Analysis of a Large French Registry-Based Study of Prescription Drugs and Road Traffic Crashes

Marta AVALOS[1,2] and Ludivine ORRIOLS[1,2]

1- Univ. Bordeaux, ISPED, Centre INSERM U897 - Epidemiologie-Biostatistique, F-33000 Bordeaux, France

2- INSERM, ISPED, Centre INSERM U897 Epidemiologie-Biostatistique, F-33000 Bordeaux, France

http://www.isped.u-bordeaux2.fr

Joint work with Emmanuel LAGARDE, Frantz THIESSARD, Benjamin CONTRAND, Nuria DURAN-ADROHER, Hélène POUYES, Yves GRANDVALET and the CESIR research group

Roadmap

1. Context
2. CESIR project
3. Design studies
4. Statistical analysis using the Lasso: justification, limitations
5. Implementation
6. Pharmacoepidemiological results
7. Conclusions

Context

Driving is a complex **psychomotor** and **cognitive** task requiring:

- Alertness, attention and processing speed
- Reaction speed and coordination
- Sensory-perceptual functions (visual, auditory,… abilities)
- Executive functions (monitoring, carrying out cognitive activities)

Drugs (**alcohol**, **illicit** drugs, **medicinal** drugs) have a potential effect on the skills needed to drive safely:

drowsiness, impaired judgment, lack of self-confidence, sedation, fainting, hypotension, hypoglycemia, blurred vision, dizziness, loss of coordination ….

Context

Medicinal drugs are a complex issue from a road safety viewpoint:

- high variety of pharmaceutical classes, different prevalence of use

- confounding underlying health conditions,

- medicinal benefits leading to improved (rather than impaired) driving

- whilst on medication, more attention to compensate for perceived risk

- dose, cumulative dose, duration of consumption (prevalent, intermittent, incident), interruption, tolerance,…

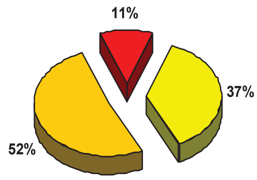
- co-consumption and interaction of drugs, …

Literature focus on hypnotics and anxiolytics (benzodiazepines), few results (frequently inconsistent) for other medicinal drugs

Context

In France

- 1983: warning message and advice on driving included in the labeling of drugs (European directive)
- 1999: unique triangular pictogram 
- 2003: visual system of grading



Effectiveness in “real life”? Reliable information?

Over-representation of certain drug classes in crash-risk statistics?

Gap of Pharmacoepidemiological data

Context

A major approach: population-registries data

UK [Barbone98,Gibson09]

Norway [Engeland07]

France [Orriols10]

Finland [Karjalainen12]

Advantage: the potential to study associations between rare exposures and outcomes in a sample large enough to provide sufficient precision

Particularity: high-dimensional and/or large-scale datasets, whose analysis needs to be addressed using appropriate statistical and computational techniques

Roadmap

1. Context
2. CESIR (Combination of Studies on Health and Road Safety) project
3. Design studies
4. Statistical analysis using the Lasso: justification, limitations
5. Implementation
6. Pharmacoepidemiological results
7. Conclusions

CESIR (Combination of Studies on Health and Road Safety)

National healthcare insurance
Reimbursement database

National healthcare insurance database (SNIIR-AM)
Covers the whole French population

- **national healthcare ID (unique)**
- sex
- birth date

- reimbursed prescription drugs
- dates of dispensing
- chronic diseases (30 recognized)

National institute for research on transport and safety
Crash database

National police database of injurious road traffic crashes (BAAC)
Standardized grid recorded for statistics purposes about crashes

-crash characteristics
-location
-vehicles
-users, role (driver, passengers..)
-eventual traffic violation

- **police report nb (not unique)**
- crash date
- crash location's zip code
- police forces who recorded
- sex
- birth date

CESIR (Combination of Studies on Health and Road Safety)

National healthcare insurance
Reimbursement database

National healthcare insurance database (SNIIR-AM)
Covers the whole French population

