Doerken *et al. BMC Medical Research Methodology* (2016) 16:103
DOI 10.1186/s12874-016-0197-0

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

The case-crossover design via penalized regression

Sam Doerken[1*], Maja Mockenhaupt[2], Luigi Naldi[3], Martin Schumacher[1] and Peggy Sekula[1]

Abstract

Background: The case-crossover design is an attractive alternative to the classical case–control design which can be used to study the onset of acute events if the risk factors of interest vary in time. By comparing exposures within cases at different time periods, the case-crossover design does not rely on control subjects which can be difficult to acquire. However, using the standard method of maximum likelihood, resulting risk estimates can be heavily biased when the prevalence to risk factors is very low (or very high).

Methods: To overcome the problem of low risk factor prevalences, penalized conditional logistic regression via the lasso (least absolute shrinkage and selection operator) has been proposed in the literature as well as related methods such as the Firth correction. We apply and compare several penalized regression approaches in the context of a case-crossover analysis of the European Study of Severe Cutaneous Adverse Reactions (EuroSCAR; 1997–2001).

Results: Out of 30 drugs, standard methods only correctly classified 17 drugs (including some highly implausible risk estimates), while penalized methods correctly classified 22 drugs.

Conclusion: Penalized methods generally yield better risk classifications and much more plausible risk estimates for the EuroSCAR study than standard methods. As these novel techniques can be easily implemented using available R packages, we encourage routine use of penalized conditional logistic regression for case-crossover data.

Keywords: Case-crossover design, Lasso, Conditional logistic regression, Penalized regression, Severe cutaneous adverse reactions

Background

The case–control design is a common study design for assessing risk factors in epidemiology. However, recruiting suitable controls is a constant challenge [1]. For such settings, case series designs have been developed. One such type of case series is the case-crossover design [2] where every subject or patient serves as his own control. The association between the event of interest (e.g. disease onset) and risk factors is estimated by comparing exposure during the time period just prior to the event of interest to the same subject's exposure during a reference period. The selection bias that case–control studies suffer from due to having to select controls is thus avoided by the case-crossover design. It also removes confounding effects from any time-invariant factors (e.g. sex, ethnicity).

Still, case-crossover studies face many of the remaining challenges that also affect case–control studies. In case–control studies, risk factors with very low (or very high) prevalence are problematic, e.g. they may result in unstable or unreliable estimates. Analogously, in case-crossover studies, risk factors become problematic if they are not very time-variant, i.e. only few subjects switch between exposed and unexposed during the observation period, leading consequently to few discordances. Such data may be referred to as sparse. A second common challenge is a potentially large number of risk factors under investigation. Furthermore, collinearity of risk factors is a typical problem. Many studies suffer from a combination of these three problems. As a result, estimation using standard maximum likelihood methods may be heavily biased or may not even be possible at all.

Penalized regression methods (also known as shrinkage, regularization, or sparse regression) have become popular and been applied plentiful since they are particularly

* Correspondence: doerken@imbi.uni-freiburg.de
[1] Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany
Full list of author information is available at the end of the article

suitable to remedy the challenges of collinear or sparse data [3, 4]. They use a modified likelihood, allowing estimation in instances where maximizing the original likelihood is not numerically possible.

Conditional logistic regression (CLR) using maximum likelihood is the standard method for the analysis of case-crossover studies; Avalos et al. [5] suggested adapting penalized methods to the CLR model and applied their methods successfully to both simulated and real data; they provided further applications in Avalos et al. [6] In our article, we aim to further investigate the use of penalized regression in case-crossover studies. We will include the same methods that were evaluated in Avalos et al. [5]. We do this using data from EuroSCAR [7], a study on the very rare and severe cutaneous adverse reactions Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS/TEN) that are often drug induced.

The EuroSCAR study is attractive for this purpose for three reasons: firstly, it may potentially benefit from a penalization approach since it suffers from some of the challenges mentioned earlier. Secondly, even though EuroSCAR was a case–control study, here we only use data from cases, but are then able to compare our results to the results of the original case–control analysis, using the latter as a benchmark. And thirdly, a successor study to EuroSCAR is now running a registry of SJS/TEN patients, thereby only includes cases, which makes an assessment of case-crossover methods in this setting is highly relevant. We argue that penalized regression, due to the benefits in estimation and ease of implementation, should be encouraged for routine use for the case-crossover design.

Methods

Study design

The motivation behind the case-crossover design is as follows: if an exposure is prevalent right before the onset of an event of interest (e.g. disease onset) but absent during other times, it is natural to suspect that the exposure may be a trigger of the event. As opposed to a case–control design, the case-crossover design relies exclusively on cases in order to make inferences about risk factors. Within every case, the exposure to a time-varying risk factor during a time interval immediately prior to the event of interest, referred to as the case period, is compared to the exposure during a different, independent time window prior (or possibly also after) the event, referred to as the reference period (see Fig. 1 for an illustration). The case-crossover design is therefore most suitable for studying acute events. Since case and reference periods are compared within the same subject, potential time-invariant confounders have no influence. This is true for both known and unknown confounders. The design is attractive because it spares the difficulty of having to recruit suitable controls.

