eAPPENDIX

LASSO SHRINKAGE LOGISTIC REGRESSION

1. Model

Let p be the number of exposures, q the number of adjustment variables, n the number of subjects, y_i the binary response variable for subject i, coded as 1 for cases and 0 for controls, $\mathbf{x}_i=(x_{i1}, \ldots, x_{ip})$ a vector where x_{ij} is the j-th exposure for the i-th subject, coded as 1 for exposed and 0 for unexposed and $\mathbf{z}_i=(z_{i1}, \ldots, z_{iq})$ a vector where z_{ik} is the k-th adjustment variable for the i-th subject. Let $D=\{(\mathbf{x}_i, \mathbf{z}_i, y_i)\}_{i=1, \ldots, n}$ be the observed data.

In ordinary logistic regression, we write the probability of observing the event given the values of the exposure and adjustment variables for subject i as

$$p_i = \Pr(y_i = 1 | \mathbf{z}_i, \mathbf{x}_i) = \frac{\exp(\alpha_0 + z_{i1}\alpha_1 + \ldots + z_{iq}\alpha_q + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p)}{1 + \exp(\alpha_0 + z_{i1}\alpha_1 + \ldots + z_{iq}\alpha_q + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p)},$$

where β_j and α_k are the regression coefficients corresponding to the log-odds ratio of the j-th exposure-outcome association and the log-odds ratio of the k-th adjustment variable-outcome association, respectively, and α_0 is the intercept. The parameter vector $(\boldsymbol{\alpha},\boldsymbol{\beta})=(\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)$ is usually estimated by maximizing the log-likelihood function

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right).$$

2. Estimation

In ordinary logistic regression, the parameter vector $(\boldsymbol{\alpha},\boldsymbol{\beta})=(\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)$ is usually estimated by maximizing the log-likelihood function

$$l((\boldsymbol{\alpha}, \boldsymbol{\beta}), D) = \sum_{i=1}^{n} \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right).$$

The general idea behind maximum likelihood estimation is to find the population that is more likely than any other to produce the observed data. The lasso method applied to logistic regression consists in maximizing the log-likelihood penalized by the sum of coefficients in absolute value (the L1 norm) corresponding to exposures:

$$(\hat{\boldsymbol{\alpha}}(\lambda),\hat{\boldsymbol{\beta}}(\lambda)) = \underset{(\boldsymbol{\alpha},\boldsymbol{\beta})}{\arg\max}\, l(\boldsymbol{\alpha},\boldsymbol{\beta}),D) - \lambda\sum_{j=1}^{p}\left|\beta_j\right|.$$

where λ is the positive tuning constant which controls the amount of shrinkage. The lasso penalty shrinks each β_j and sets some of them to zero, leading to parsimonious models, and enabling continuous model selection. Adjustment variables and the intercept are not penalized and thus forced into the model. Adjustment variables are potential confounders found to be associated with responsibility in the crash (according to 95% confidence intervals from univariate logistic models) including crash-related and socio-demographic factors and the presence of chronic diseases (Table 1). We forced them into the model to ensure that the apparent differences between exposed and unexposed to drugs are not misleadingly created by confounding covariates and to enhance the comparability of models constructed from the two different strategies or within the same strategy.

It should be noted that, even if exposures are in the same units already (as in the case of binary exposures), it is useful to standardize since penalization techniques based on norms are sensitive to scaling. Without scaling, the lasso estimate has a tendency to disregard exposures with small variability on the sample used for estimation, corresponding here to rarely prescribed drugs, which are as important regarding inference as more commonly prescribed ones.

3. Selection of the tuning constant

The positive tuning constant λ controls the amount of shrinkage. In general, the smaller λ, the more the penalty is relaxed, and the more exposures are selected. Inversely, the higher λ, the more exposures are eliminated. The *regularization path* is the continuous trace of the shrinkage estimates of the regression coefficients obtained when varying λ from 0 (the maximum-likelihood solution for the full logistic model) to a certain threshold, which depends on data, beyond which no exposures are retained in the model.

To select the proper amount of shrinkage, we conducted a grid-search with cross-validation. We considered a log-scale grid of 50 λ-values ranging from 0 to the (data derived) smallest value for which all coefficients are zero. Then, we applied the 10-fold cross-validated area under the curve (AUC) criteria. The data set D is first randomly chunked into K=10 disjoint blocks of approximately equal size. For each value, the logistic regression coefficients are estimated K times, one for each union of (K-1) blocks of data used for estimation, thus each time leaving out one block, which is then used to compute the AUC from data that were not used in the estimation process.

Shrinkage estimation goes through the introduction of a bias on the estimated coefficients. To correct bias, we fitted the unpenalized logistic regression model with the adjustment variables and the exposures retained in the model (those having a nonzero point estimate of log-odds ratio). [28]