- **national healthcare ID (unique)**
- sex
- birth date

- reimbursed prescription drugs
- dates of dispensing
- chronic diseases (30 recognized)

National institute for research on transport and safety
Crash database

National police database of injurious road traffic crashes (BAAC)
Standardized grid recorded for statistics purposes about crashes

-crash characteristics
-location
-vehicles
-users, role (driver, passengers..)
-eventual traffic violation

- **police report nb (not unique)**
- crash date
- crash location's zip code
- police forces who recorded
- sex
- birth date

CESIR (Combination of Studies on Health and Road Safety)

National healthcare insurance
Reimbursement database

National healthcare insurance database (SNIIR-AM)
Covers the whole French population

- **national healthcare ID (unique)**
- sex
- birth date

- reimbursed prescription drugs
- dates of dispensing
- chronic diseases (30 recognized)

Police reports (PV)
Mandatory for injury crashes.
Handwritten, stored as image files for insurance purposes.

Automatic extraction using Optical Character Recognition

- **national healthcare ID** (**unique**
- sex
- birth date

- **police report nb (not unique)**
- crash date
- crash location's zip code
- police forces who recorded
- sex
- birth date

National institute for research on transport and safety
Crash database

National police database of injurious road traffic crashes (BAAC)
Standardized grid recorded for statistics purposes about crashes

-crash characteristics
-location
-vehicles
-users, role (driver, passengers..)
-eventual traffic violation

- **police report nb (not unique)**
- crash date
- crash location's zip code
- police forces who recorded
- sex
- birth date

CESIR (Combination of Studies on Health and Road Safety)

National healthcare insurance
Reimbursement database

National healthcare insurance database (SNIIR-AM)
Covers the whole French population

- **national healthcare ID (unique)**
- sex
- birth date

- reimbursed prescription drugs
- dates of dispensing
- chronic diseases (30 recognized)

Police reports (PV)
Mandatory for injury crashes.
Handwritten, stored as image files for insurance purposes.

Automatic extraction using Optical Character Recognition

- **national healthcare ID (unique**
- sex
- birth date

- **police report nb (not unique)**
- crash date
- crash location's zip code
- police forces who recorded
- sex
- birth date

National institute for research on transport and safety
Crash database

National police database of injurious road traffic crashes (BAAC)
Standardized grid recorded for statistics purposes about crashes

- crash characteristics
- location
- vehicles
- users, role (driver, passengers..)
- eventual traffic violation

- **police report nb (not unique)**
- crash date
- crash location's zip code
- police forces who recorded
- sex
- birth date

CESIR (Combination of Studies on Health and Road Safety)

The combined police data sources complete each other and allows for verification

Police Reports

210,818 reports (each including at least one driver and possibly several drivers and/or pedestrians and/or passengers)

↓

109,078 national IDs collected in police reports

Police national database of injurious crashes

234,679 records (each including at least one driver and possibly several drivers and/or pedestrians and/or passengers)

↓

527,591 individuals among which 392,169 drivers

↘ ↙

97,438 national IDs matched (drivers, pedestrians or passengers)