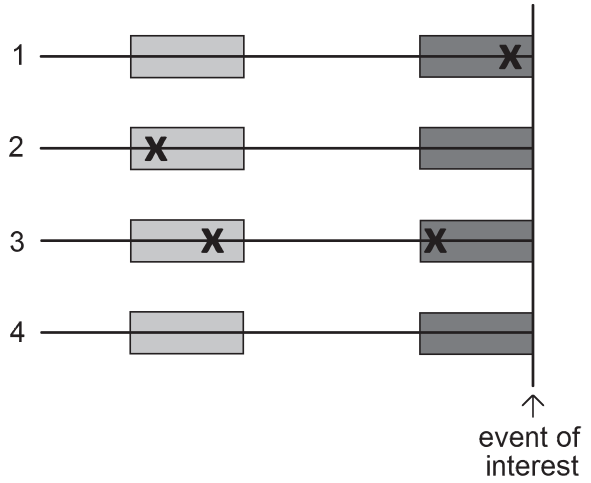

Fig. 1 Illustration of the timeline of the case-crossover design, shown for four hypothetical subjects with respect to one risk factor. The *dark gray* area just prior to the event of interest marks the case period, the *light gray* area marks the reference period. Exposure to the risk factor is marked by "x". Note that subjects 1 and 2 are discordant (differing exposures between case and reference periods), whereas subjects 3 and 4 are concordant (same exposures in case and reference periods)

Case-crossover was evaluated in the context of SJS/TEN already in an earlier study [8]. The authors reported good efficiency of the case-crossover estimates for the risk factors under investigation, but with a few exceptions.

Since the case and reference periods within every subject are matched, data from a case-crossover study can be analyzed like a conventional matched case–control study using CLR.

Statistical analysis

For our case-crossover risk estimation, we include the same methods that were suggested by Avalos et al. [5]. In addition, we considered two further methods not treated by them, namely sublasso and Firth correction.

Univariable conditional logistic regression

In a 1–1 matched study of N pairs (in the case-crossover setting, "pair" denotes a subject's case period that is matched with his reference period), an estimate of a risk factor using univariable CLR is obtained through the log-likelihood function, which writes as

$$\begin{aligned}\ell(\beta) &= \sum_{n=1}^{N}\ell_n(\beta)\\ &= \sum_{n=1}^{N}[\beta x_{1n}-\log(\exp(\beta x_{1n})+\exp(\beta x_{0n}))],\end{aligned}$$

where x_{1n} denotes the exposure during the case period and x_{0n} the exposure during the reference period in the nth subject, and β is the regression coefficient [9]. The principle of maximum likelihood, which is most commonly used, states to use $\hat{\beta}$ as an estimate of β which maximizes the log-likelihood, i.e.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ l(\beta).$$

The odds ratio (OR) of exposure to a risk factor can be directly obtained from a logistic regression model through $\exp(\hat{\beta})$. For risk factor analysis in the case-crossover study of EuroSCAR, we fit separate univariable CLR models for every risk factor.

Multivariable conditional logistic regression

In contrast to the univariable model with just one risk factor and one estimate β, the multivariable model considers a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ of p risk factors with exposure vectors $\boldsymbol{x}_{1\boldsymbol{n}}$ and $\boldsymbol{x}_{0\boldsymbol{n}}$ of length p in the log-likelihood function, using vector multiplication accordingly [9]. This allows for the simultaneous estimation of all risk factors in one model, thereby adjusting their effects for each other.

Lasso

A popular penalization method is the least absolute shrinkage and selection operator (lasso) [10]. Its popularity stems from its low computation cost and it has the property that as estimates are penalized, some are set to exactly 0. This is the main attraction of the lasso: it performs estimation and variable selection simultaneously. Estimation is based on the penalized log-likelihood

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right)$$

where a penalty term is added to the log-likelihood that is tuned with a penalization parameter $\lambda \geq 0$. The parameter λ controls the complexity of the model: if $\lambda = 0$, estimation is the same as with multivariable CLR, but as $\lambda \to \infty$, estimates are shrunk and eventually set to 0. To determine an optimal value of λ, K-fold cross-validation is used. For this, the observations of the dataset are split into K evenly-sized blocks (while preserving the matching of the pairs). Given a value of λ, estimates are obtained by fitting a lasso model to the dataset with observations from one of the K blocks removed (the training set); the likelihood is then evaluated using observations from the left-out block (the test set, see Fig. 2 for clarification). This cross-validation step is performed K times, using every block once for evaluation. Typically, $K = 5$ or $K = 10$ is chosen; here, we use $K = 10$. The likelihood contributions from the K cross-validation steps are then summarized to obtain a likelihood given the value of λ. If this procedure is performed for a range of values for λ, we then choose the value for λ which maximizes the cross-validation likelihood.

