	B03AA		1.20
	B03BB	1.24	
Peripheral vasodilators	C04AX	1.15	
Antifungals for dermatological use	D01AE	1.13	
Corticosteroids	D07AB		1.16
Sex hormones and modulators of the genital system	G03CA	1.20	
Muscle relaxants	M03BX	1.23	
Analgesics	N02BG		1.09
Antiepileptics	N03AA	2.93	
	N03AF	1.34	2.11
	N03AX	1.19	
Psycholeptics	N05BA		1.11
	N05CD	1.37	1.09
	N05CX	1.01	1.46
Psychoanaleptics	N06AB		1.06
	N06AX	1.05	1.11
Drugs for obstructive airway diseases	R03BB	1.23	
Cough and cold preparations	R05DA	1.08	
Antihistamines for systemic use	R06AX	1.06	

Odds ratio estimates are displayed only for selected risk factors.

straightforward interpretation, are the most important statistical techniques used in analytical epidemiology. Thus, these techniques appear to be a good compromise between traditional and data-driven approaches since modeling is based on standard regression models, rather than a black-box. Second, controlling for potential confounding is a critical point in epidemiology, thus multivariate modeling approaches are preferable to separate univariate tests. Third, it is expected that only few drugs will be truly associated with the risk of being involved in a road traffic crash, thus sparsity-inducing penalties seem to be appropriate. It is also expected that most of these relevant drugs will have a weakly strength of association, however, only predictors with effect sizes above the noise level can be detected using Lasso-related techniques. Nevertheless, this limitation is shared by any model selection method [56-58].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MA developed the algorithms and revised the R code, performed the analysis on the datasets and wrote the Manuscript. HP developed the R code. YG revised the algorithms, helped conduct the statistical/machine learning/bioinformatics literature review and revised the Manuscript. LO helped collect the data, performed the analysis on the datasets, interpreted the results of the analysis and conducted the epidemiological literature review. EL designed and supervised the epidemiological research, collected the data, and interpreted the results of the analysis. All authors read and approved the final manuscript.

Acknowledgements
We thank the CESIR research group for its collaborative support: Fabienne Bazin (INSERM U657), Sylvie Blazejewski (CIC 0005, Bordeaux), Anne Castot (AFSSAPS), Bernard Delorme (AFSSAPS), Geneviève Durrieu (Service de pharmacologie médicale et clinique, CHU Toulouse), Blandine Gadegbeku (Univ. de Lyon, IFSTTAR), Pierre-Olivier Girodet (CIC 0005, Bordeaux), Marcel Goldberg (INSERM U687-UVSQ), Bernard Laumon (Univ. de Lyon, IFSTTAR), Dominique Lauque (CHU Toulouse), Nathalie Lecoules (CHU Toulouse), Laurence Memes (CIC 0005, Bordeaux), Louis Merle (CHU Limoges), Yvon Merlière (CNAM-TS), Jean-Louis Montastruc (Service de pharmacologie médicale et clinique, CRPV, INSERM U 1027, Univ. de Toulouse, CHU Toulouse), Nicholas Moore (INSERM U657, CIC 0005, Bordeaux), Pernelle Noize (Inserm U657), Nathalie Orsoni (CHU Limoges), Antoine Pariente (INSERM U657, CIC 0005, Bordeaux), Pierre Philip (Clinique du sommeil, CHU Bordeaux), Régis Ribéreau-Gayon (CHU Bordeaux), Louis-Rachid Salmi (INSERM U897, Univ. Bordeaux Segalen), Aurore Tricotel (AFSSAPS). We also acknowledge the French National Health Insurance (CNAM-TS), the National Interministerial Road Safety Observatory (ONISR), and Agira-TransPV for providing data related to health insurance reimbursements and road traffic accidents, as well as the Public Health Research Federative Institute (IFR 99).