↓

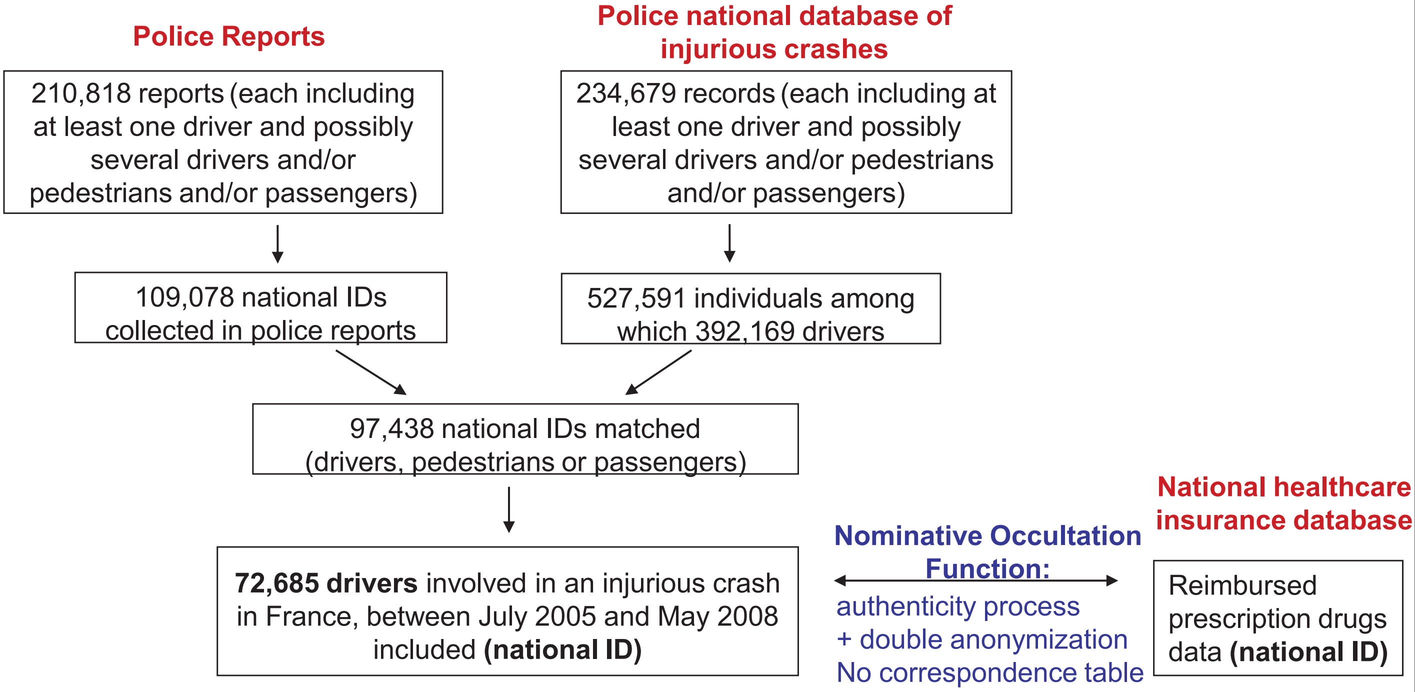
72,685 drivers involved in an injurious crash in France, between July 2005 and May 2008 included **(national ID)**

Matching procedure based on:
police report nb, crash date, zip code, police forces, birth date, sex.

- If all variables concordant => match
- If 3 or + discordant variables => unmatch
- Otherwise => probabilistic linkage method [Jaro95]

CESIR (Combination of Studies on Health and Road Safety)

The combined police data sources complete each other and allows for verification



Available Data

→ Socio-demographic characteristics: age, sex, socioeconomic category

→ Crash-related characteristics: year, season, day of week, time of the day, injury severity, location (inhabitants), vehicle type, alcohol level

→ responsibility for the crash deduced

→ Chronic diseases 30 recognized in France (ICD-10 code)

→ Reimbursement dates of prescribed drugs (the 6 months prior the crash) coded by the Anatomical Therapeutic Chemical (ATC) classification

→ daily medication exposure (exposed=1, unexposed=0) estimated from median value of treatment duration for each ATC class (4th level) [IMS2008]

Available Data

For example: ATC code for "Aspirin"