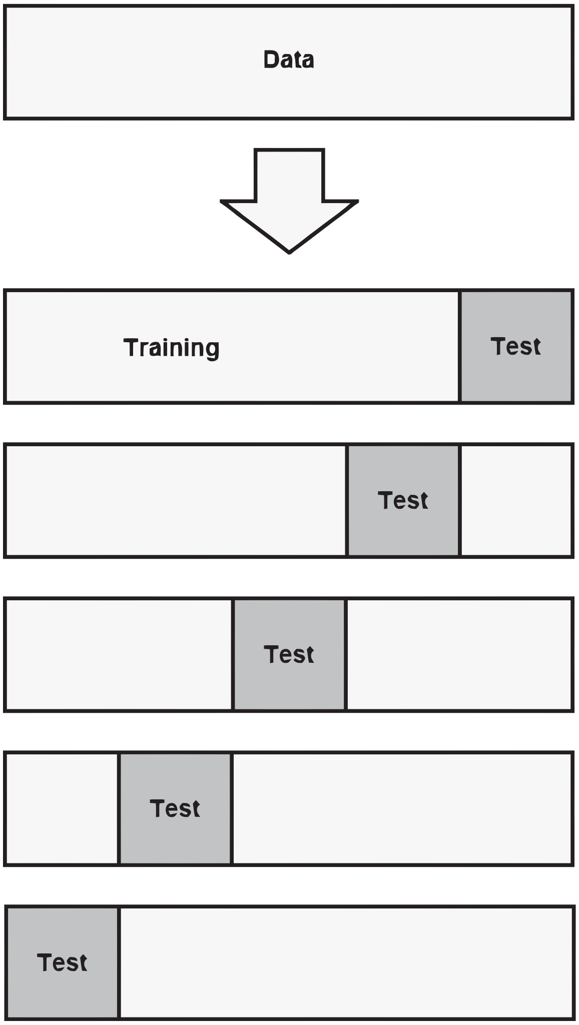

Fig. 2 Illustration of cross-validation. The data is randomly split into K evenly-sized blocks (in this Figure, K = 5), partitioning the data into a training set and a test set. For each fold, a model is fit on the training set and evaluated on the test set. The model that performs best (i.e. has the highest likelihood) on average over all K folds is the one that is selected

Elastic net

A common alternative to the lasso is the elastic net [10]. The penalized log-likelihood

$$\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}) - \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| - \lambda_2 \left(\sum_{j=1}^{p} \beta_j^2 \right)^{\frac{1}{2}} \right)$$

contains two penalty terms that are tuned via parameters $\lambda_1, \lambda_2 \geq 0$. If $\lambda_2 = 0$, $\hat{\boldsymbol{\beta}}$ becomes the lasso estimate; if $\lambda_1 = 0$, $\hat{\boldsymbol{\beta}}$ is the estimator of the ridge regression [10]. Elastic net is therefore a compromise between the lasso and ridge regression. Estimation of $\hat{\boldsymbol{\beta}}$ is again done through K-fold cross-validation, with the extension that maximization is done over a grid of values for λ_1 and λ_2, thereby increasing computation cost. The benefit of using the elastic net is that it tends to be better than lasso in selecting important variables when high correlations are present.

Bolasso

In general, the lasso selects all the relevant variables but additionally also a few irrelevant ones. As a remedy, Bach suggested the bolasso [11]. With the bolasso, a number of bootstrap samples are drawn from the dataset, where each bootstrap sample is generated by sampling N pairs from the dataset with replacement. Here, we will draw 1000 bootstrap samples. A lasso model is then fit to each bootstrap sample. Risk factors not selected in at least 75 % of the bootstrap samples are set to zero, the remaining risk factor estimates are set to their average estimates from all the bootstrap samples. Using a threshold of 75 % had the best results in Avalos et al. [5] and was thus used by us.

Sublasso

Analogous to the bolasso method, but instead of drawing bootstrap samples, sublasso uses subsampling by drawing 75 % of the observations without replacements. Though sublasso was not part of the investigation by Avalos et al. [5], we included it here as an ad hoc extension because there are many contentions to favor subsampling over bootstrapping [12] and because bolasso is easily extendable to accommodate sublasso.

Random lasso

Similar in name and principle to random forest, random lasso is an extension of bolasso which consists of two steps [13]. The first step is the same as the bolasso procedure, but instead of including all risk factors, for each bootstrap sample only a random subset of size q_1 of the risk factors are used for fitting. The resulting estimates then serve as weights in step two: lasso is again fitted to bootstrap samples by only considering a random subset of size q_2 of the risk factors, where the risk factors are selected proportionally to the weights determined in the first step. In the EuroSCAR study, $p = 30$, therefore we select the optimal q_1 and q_2 from the set 15, 20, 25 and 30 using cross-validation. As in bolasso, variables are set to their average estimate from all bootstrap samples if their selection is at least 75 %.

Firth correction

Similar to the lasso, Firth [14] suggested maximizing a modified log-likelihood function,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}) + \frac{1}{2}\log(\det I(\boldsymbol{\beta})) \right),$$

where $I(\boldsymbol{\beta})$ is the observed information matrix, to overcome the challenge of data separation when sample size is small. Firth correction is another effective bias-correction method which has gained some popularity. It was not used by Avalos et al. [5] but it has shown good results in a study design very similar to case-crossover [15]. The adaptation of the Firth correction for CLR is described by Heinze & Puhr [16] and Sun et al. [17].

Analyses were performed using R version 3.1.2. To calculate correlations of risk factors, we employ the Pearson/Spearman correlation, the two being equivalent for binary data such as encountered here. Standard CLR was done using the R package `survival`. CLR with lasso penalization was performed using the R package `clogitL1` by Reid & Tibshirani [18] which is described in Reid & Tibshirani [19] and is available at CRAN. Alternative R packages for CLR with lasso are `clogitLasso`, also available at CRAN, which is described in Avalos & Pouyes [20], and in more detail in Avalos et al. [21], and `pclogit` [22] available at www.columbia.edu/~sw2206/softwares.htm. To implement the Firth correction, we used a macro `CFL` by Heinze available at http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/fllogistf/ which was run using SAS version 9.2. The package `coxphf` [23] provides an implementation of the Firth regression in R for the Cox model. Since the likelihood for a CLR model is equivalent to that of a Cox model with a particular data structure, this package is a suitable alternative. Our R and adapted SAS scripts are available in the supplementary material.