Declaration statement
Publication costs for this work were funded by the last author's institution: "Injury prevention and control" team of the French Institute of Health and Medical Research INSERM U897.

This article has been published as part of *BMC Bioinformatics* Volume 16 Supplement 6, 2015: Selected articles from the 10th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/16/S6.

Authors' details
[1]Univ. Bordeaux, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France. [2]INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France. [3]SISTM team, INRIA, F-33000 Bordeaux, France. [4]Univ. de Pau et des Pays de l'Adour, F-64012 Pau, France. [5]Univ. de Technologie de Compiègne, CNRS, Heudiasyc UMR7253, F-60203 Compiègne, France.

Published: 17 April 2015

References

1. Bland JM, Altman DG: **Matching.** *BMJ* 1994, **309(6962)**:1128.
2. Maclure M: **The case-crossover design: A method for studying transient effects on the risk of acute event.** *Am J Epidemiol* 1991, **133(2)**:144-153.
3. Delaney JA, Suissa S: **The case-crossover study design in pharmacoepidemiology.** *Stat Methods Med Res* 2009, **18(1)**:53-65.
4. Mittleman MA, Maclure M, Robins JM: **Control sampling strategies for case-crossover studies: An assessment of relative efficiency.** *Am J Epidemiol* 1995, **142(1)**:91-98.
5. Janes H, Sheppard L, Lumley T: **Overlap bias in the case-crossover design, with application to air pollution exposures.** *Stat Med* 2005, **24(2)**:285-300.
6. Toh S, Platt R: **Is size the next big thing in epidemiology?** *Epidemiology* 2013, **24(3)**:349-351.
7. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA: **High-dimensional propensity score adjustment in studies of treatment effects using health care claims data.** *Epidemiology* 2009, **20(4)**:512-522.
8. Rassen JA, Schneeweiss S: **Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system.** *Pharmacoepidemiol Drug Saf* 2012, **Suppl 1:** 41-49.
9. Epstein MP, Broadaway KA, He M, Allen AS, Satten GA: **Stratification-score matching improves correction for confounding by population stratification in case-control association studies.** *Genet Epidemiol* 2012, **36(3)**:195-205.
10. Le HV, Poole C, Brookhart MA, Schoenbach VJ, Beach KJ, Layton JB, Sturmer T: **Effects of aggregation of drug and diagnostic codes on the performance of the high-dimensional propensity score algorithm: An empirical example.** *BMC Med Res Methodol* 2013, **13**:142.
11. Schroder M, Husing J, Jockel KH: **An implementation of automated individual matching for observational studies.** *Methods Inf Med* 2004, **43(5)**:516-520.
12. Austin PC: **A comparison of 12 algorithms for matching on the propensity score.** *Stat Med* 2014, **33(6)**:1057-1069.
13. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J Roy Statist Soc Ser B* 1996, **58(1)**:267-288.
14. Walter S, Tiemeier H: **Variable selection: Current practice in epidemiological studies.** *Eur J Epidemiol* 2009, **24(12)**:733-736.
15. Avalos M: **Model selection via the lasso in conditional logistic regression.** *Proceedings of the Second International Biometric Society Channel Network Conference* Ghent, Belgium; 2009.
16. Avalos M, Grandvalet Y, Duran-Adroher N, Orriols L, Lagarde E: **Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood.** *Stat Med* 2012, **31(21)**:2290-2302.
17. Orriols L, Delorme B, Gadegbeku B, Tricotel A, Contrand B, Laumon B, Salmi LR, Lagarde E: **Prescription medicines and the risk of road traffic crashes: A French registry-based study.** *PLoS Med* 2010, **7(11)**:e1000366.
18. Yu Z, Deng L: **Pseudosibship methods in the case-parents design.** *Stat Med* 2011, **30(27)**:3236-3251.
19. Qian J, Payabvash S, Kemmling A, Lev MH, Schwamm LH, Betensky RA: **Variable selection and prediction using a nested, matched case-control study: Application to hospital acquired pneumonia in stroke patients.** *Biometrics* 2014, **70(1)**:153-163.
20. Green PJ: **Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives.** *J Roy Statist Soc Ser B* 1984, **46(2)**:149-192.
21. Sun H, Wang S: **Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data.** *Stat Med* 2013, **32(21)**:2127-2139.