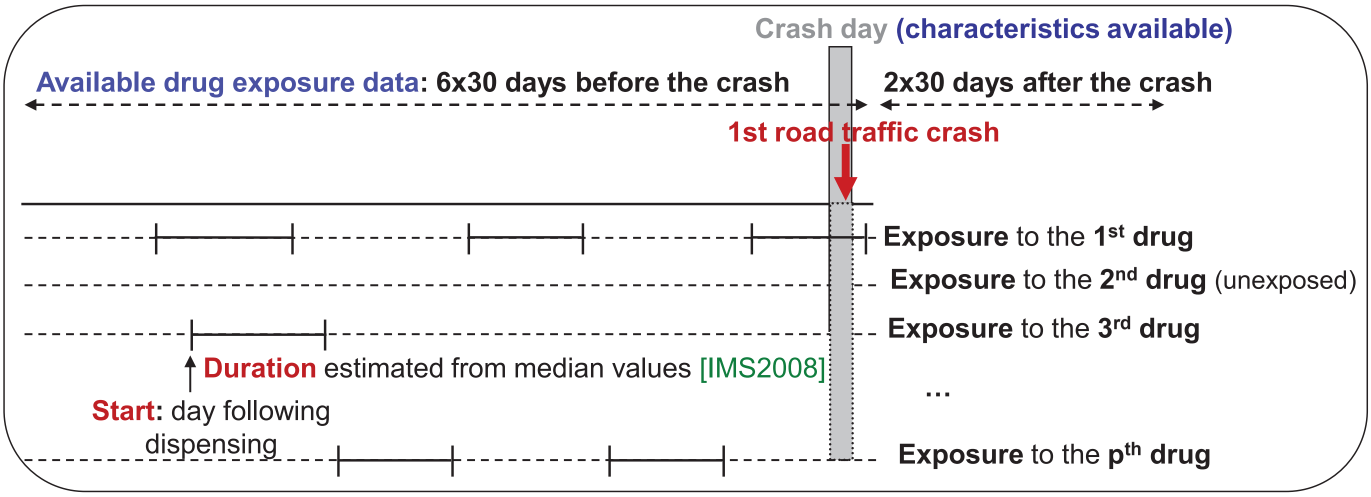
1st level (anatomical main group):	N Nervous system
2nd (therapeutic main group):	N02 Analgesics
3rd (pharmacological subgroup):	N02B No opioids analgesics and antipyretics
4th (chemical subgroup):	N02BA Salicylic acid and derivatives
5th (chemical substance):	N02BA01 Acetylsalicylic acid

→ Reimbursement dates of prescribed drugs (the 6 months prior the crash) coded by the Anatomical Therapeutic Chemical (ATC) classification

→ daily medication exposure (exposed=1, unexposed=0) estimated from median value of treatment duration for each ATC class (4th level) [IMS2008]

Available Data

ith subject (**Socio-demographic characteristics available**)



72,685 drivers involved in an injurious crash in France, 2005 – 2008

62,766 without missing values for the alcohol level

Exposure to ~400 prescription drugs coded with the 4th ATC level

Objective: to identify prescription drugs associated with an increased risk of crash

Study Designs: Case-Control

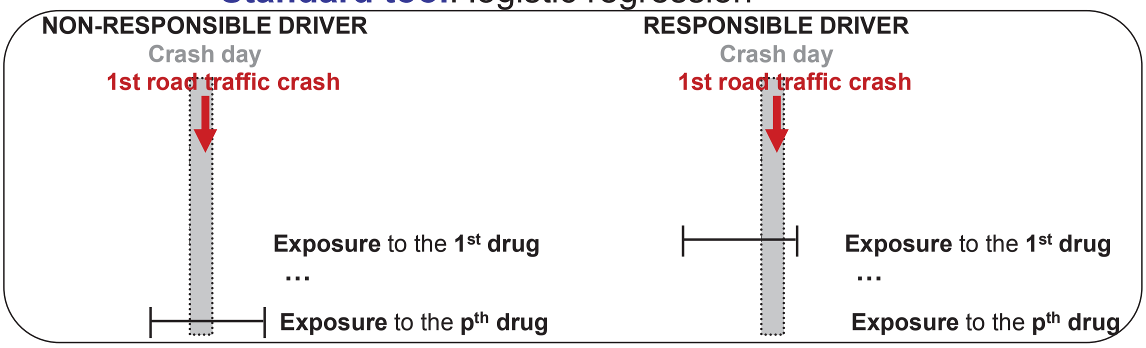
→ Automated scoring tool based on factors likely to reduce driver responsibility (road/driving conditions, traffic rule obedience, including alcohol…) [Robertson94]