Study population

We apply the methods under investigation using patient data from the EuroSCAR study [7]. The study was designed as a multinational case–control study on patients with SJS/TEN. As these events are mainly caused by a variety of drugs, the main aim was to assess the risk of drugs or drug groups.

The recruitment of cases and controls took place in six countries (Austria, France, Germany, Israel, Italy, and The Netherlands) within the period from April 1997 to December 2001. Altogether, 379 cases and 1,505 controls were included. From the obtained data, 30 drug groups (henceforth simply referred to as drugs) were defined as either “highly suspected” ($n = 9$), “suspected” ($n = 10$), or not suspected (“other”, $n = 11$) of causing SJS/TEN. These classifications were corroborated by Papay et al. [24]. For more details on the EuroSCAR study, we refer the reader to Mockenhaupt et al. [7].

For the case-crossover analysis, only the 379 cases of the EuroSCAR study were extracted. 22 patients were excluded because they had not taken any of the 30 relevant drugs during the observation period and thus could not contribute any information to the estimations. A further 6 patients were excluded because their observation period was too short to accommodate the minimum required number of days for both the case and reference period, leaving a total of 351 subjects for our analysis. The data used in our analysis is available in the supplementary material (see Additional file 1).

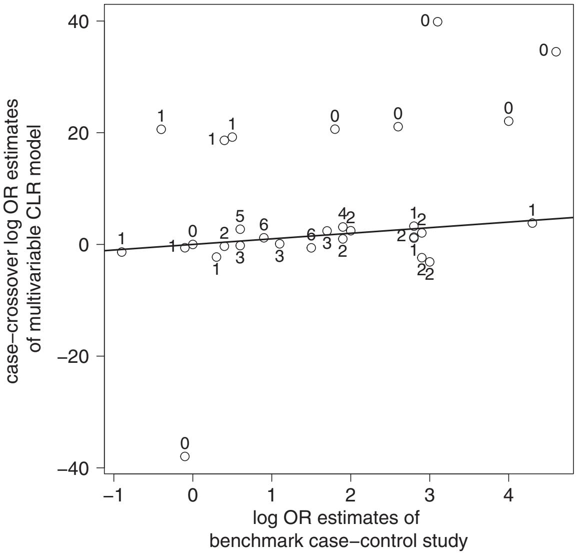

Fig. 3 Estimates from the multivariable CLR plotted against the estimates from the case–control study (Mockenhaupt et al. [7]). The line marks the 45-degree line of equality (note the different scales on the x- and the y-axis). Ideally, the estimates from the multivariable CLR would resemble the estimates from the case–control study; however, some very large discrepancies are apparent. Displayed next to the estimates is the minimum of the two possible discordances between case and reference period of the risk factor. It can be seen that drugs with extreme estimates have only one or zero such discordances, while drugs with at least two discordances have plausible estimates

Evaluation of methods

In our work, the goal of the competing methods is to perform a new classification, herein called reclassification, of the drugs that would ideally resemble the classifications of the case–control study of EuroSCAR [7], thus the latter will serve as a benchmark. Using solely the estimates of the case–control study (Additional file 2: Table S1), the best resemblance of risk classes is obtained by deeming drugs “highly suspected” if log OR > 2.4, “suspected” if 0.45 < log OR < 2.4 and “other” if log OR < 0.45. Further, we employ receiver operating characteristic (ROC) curves by discriminating which drugs are correctly reclassified as “highly suspected”, and alternatively by discriminating which drugs are correctly reclassified as “highly suspected” or “suspected”.

Results

Correlations of drugs within case periods, and separately within reference periods, were all considerably low (<0.32 in absolute value).

Estimates from a multivariable CLR, arguably the most common technique for analyzing matched data, suffer from drugs that are prevalent in subjects’ case periods but infrequent in the reference periods, or vice versa. Only cases with discordant exposure in both periods contribute to the likelihood, thus drugs with only few discordances prove to be problematic for analysis. For such drugs, resulting risk estimates tend to be implausible, either because they are too large or too small.

Figure 3 displays the estimates of the multivariable CLR model plotted against the estimates from the benchmark case–control study. Ideally, the estimates of the former would agree with the latter, in which case the plotted points would fall on the 45 degree line. However, there are some very large disagreements between the two. As indicated by the numbers next to the plotted points, it appears to occur only (though not necessarily) for drugs with infrequent discordances. Therefore, multivariable CLR is very unsatisfactory for obtaining plausible risk estimates.

Figure 4a shows the risk estimate paths of the lasso for different values of the tuning parameter λ. As the tuning parameter increases, the risk estimates shrink and are eventually zero. Cross-validation determines the optimal value of λ and is marked by the vertical dashed line (see Additional file 2: Figure S1 for illustration). These optimal risk estimates are shown in Fig. 4b. They agree much better with the estimates of the benchmark case–control study. There no longer is a problem of highly implausible values.