4. Confidence intervals

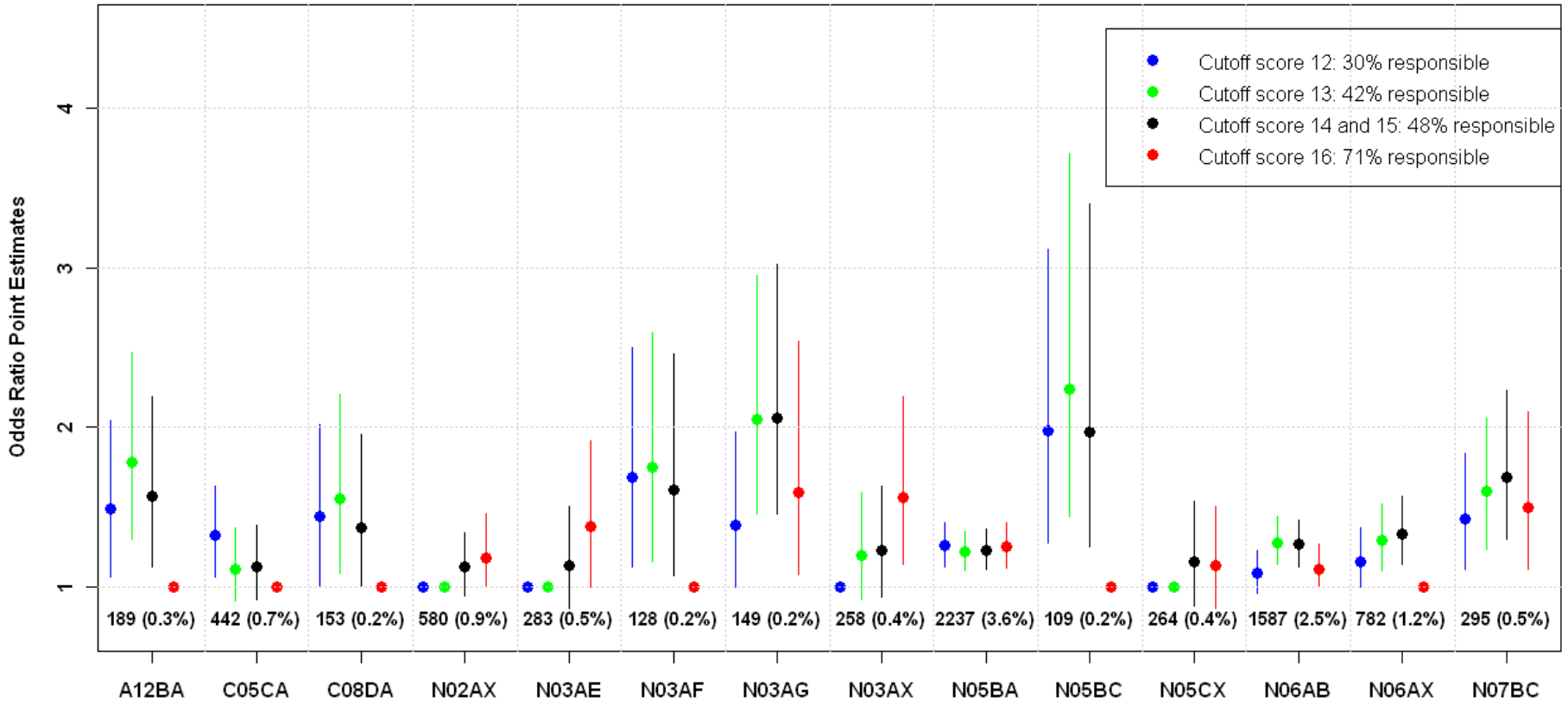
To account for uncertainty in selection, we built 95% bootstrap percentile confidence intervals, using 5000 replicates. [26] The bootstrap distribution of each bias-corrected coefficient was computed, and the 2.5th and 97.5th percentiles of the empirical distribution formed the limits for the 95% bootstrap percentile confidence interval.

5. Computation and software

Several algorithms have been proposed for solving the lasso shrinkage logistic regression. Here we used the R package *glmnet* since it has been shown to be faster than competing methods. [12,13] Also, practical options are available, for example, sparse data matrices, resulting from rare exposures, can be stored in sparse format; exposures (penalized) and adjustment (unpenalized) variables can be specified; model selection is performed by several criteria; when variable standardization is performed prior to fitting, the coefficients are returned on the original scale.

eTABLE: Adjusted Odds Ratio Lasso Estimates and 95% Bootstrap Percentile Confidence Intervals, Using 5000 Replicates, for Potential Traffic Crash and Socio-demographic Confounders Showing the Most Relevant Associations.

Adjustment variable	**Category**	**Adjusted OR lasso estimates and 95% bootstrap CIs**
Age (y)	<18	1.67 (1.51-1.86)
	18-24	1.61 (1.52-1.71)
	25-34	1.23 (1.17-1.30)
	35-44	1.02 (0.97-1.08)
	45-54	1.00
	55-64	1.01 (0.93-1.08)
	65-74	1.40 (1.23-1.58)
	>=75	2.68 (2.33-3.09)
Alcohol level (g/L)	<0.5	1.00
	[0.5-0.8[	4.29 (3.49-5.35)
	[0.8-1.2[	7.58 (6.18-9.60)
	[1.2-2.0]	11.04 (9.19-13.67)
	> 2	13.21 (10.88-16.73)



eFIGURE. Odds ratio point estimates and 95% bootstrap confidence intervals obtained with the lasso method (due to space constraints, only some of them are represented), when the cutoff for responsibility varies around the selected cutoff value: 15 (this choice is based on the concordance with decision maker experts). Values above the drug names indicate the number (and percentage) of exposed subjects. Only exposures appearing associated with responsibility (from confidence intervals) for at least one of the cutoff values are presented. Point estimates equal to 1, without confidence intervals, correspond to exposures not selected in the corresponding model.