22. Avalos M, Orriols L, Pouyes H, Grandvalet Y, Thiessard F, Lagarde E: **Variable selection on large case-crossover data: Application to a registry-based study of prescription drugs and road-traffic crashes.** *Pharmacoepidemiology and Drug Safety* 2014, **23(2)**:140-151.
23. Reid S, Tibshirani R: **Regularization paths for conditional logistic regression: The clogitl1 package.** *Journal of Statistical Software* 2014, **58(12)**:1-23.
24. Friedman J, Hastie T, Tibshirani R: **Regularized paths for generalized linear models via coordinate descent.** *Journal of Statistical Software* 2010, **33(1)**:1-22.
25. Jörnsten R, Abenius T, Kling T, Schmidt L, Johansson E, Nordling T, Nordlander B, Sander C, Gennemark P, Funa K, Nilsson B, Lindahl L, Nelander S: **Network modeling of the transcriptional effects of copy number aberrations in glioblastoma.** *Molecular Systems Biology* 2011, **7(486)**.
26. Greenland S: **Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators.** *Biostatistics* 2000, **1(1)**:113-122.
27. Corcoran C, Mehta C, Patel N, Senchaudhuri P: **Computational tools for exact conditional logistic regression.** *Stat Med* 2001, **20(17-18)**:2723-2739.
28. Agresti A, Min Y: **Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data.** *Stat Med* 2004, **23(1)**:65-75.
29. Bartolucci F: **On the conditional logistic estimator in two-arm experimental studies with non-compliance and before-after binary outcomes.** *Stat Med* 2010, **29(13)**:1411-1429.
30. Heinze G, Puhr R: **Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets.** *Stat Med* 2010, **29(7-8)**:770-777.
31. Sun JX, Sinha S, Wang S, Maiti T: **Bias reduction in conditional logistic regression.** *Stat Med* 2011, **30(4)**:348-355.
32. Schmidt M, Fung G, Rosales R: **Fast optimization methods for L1 regularization: A comparative study and two new approaches.** *European Conference on Machine Learning (ECML)* 2007, **4107**:286-297.
33. Yuan GX, Chang KW, Hsieh C.-J, Lin CJ: **A comparison of optimization methods and software for large-scale L1-regularized linear classification.** *Journal of Machine Learning Research* 2010, **11**:3183-3234.
34. Lokhorst J: **The lasso and generalised linear models.** *Honors project* Department of Statistics, The University of Adelaide, South Australia, Australia; 1999.
35. Lee S, Lee H, Abbeel P, Ng AY: **Efficient L1-regularized logistic regression.** *Proceedings of the 21th National Conference on Artificial Intelligence (AAAI)* 2006.
36. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least angle regression.** *Ann Statist* 2004, **32(2)**:407-499.
37. Candes EJ, Wakin M, Boyd S: **Enhancing sparsity by reweighted l1 minimization.** *J Fourier Anal Appl* 2008, **14(5)**:877-905.
38. Daubechies I, DeVore R, Fornasier M, Gunturk C: **Iteratively reweighted least squares minimization for sparse recovery.** *Comm Pure Appl Math* 2010, **63(1)**:1-38.
39. Wipf DP, Nagarajan S: **Iterative reweighted l1 and l2 methods for finding sparse solutions.** *IEEE Journal of Selected Topics in Signal Processing (Special Issue on Compressive Sensing)* 2010, **4**:317-329.
40. Rosset S, Zhu J: **Piecewise linear regularized solution paths.** *Ann Statist* 2007, **35(3)**:1012-1030.
41. Keerthi SS, Shevade S: **A fast tracking algorithm for generalized lars/lasso.** *IEEE Transactions on Neural Networks* 2007, **18(6)**:1826-1830.
42. Park MY, Hastie T: **l_1-regularization path algorithm for generalized linear models.** *J Roy Statist Soc Ser B* 2007, **69(4)**:659-677.
43. Avalos M, Pouyes H: **clogitLasso: An R package for L1 penalized estimation of conditional logistic regression models.** *Proceedings of the 1ères Rencontres R (in French)* Bordeaux, France; 2012, , **:** 99-100.
44. Avalos M, Grandvalet Y, Pouyes H, Orriols L, Lagarde E: **High-dimensional sparse matched case-control and case-crossover data: A review of recent works, description of an R tool and an illustration of the use in epidemiological studies.** In *Computational Intelligence Methods for Bioinformatics and Biostatistics. Volume 8452.* Lecture Notes in Computer Science;Formenti, E., Tagliaferri, R., Wit, E 2014:109-124.
45. Gui J, Li H: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.** *Bioinformatics* 2005, **21(13)**:3001-3008.