Table 1 is a summary of the number of correctly reclassified drugs according to the classification by Mockenhaupt et al. [7] (see Additional file 2: Table S2 for a detailed version). Multivariable CLR performs worst overall among all the methods and only correctly reclassifies 17 out of 30 drugs. The penalized methods show large improvements, with lasso and random lasso performing best, and similar results from elastic net, sublasso and Firth correction.

The arguably most important aspect of the EuroSCAR study was to correctly identify drugs that are highly suspected of causing SJS/TEN. This poses a binary discrimination problem (“highly suspected” versus “suspected” or “other”) for which we use ROC curves that do not rely on a single fixed cut-off value to discriminate between risk classes.

Figure 5 shows a comparison of ROC curves for multivariable CLR and the lasso. Note that due to the shrinkage of its estimates, the lasso uses a very different

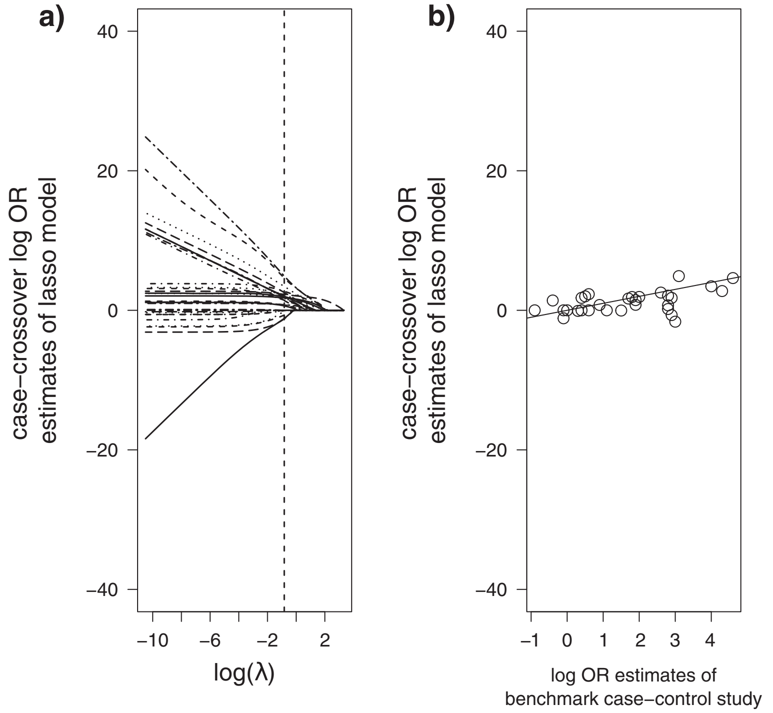

Fig. 4 a Estimates of the lasso for different values of λ. As λ increases, risk estimates shrink and are eventually 0. The dashed line marks the optimal λ-value and the corresponding optimal estimates, as determined by cross-validation. **b** Similarly to Fig. 3, the optimal lasso estimates are plotted against the benchmark case–control estimates. Agreement between the two is evidently much better than for multivariable CLR. The line marks the 45-degree line of equality. The y-scale is chosen to accommodate comparisons with Fig. 3

discrimination scheme: where multivariable CLR and the lasso have the same or similar specificity, the lasso uses a much lower log OR cut-off. It can be seen that the ROC curve of lasso is largely superior to the ROC curve of the multivariable CLR model. Therefore, the lasso outperforms multivariable CLR in discriminating highly suspected drugs.

Table 1 Number of drugs correctly identified under new reclassification rule by each method

Original Classification	Highly suspected ($n = 9$)	Suspected ($n = 10$)	Other ($n = 11$)	Total ($n = 30$)
Rule for new reclassification	log OR > 2.4	0.45 < log OR < 2.4	log OR < 0.45	
Univariable CLR	5	7	7	19
Multivariable CLR	6	2	9	17
Lasso	5	8	9	22
Elastic net	4	8	9	21
Bolasso	4	6	9	19
Sublasso	4	8	9	21
Random lasso	6	6	10	22
Firth correction	5	8	7	20
Case-control[a]	9	10	8	27

[a] Note that even though the case–control study was the basis for the risk classification in Mockenhaupt et al. [7], its estimates do not fully agree with the reclassification scheme used by us. This is because experts classified some of the drugs differently than its risk estimates would suggest based on their experience and opinion, whereas we employ a reclassification scheme using only the estimates

Summarizing ROC curves for all methods with their area under the curve, Table 2a shows that lasso, elastic net and Firth correction perform best for correctly identifying highly suspected drugs. Multivariable CLR again performs worst, and the resampling methods fall in-between.

We also look at which methods best discriminate between risk classes “highly suspected” or “suspected” versus “other”. Again the lasso performs better overall than multivariable CLR. Among all methods, Table 2b shows that Firth correction and elastic net have the best ability to discriminate in this setting. The robustness of these results is corroborated by a sensitivity analysis wherein the AUCs were calculated for 100 subsamples drawn without replacement (Additional file 2: Figure S2). Interestingly, the benchmark case–control study performs rather poorly here. A possible explanation could be that in the case–control study by Mockenhaupt et al. [7] some drugs were classified as “other” simply because they were newly introduced, when in fact their estimates would suggest that they are “suspected” or “highly suspected”. Therefore, an assessment of the supposedly “correct” case–control risk estimates alone need not necessarily lead to the best discrimination scheme.