→ Selection of a cutoff score based on the consensus with traffic safety experts (using a validation sample) [Laumon05]

→ **Responsibility analysis**: responsible (cases) / non-responsible (controls)

→ Exposure probabilities on the day of crash compared between responsible drivers and non-responsible drivers

Standard tool: logistic regression



Study Designs: Case-Crossover

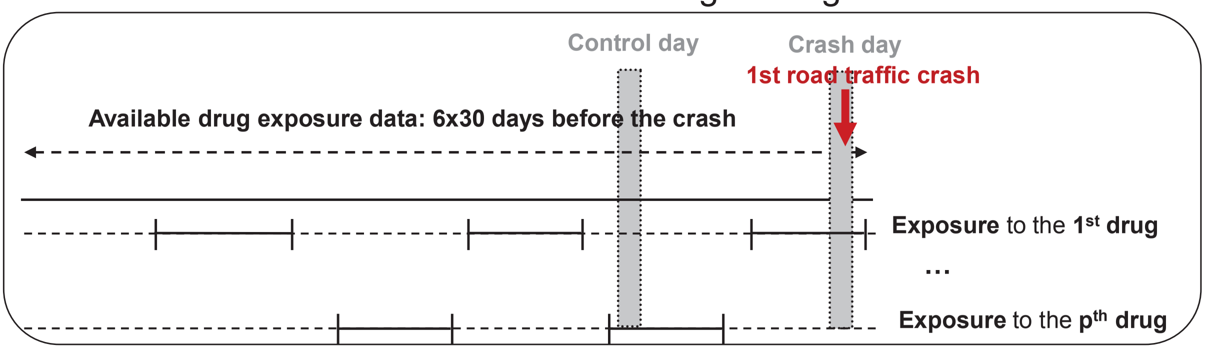
→ Effects of transient exposures on the risk of acute-onset events [Maclure91]

→ Exposure during the period just prior to the event (case period) compared to the same subject's exposure during one or more control periods

→ Inherently eliminates the bias in control selection and removes the confounding effects of time-invariant factors

→ To identify the short-term, transient triggers of event (Why now?), rather than to identify who is at highest risk (Why them?)

Standard tool: conditional logistic regression



Study Designs: limitations

→ Sample over-represented drivers injured in severe crashes

→ residual confounding by underlying health status (causal inference is not possible)

→ exposure periods estimation (misuse of prescribed drugs, dosage changes, use of illicit or over-the-counter drugs,...)

Case-Control

→ determination of responsibility tool have to be evaluated

Case-crossover

→ no data ensures driving during the control window

→ sensitivity to the effects of unmeasured time-varying risk factors

Roadmap

1. Context
2. CESIR project
3. Design studies
4. Statistical analysis using the Lasso: justification, limitations
5. Implementation
6. Pharmacoepidemiological results
7. Conclusions

L_1-Penalization

The parameters are defined by the penalization problem:

$$\hat{\beta}(\lambda) = \arg\max_{\beta} \Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta}) - \lambda T(\boldsymbol{\beta})$$

$\Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta})$ is the log likelihood

λ is a tuning parameter that controls the model complexity

$T(\boldsymbol{\beta})$ is a penalty function on the coefficient vector

L_1-Penalization

The parameters are defined by the penalization problem:

$$\hat{\beta}(\lambda) = \arg\max_{\beta} \Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta}) - \lambda T(\boldsymbol{\beta})$$

$\Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta})$ is the log likelihood

λ is a tuning parameter that controls the model complexity

- $\lambda = 0$ unpenalized problem ie max likelihood estimates
- λ high enough: all coefficients=0

$T(\boldsymbol{\beta})$ is a penalty function on the coefficient vector

L_1-Penalization

The parameters are defined by the penalization problem:

$$\hat{\beta}(\lambda) = \arg\max_{\boldsymbol{\beta}} \Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta}) - \lambda T(\boldsymbol{\beta})$$