46. Engler D, Li Y: **Survival analysis with high-dimensional covariates: An application in microarray studies.** *Statistical Applications in Genetics and Molecular Biology* 2009, **8**:14.
47. Sohn I, Kim J, Jung SH, P C: **Gradient lasso for Cox proportional hazards model.** *Bioinformatics* 2009, **25(14)**:1775-1781.
48. Wang S, Nan B, Zhou N, Zhu J: **Hierarchically penalized Cox regression with grouped variables.** *Biometrika* 2009, **96(2)**:307-322.
49. Goeman J: **l_1 penalized estimation in the Cox proportional hazards model.** *Biometrical Journal* 2010, **52(1)**:70-84.
50. Simon N, Friedman J, Hastie T, T R: **Regularization paths for Cox's proportional hazards model via coordinate descent.** *Journal of Statistical Software* 2011, **39(5)**:1-13.
51. Avalos M, Grandvalet Y, Pouyes H, Orriols L, Lagarde E: **clogitLasso: An R package for high-dimensional analysis of matched case-control and case-crossover data.** *Proceedings of the Tenth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2013)* Nice, France; 2013, **.**
52. Barbone F, McMahon AD, Davey PG, Morris AD, Reid IC, McDevitt DG, MacDonald TM: **Association of road-traffic accidents with benzodiazepine use.** *Lancet* 1998, **352(9137)**:1331-1336.
53. Gibson JE, Hubbard RB, Smith CJ, Tata LJ, Britton JR, Fogarty AW: **Use of self-controlled analytical techniques to assess the association between use of prescription medications and the risk of motor vehicle crashes.** *Am J Epidemiol* 2009, **169(6)**:761-768.
54. Engeland A, Skurtveit S, Morland J: **Risk of road traffic accidents associated with the prescription of drugs: A registry-based cohort study.** *Annals of Epidemiology* 2007, **17(8)**:597-602.
55. Karjalainen K, Blencowe T, Lillsunde P: **Substance use and social, health and safety-related factors among fatally injured drivers.** *Accid Anal Prev* 2012, **45**:731-736.
56. Candes EJ, Plan Y: **Near-ideal model selection by L1 minimization.** *Ann. Statist* 2009, **37(5A)**:2145-2177.
57. Bunea F: **Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization.** *Electron J Statist* 2008, **2**:1153-1194.
58. Bunea F, She Y, Ombao H, Gongvatana A, Devlin K, Cohen R: **Penalized least squares regression methods and applications to neuroimaging.** *Neuroimage* 2011, **55(4)**:1519-1527.

doi:10.1186/1471-2105-16-S6-S1
Cite this article as: Avalos *et al.*: Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm. *BMC Bioinformatics* 2015 **16**(Suppl 6):S1.