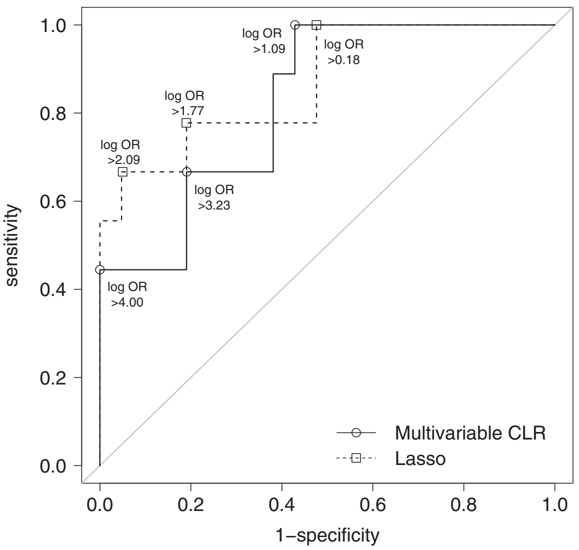

Fig. 5 ROC curves of multivariable CLR and the lasso for reclassifying highly suspected drugs. Cut-off values of the log ORs for discriminating between highly suspected drugs versus suspected or other drugs range from 0 to 4; the sensitivity and 1-specificity of several cut-off values are highlighted. A 45-degree line of equality is added for reference

Discussion

Summary of our work

Penalized regression has been introduced in the literature as a method for bias-correction and has many applications; here, we argue in favor of it in the context of the case-crossover design. Similarly to a previous study, we evaluated the case-crossover design in the context of Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS/TEN) [8]. Using data from the EuroSCAR study [7], the superiority of the lasso over multivariable conditional logistic regression (CLR) was apparent, since the latter seemingly suffered from the problem of sparse data.

Table 2 Area under the curve for binary discrimination of drugs

	a) AUC for “highly suspected drugs“	b) AUC for “highly suspected drugs“ or “suspected drugs“
Univariable CLR	0.836	0.871
Multivariable CLR	0.741	0.775
Lasso	0.862	0.861
Elastic net	0.868	0.909
Bolasso	0.804	0.813
Sublasso	0.815	0.823
Random lasso	0.820	0.847
Firth correction	0.852	0.928
Case-control[a]	0.947	0.809

[a] see footnote in Table 1

We also investigated several resampling approaches together with penalization methods. They did not demonstrate any improvement, thus our results suggest that the increased computation cost is not justified. The Firth correction performed reasonably compared to the penalized methods and was even best in discriminating “highly suspected” or “suspected” drugs. This is particularly noteworthy when considering that the other penalized methods tune their penalty terms whereas the Firth correction uses a constant factor of 0.5. It may be an interesting investigation to explore if a similarly tuned Firth correction would be an appropriate extension.

There are several limitations to our study. First and foremost, although we enjoyed the advantage of having benchmark case–control estimates for our case-crossover analysis, the benchmark is not perfect since the original case–control study had its own limitations. Because of its limitations, some of the risk factors of the case–control study could be estimated only through univariable CLR. However, univariable estimates may not be the best benchmark for comparing multivariable methods. Also, when single cut-off values are used to compare all methods (as was done for Table 1), direct comparisons of estimators may not be suitable when some methods shrink estimates and some do not. For this reason, we also used ROC curves which do not rely on single cut-off values. Further, assessment solely based on estimates may be incomplete without also considering standard errors of the estimates. This, however, is a weakness of the lasso, as standard errors are not directly available [25, 26]. For a further discussion on the limitations of penalized regression, see Greenland et al. [27].

Penalized regression in epidemiology and case-crossover

A good introduction to penalized regression is given by Cole et al. [4]. However, epidemiologic literature on penalized regression is still relatively scarce. An example of an exception is Rose [28] who developed risk scores for mortality prediction using, among other techniques, penalized regression. In another study, Burgette et al. [29] implement the lasso and elastic net to model adverse birth outcomes. Further, Smith et al. [30] apply penalized regression to model the BMI in a longitudinal study. In a large study that uses case-crossover with CLR, Mostofsky et al. [31] investigate the association between particle constituents of air pollution and health outcomes. Theirs is a great example where penalized regression could be applied since their risk factors suffer from high collinearity.

Sullivan & Greenland [32] point out that sparse-data artefacts often go unrecognized in study reports. They provide an example of a case–control study which reported an odds ratio between ever having smoked and ICU admission of 65 (95 % CI 6.3–672) without ever considering the plausibility of the estimate. Cole et al. [4] stress the importance of penalized regression by pointing out that

> "[…] epidemiologists are aware of problems due to sparse-data bias in very small studies, but sparse-data bias appears to be less widely recognized when it occurs in larger studies." (p. 257)

Among the attractive features in favor of penalized methods is that the shrinkage of coefficients is proportional to the estimated variance of the coefficients, thus unstable estimates are shrunk more than stable ones. And further, penalized regression can be applied when covariates are collinear and conventional methods fail completely [3].