$\Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta})$ is the log likelihood

λ is a tuning parameter that controls the model complexity

- $\lambda = 0$ unpenalized problem ie max likelihood estimates
- λ high enough: all coefficients=0

$T(\boldsymbol{\beta})$ is a penalty function on the coefficient vector

→ All coefficient estimates are shrunk individually towards 0

→ By introducing a (relatively small) bias, we reduce the variance

L_1-Penalization

The parameters are defined by the penalization problem:

$$\hat{\beta}(\lambda) = \arg\max_{\boldsymbol{\beta}} \Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta}) - \lambda T(\boldsymbol{\beta})$$

$\Sigma_i L(\mathbf{x}_i, y_i, \boldsymbol{\beta})$ is the log likelihood

λ is a tuning parameter that controls the model complexity

- $\lambda = 0$ unpenalized problem ie max likelihood estimates
- λ high enough: all coefficients=0

$T(\boldsymbol{\beta})$ is a penalty function on the coefficient vector

→ All coefficient estimates are shrunk individually towards 0

→ By introducing a (relatively small) bias, we reduce the variance

In particular, $T(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1 = \Sigma_i |\beta_i|$ is called the Lasso [Tibshirani96,Friedman10]:

→ Attractive feature: the ability to shrink some coefficients to exactly 0, performing both estimation and variable selection simultaneously

Justification of using the Lasso

1. Good compromise between traditional and data-driven approaches: modeling is based on standard regression models, rather than blackbox

Regression models, with straightforward interpretation, are the most important statistical techniques used in analytical epidemiology

Justification of using the Lasso

1. Good compromise between traditional and data-driven approaches: modeling is based on standard regression models, rather than blackbox

Regression models, with straightforward interpretation, are the most important statistical techniques used in analytical epidemiology

2. Multivariate modeling approach, rather than separate univariate tests

Controlling for potential confounding is an critical point in epidemioloy

Justification of using the Lasso

1. Good compromise between traditional and data-driven approaches: modeling is based on standard regression models, rather than blackbox

Regression models, with straightforward interpretation, are the most important statistical techniques used in analytical epidemiology

2. Multivariate modeling approach, rather than separate univariate tests

Controlling for potential confounding is an critical point in epidemioloy

3. Stability: the variable selection process is not discrete (all regression coefficients are “smoothly” shrunk towards 0)

The procedure is more stable than stepwise selection procedures

Justification of using the Lasso

1. Good compromise between traditional and data-driven approaches: modeling is based on standard regression models, rather than blackbox

Regression models, with straightforward interpretation, are the most important statistical techniques used in analytical epidemiology

2. Multivariate modeling approach, rather than separate univariate tests

Controlling for potential confounding is an critical point in epidemioloy

3. Stability: the variable selection process is not discrete (all regression coefficients are “smoothly” shrunk towards 0)

The procedure is more stable than stepwise selection procedures

4. Sparsity: the L_1-penalty allows to produce some coefficients that are exactly 0, thus estimation and variable selection performed simultaneously (unlike ridge)

We expect that only few drugs affect driving

Justification of using the Lasso

5. Only detects effect sizes > the noise level [Bunea08].

We expect relevant but small drugs effect sizes. Shared by any selection method.

Justification of using the Lasso

5. Only detects effect sizes > the noise level [Bunea08].

We expect relevant but small drugs effect sizes. Shared by any selection method.

6. Convexity: L_1-penalty yields a convex optimization problem $\Rightarrow$ algorithms remain approachable and mathematical guarantee of a unique global max. Some fast and efficient algorithms compute the entire regularization path of solutions: for all λ, instead of varying λ on a grid (approximately for logistic regression)

We can handle large databases (N~70,000)

Justification of using the Lasso

5. Only detects effect sizes > the noise level [Bunea08].

We expect relevant but small drugs effect sizes. Shared by any selection method.