In the context of case-crossover, there has been hardly any work on using penalized regression. Walter & Tiemeier [33] conducted a survey of 300 articles published in four major epidemiologic journals, not one of which used penalized regression. Though their work is no longer up to date, a recent literature search revealed only three articles that use penalized regression in the context of case-crossover, all by Avalos and her colleagues (Additional file 2: Table S3).

In a study design closely related to case-crossover, the self-controlled case series design, work in the literature on penalization methods has been equally scarce. The self-controlled case series is tantamount to the classical cohort study in the same way that case-crossover is the case series equivalent of a classical case–control study [34]. In a work using the self-controlled case series design, Zeng et al. [15] study bias correction methods for datasets with a small number of adverse events (i.e. a sparse data), such as a vaccine safety study. In it, the Firth estimate consistently outperformed the classical maximum likelihood estimate.

In a recent work by Avalos et al. [35], the authors describe in detail an algorithm more efficient than the one in their initial paper [5], making it suitable even for large datasets. The authors also provide a good comparison of the different R packages available that implement the lasso in CLR (and also Cox and unconditional logistic regression).

Conclusion

Standard maximum likelihood is the default in most statistical software packages and has many desirable large-sample properties, among them asymptotic unbiasedness. The large-sample condition, however, is difficult to achieve satisfactorily, and consequently, bias can be substantial for small-sample studies.

Therefore, arguments are strong for wider use of penalized regression in epidemiological studies. Of great practical importance is that penalized regression is easily implemented for linear, Poisson, Cox, or CLR models. For comparison, running a CLR requires two lines of code in R:

```
library(survival)
clogit(outcome ~ risk.factor + strata(id),
        data = mydata)
```

Running the lasso is only marginally more work, requiring three lines:

```
library(clogitL1)
lasso <- clogitL1(y = mydata$outcome,
            x = mydata$risk.factor, strata = mydata$id)
cv.clogitL1(lasso)
```

Avalos et al. [36] recommend penalized regression as an alternative to conventional strategies, and Cole et al. [4] even suggest that penalized likelihood should arguably replace standard maximum likelihood as the default method. For the case-crossover design, we also encourage penalized regression for routine use.

Additional files

Additional file 1: Dataset and scripts used for analysis. (ZIP 28.0 kb)

Additional file 2: Table S1. Results for the 30 drugs and drug groups of the EuroSCAR study, a Multinational Case–control study in Europe and Israel between 1997 and 2001 on patients with SJS/TEN [4]. **Table S2.** Reclassification of individual drugs by each method. **Table S3.** Relevant results of literature search for penalized regression used with case-crossover studies, performed on November 21st, 2015. **Figure S1.** Deviance (−2 × log-likelihood) of the lasso for different values of λ using 10-fold cross-validation. **Figure S2.** Sensitivity analysis of Table 2a. (DOCX 121 kb)

Abbreviations

AUC, area under the curve; CLR, conditional logistic regression; EuroSCAR, European Study of Severe Cutaneous Adverse Reactions; Lasso, least absolute shrinkage and selection operator; OR, odds ratio; ROC, receiver operating characteristic; SJS/TEN, Stevens-Johnson syndrome and toxic epidermal necrolysis

Acknowledgements

We thank Jean-Claude Roujeau, Cecile Viboud, Ariane Dunant, Sima Halevy, Jan Nico Bouwes Bavinck, Alexis Sidoroff, Jürgen Schneck and Antoine Flahault on behalf of the EuroSCAR team for providing access to their data. We also thank Martin Wolkewitz for valuable comments.

Funding

EuroSCAR was funded (unrestricted grants) by the following institutions/companies: ADIR \& Cie, Bayer Pharma/AG/Vital, Boehringer Ingelheim, Cassenne, Ciba Geigy/Novartis, Cilag GmbH, Dr. Willmar Schwabe GmbH, Goedecke Parke Davis, Glaxo Wellcome/GlaxoSmithKline, Hoechst AG/Hoechst Marion Roussel/Aventis, Hoffmann-La-Roche, IRIS Servier, Jouveinal Lab., LEO, LILLY, MSD Sharp \& Dohme, Pfizer, Rhone Poulenc Rorer, Sanofi Winthrop/Sanofi Synthelabo GmbH, Schering AG; Funding from Pharmaceutical Companies in France were managed through contract with INSERM (Institut National de la Santé et de la Recherche Médicale), French Ministry of Health (PHRC AOM 98027).
The article processing charge was funded by the German Research Foundation (DFG) and the University of Freiburg in the funding programme Open Access Publishing. The authors have no further funding to report.

Availability of data and materials

The dataset supporting the conclusions of this article is included within the article. The data used in our analysis along with the R and SAS scripts are available in the supplementary material.

Authors' contributions

Author SD did the programming and writing of the study. MM and LN are responsible for the data. MS is responsible for the study design. PS reviewed and interpreted the analysis. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

As reported in the original publication, the EuroSCAR-study had been approved by the ethical committee and institutional review board of each center participating in the project in the 6 countries. Informed consent was obtained from participants before collecting their data.

Author details

[1]Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany. [2]Dokumentationszentrum schwerer Hautreaktionen (dZh), Medical Center, University of Freiburg, Freiburg, Germany. [3]USC di Dermatologia, Azienda Ospedaliero Papa Giovanni XXIII, Bergamo, Italy.