6. Convexity: L_1-penalty yields a convex optimization problem $\Rightarrow$ algorithms remain approachable and mathematical guarantee of a unique global max. Some fast and efficient algorithms compute the entire regularization path of solutions: for all λ, instead of varying λ on a grid (approximately for logistic regression)

We can handle large databases (N~70,000)

7. Adapted to challenging estimation situations: when the nb of predictors is high (even p>N) or in case of moderate correlation. But selects only few of the relevant variables if highly correlated.

We have p~400; some drugs usually prescribed together

A solution if correlation is high: elastic net (L1+L2) [Zou05]

Justification of using the Lasso

8. Prediction-based criteria (CV, AIC) to choose λ select the relevant predictors but also a few irrelevant ones. No consequences on prediction but not the best selection. Research on consistent criteria is open (BIC in high-dimensionality, minimization of the expected nb of false positives, consistent versions of CV).

Our main goal is not prediction but identification of drugs impacting driving

A solution: **Bolasso** [Bach08] (only predictors frequently chosen by the lasso over the bootstrap samples are selected)

Justification of using the Lasso

8. Prediction-based criteria (CV, AIC) to choose λ select the relevant predictors but also a few irrelevant ones. No consequences on prediction but not the best selection. Research on consistent criteria is open (BIC in high-dimensionality, minimization of the expected nb of false positives, consistent versions of CV).

Our main goal is not prediction but identification of drugs impacting driving

A solution: **Bolasso** [Bach08] (only predictors frequently chosen by the lasso over the bootstrap samples are selected)

9. Research on uncertainty measures is open (discontinuity of the sampling distribution problematic) [Pötscher10,Chatterjee11]

Because of potential bias, epidemiologists are attached to confidence intervals (rather than p-values or other “binary” decision criteria).

Justification of using the Lasso

8. Prediction-based criteria (CV, AIC) to choose λ select the relevant predictors but also a few irrelevant ones. No consequences on prediction but not the best selection. Research on consistent criteria is open (BIC in high-dimensionality, minimization of the expected nb of false positives, consistent versions of CV).

Our main goal is not prediction but identification of drugs impacting driving

A solution: **Bolasso** [Bach08] (only predictors frequently chosen by the lasso over the bootstrap samples are selected)

9. Research on uncertainty measures is open (discontinuity of the sampling distribution problematic) [Pötscher10,Chatterjee11]

Because of potential bias, epidemiologists are attached to confidence intervals (rather than p-values or other “binary” decision criteria).

10. Linearity of the Logit function: We have only binary predictors

Roadmap

Implementation

Case-Control study

Glmnet R-package [Friedman10]. Practical options:

→ Faster than competing methods (appreciable when bootstrapping)

→ Sparse data matrices stored in sparse format (73% unexposed to any drug)

→ Penalized (drugs) and unpenalized (adjustment) variables can be specified

→ Model selection is performed by several criteria (we used 10-fold CV AUC [Avalos12b])

Implementation

Case-Crossover study

For moderate N (<3000), algorithms based on the stratified discrete-time Cox proportional hazards model compute lasso estimates of conditional logistic regression [Avalos12a]: we can use the penalized R-package [Goeman10]

For large datasets, **clogitLasso** R-package [Avalos13a]:

→ Algorithm based on an iteratively re-weighted least squares (IRLS) algorithm and depending on the efficient for large N package lassoshooting [Jornsten11]

→ Model selection is performed by several criteria (we used 10-fold likelihood-based CV, accounting for data dependence)

→ Bootstraping (accounting for data dependence): Bolasso, CI

Roadmap

1. Context
2. CESIR project
3. Design studies
4. Statistical analysis using the Lasso: justification, limitations
5. Implementation
6. Pharmacoepidemiological results
7. Conclusions

(Some) Pharmacoepidemiological results

Some results are consistent with literature [Avalos12b, Avalos13b] : benzodiazepine hypnotics (N05CD), anxiolytics (N05BA, N05BC)

Some results complete results of other studies: Antiepileptics (N03AF, N03AG, N03AX), antidepressants (N06AB, N06AX), drugs used in opioid dependence (N07BC) and drugs used in diabetes (A10AD)

Cardiovascular drugs (C10BA, C10BX) are markers of cardiovascular diseases which themselves may increase the risk of accident involvement.