Received: 14 April 2016 Accepted: 28 July 2016
Published online: 22 August 2016

References

1. Breslow NE, Day NE. Statistical Methods in Cancer Research. Vol. 1. The Analysis of Case–control Studies. (IARC Scientific Publication no. 32). Lyon: International Agency for Research on Cancer; 1980.
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):144–53.
3. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol. 2008;167:523–9.
4. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. Am J Epidemiol. 2014;179(2):252–60.
5. Avalos M, Grandvalet Y, Adroher ND, et al. Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood. Stat Med. 2012;31(21):2290–302.
6. Avalos M, Orriols L, Pouyes H, et al. Variable selection on large case-crossover data: application to a registry-based study of prescription drugs and road traffic crashes. Pharmacoepidemiol Drug Saf. 2014;23:140–51.
7. Mockenhaupt M, Viboud C, Dunant A, et al. Stevens-Johnson syndrome and toxic epidermal necrolysis: assessment of medication risks with emphasis on recently marketed drugs. The EuroSCAR study. J Invest Dermatol. 2008;128(1):35–44.
8. Viboud C, Boëlle PY, Kelly J, et al. Comparison of the statistical efficiency of case-crossover and case–control designs: Application to severe cutaneous adverse reactions. J Clin Epidemiol. 2001;54:1218–27.
9. Hosmer DW, Lemeshow S. Applied Logistic Regression. 3rd ed. Hoboken: Wiley-Interscience; 2013.
10. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer Publishing Company; 2009.
11. Bach F. Bolasso: model consistent lasso estimation through the bootstrap. McCallum A, Roweis S, Cohen W, (eds). In Proceedings of the 25th International Conference on Machine Learning (ICML 2008). Helsinki, Finland; 2008.
12. De Bin R, Janitza S, Sauerbrei W, et al. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. Biom. 2016;72(1):272–280.
13. Wang S, Nan B, Rosset S, et al. Random lasso. Ann Appl Stat. 2011;5(1):468–85.
14. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.
15. Zeng C, Newcomer SR, Glanz JM, et al. Bias correction of risk estimates in vaccine safety studies with rare adverse events using a self-controlled case series design. Am J Epidemiol. 2013;178(12):1750–9.
16. Heinze G, Puhr R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. Stat Med. 2010;29(7–8):770–7.
17. Sun JX, Sinha S, Wang S, et al. Bias reduction in conditional logistic regression. Stat Med. 2011;30(4):348–55.
18. Reid S, Tibshirani R. clogitL1. R package version 1.4. 2014.
19. Reid S, Tibshirani R. Regularization paths for conditional logistic regression: the clogitL1 package. J Stat Softw. 2014;58(12):1–23.
20. Avalos M, Pouyes H. clogitLasso: an R package for L1 penalized estimation of conditional logistic regression models. 1eres Rencontres R: Bordeaux, France; 2012. (available at https://hal.archives-ouvertes.fr/file/index/docid/717505/filename/gabarit_RR.pdf. Accessed 4 Aug 2016).
21. Avalos M, Grandvalet Y, Pouyes H, et al. 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: Formenti E, Tagliaferri R, Wit E, editors. Computational Intelligence Methods for Bioinformatics and Biostatistics, vol. 8452. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing; 2014. p. 109–24.
22. Sun H, Wang S. Network-based regularization for matched case–control analysis of high-dimensional DNA methylation data. Stat Med. 2013;32(21):2127–39.
23. Ploner M, Heinze G. coxphf: Cox regression with Firth's penalized likelihood. R package version 1.11. 2015.
24. Papay J, Yuen N, Mockenhaupt M, et al. Spontaneous adverse event reports of Stevens-Johnson syndrome/toxic epidermal necrolysis: detecting associations with medications. Pharmacoepidemiol Drug Saf. 2012;21(3):289–96.
25. Herring AH. Nonparametric bayes shrinkage for assessing exposures to mixtures subject to limits of detection. Epidemiology. 2010;21 Suppl 4:71–6.
26. Chatterjee A, Lahiri SN. Bootstrap lasso estimators. JASA. 2011;106(494):608–25.
27. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol. 2000;151(5):531–9.
28. Rose S. Mortality risk score prediction in an elderly population using machine learning. Am J Epidemiol. 2013;177(5):443–52.
29. Burgette LF, Reiter JP, Miranda ML. Exploratory quantile regression with many covariates: an application to adverse birth outcomes. Epidemiology. 2011;22(6):859–66.
30. Smith ADAC, Heron J, Mishra G, et al. Model selection of the effect of binary exposures over life course. Epidemiology. 2015;26(5):719–26.
31. Mostofsky E, Schwartz J, Coull BA, et al. Modeling the association between particle constituents of air pollution and health outcomes. Am J Epidemiol. 2012;176(4):317–26.
32. Sullivan SG, Greenland S. Bayesian regression in SAS software. Int J Epidemiol. 2013;42(1):308–17.
33. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6.
34. Farrington CP. Relative incidence estimation from case series for vaccine safety. Biometrics. 1995;51(1):228–35.
35. Avalos M, Pouyes H, Grandvalet Y, et al. Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm. BMC Bioinformatics. 2015;16 Suppl 6:51.
36. Avalos M, Adroher ND, Lagarde E, et al. Prescription-drug-related risk in driving: comparing conventional and lasso shrinkage logistic regression. Epidemiology. 2012;23(5):706–12.