Antithrombotic agents (B01AB): cardiovascular events may occur when the treatment is interrupted. This result oriented new researches.

Conclusion

→ National registries exist. Exceptional sources of information. But, extraction of reliable and useful information is not an obvious process

→ When the data are (finally!) available, analysis needs to be addressed using appropriate statistical and computational techniques

→The Lasso appears to be a good compromise between traditional and data-driven approaches for analyzing high-dimensional and/or large epidemiological data

→ Some theoretical properties of lasso-type estimators are now well established, others are, however, still heuristic-based (uncertainty measures)

References

[Bach08] Bolasso: model consistent lasso estimation through the bootstrap. ICML, Helsinki, Finland, 2008.

[Barbone98] Association of road-traffic accidents with benzodiazepine use. Lancet 1998; 352:1331–1336.

[Bunea08] Honest variable selection in linear and logistic regression models via l1 and l1 + l2 penalization. Electron. J Statist 2008

[Chatterjee11] Bootstrapping Lasso estimators. J Am Stat Assoc. 2011

[Engeland07] Risk of road traffic accidents … with the prescription of drugs: a registry-based cohort study. Ann Epidemiol. 2007

[Friedman10] Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010

[Gibson09] Use of self-controlled … use of prescription medications and the risk of motor vehicle crashes. Am J Epidemiol 2009

[Goeman10] L1 penalized estimation in the Cox proportional hazards model. Biometrical Journal 2010

[IMS08] IMS Health. Enquête Permanente sur la Prescription Médicale (EPPM). Danbury, CT: IMS Health; 2005–2008

[Jaro95] Probabilistic linkage of large public health data files. Stat Med 1995

[Jornsten11] Network modeling of the transcriptional …copy number aberrations in glioblastoma. Molecular Systems Biology 2011

[Karjalainen12] Substance use and social, health and safety related factors among fatally injured drivers. Accid Anal Prev. 2012

[Laumon05] Cannabis intoxication and fatal road crashes in France: population based case-control study. BMJ. 2005

[Maclure91] The case-crossover design: a method for studying transient effects on the risk of acute event. Am J Epidemiol 1991

[Pötscher10] Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. Electron J Stat. 2010

[Robertson94] Responsibility analysis a methodology to study the effects of drugs in driving. Accid Anal Prev 1994

[Tibshirani96] Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 1996

[Zou05] Regularization and variable selection via the elastic net. J. Royal. Statist. Soc B 2005

CESIR References

[Avalos12a] Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood. Stat. Med. 2012

[Avalos12b] Prescription-drug-related risk in driving comparing conventional and Lasso shrinkage logistic regressions. Epidemiology 2012

[Avalos13a] clogitLasso: an R package for high-dimensional analysis of matched case-control and case-crossover data. CIBB 2013

[Avalos13b] Variable selection on large case-crossover data: application to a registry-based study of prescription drugs and road traffic crashes. Pharmacoepidemiol Drug Saf 2013

[Blazejewski12] Factors associated with serious traffic crashes: a prospective study in southwest France. Arch Intern Med. 2012

[Corsenac12] Road traffic crashes and prescribed methadone and buprenorphine: a French registry-based case-control study. Drug Alcohol Depend 2012

[Orriols10] Prescription medicines and the risk of road traffic crashes: A French registry-based study. PLoS Med 2010

[Orriols11] Benzodiazepine-like hypnotics and the associated risk of road traffic accidents. Clin Pharmacol Ther 2011

[Orriols12] Risk of injurious road traffic crash after prescription of antidepressants. J Clin Psychiatry. 2012

[Orriols13] Prescription of antiepileptics and the risk of road traffic crash. J Clin Pharmacol. 2013