Bayesian Logistic Regression for Medical Claims Data

Ivan Zorych, Patrick Ryan, David Madigan

Outline

- Drug safety data
- Problems with 2x2 approaches
- Bayesian logistic regression for SRS data
- Bayesian logistic regression approach to observational data

Typical Entry in SRS database

Age	*Sex*	*Drug 1*	*Drug 2*	…	*Drug 15000*	*AE 1*	*AE 2*	…	*AE 16000*
42	Male	No	Yes	…	No	Yes	No	…	Yes

- SRS datasets resemble spreadsheets with up to millions of rows (one per report) and tens of thousands of columns

Existing Methods

- Multi-item Gamma Poisson Shrinker (MGPS)
  - US Food and Drug Administration (FDA)
- Bayesian Confidence Propagation Neural Network
  - WHO Uppsala Monitoring Centre (UMC)
- Proportional Reporting Ratio (PRR)
  - UK Medicines Control Agency (MCA)
- Reporting Odds Ratios and Incidence Rate Ratios
  - Other national spontaneous reporting centers and drug safety research units

Different Measures

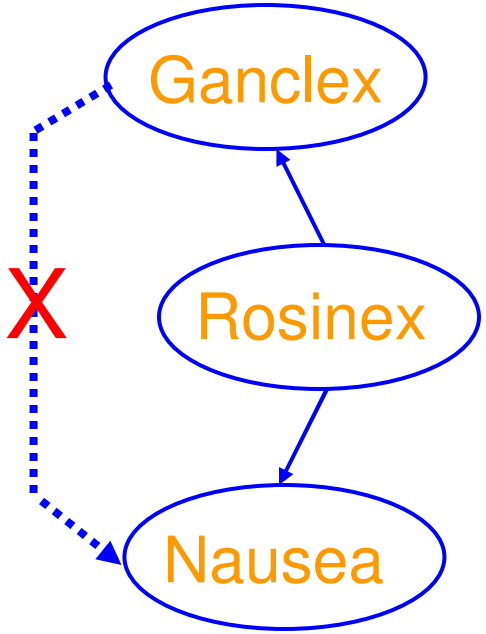
	AE	*All other AE*
Drug	*a*	*b*
All other drugs	*c*	*d*

Measure of Association	*Formula*	*Probabilistic Interpretation*
RR Relative Risk*	$\frac{a * (a+b+c+d)}{(a+c) * (a+b)}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae)}$
PRR Proportional Reporting Ratio	$\frac{a / (a+b)}{c / (c+d)}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae \mid \neg drug)}$
ROR Reporting Odds Ratio	$\frac{a / c}{b / d}$	$\frac{\Pr(ae \mid drug) / \Pr(\neg ae \mid drug)}{\Pr(ae \mid \neg drug) / \Pr(\neg ae \mid drug)}$
Information Component	$Log_2 \frac{a * (a+b+c+d)}{(a+c) * (a+b)}$	$\log_2 \frac{\Pr(ae \mid drug)}{\Pr(ae)}$

Innocent bystander problem

- Contingency table analysis ignores effects of drug-drug association on drug-AE association

	Rosinex		No Rosinex		Total	
	Nausea	No Nausea	Nausea	No Nausea	Nausea	No Nausea
Ganclex	81	9	1	9	82	18
No Ganclex	9	1	90	810	99	811
OR	1		1		37.3	



Logistic Regression

$$\log \frac{\Pr(Nausea)}{\Pr(Not\ Nausea)} = \beta_0 + \beta_1 \times \text{Rosinex} + \beta_2 \times \text{Ganclex}$$

- SAS and R-package give the following estimations:

beta1=4.4, beta2=0

Adjusted odds ratio exp(beta2)=1 indicates no association between Ganclex and Nausea.

Regression and health data

- Drug safety data sets are sparse
- Typical AERS report contains just a few drugs and a few adverse events
- Dependent variable in the regression:

  Y=1 if AE is present, 0 otherwise;
- Number of independent variables = number of drugs + sex/age/year info

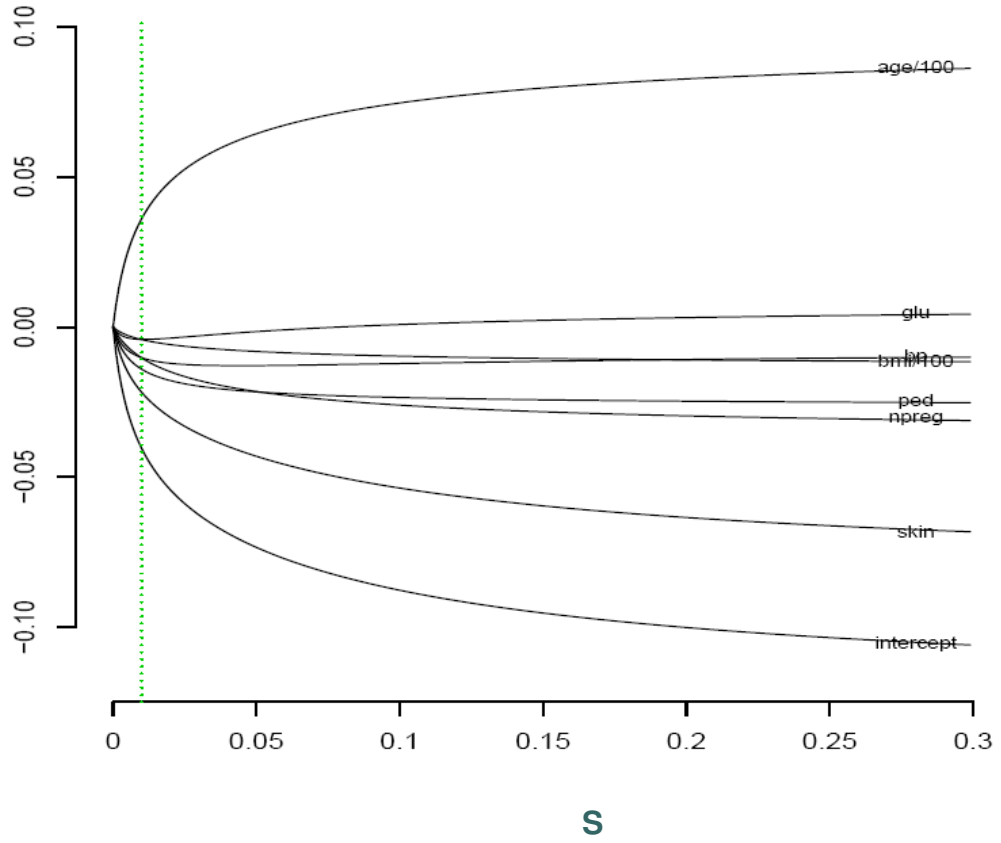
Ridge logistic regression*

- Maximum likelihood
- and restrictions on coefficients:

$$\sum_{j=1}^{P} \beta_j^2 \leq s$$

*Hoerl and Kennard (1970).

Profiles of the regression coefficients



Lasso*

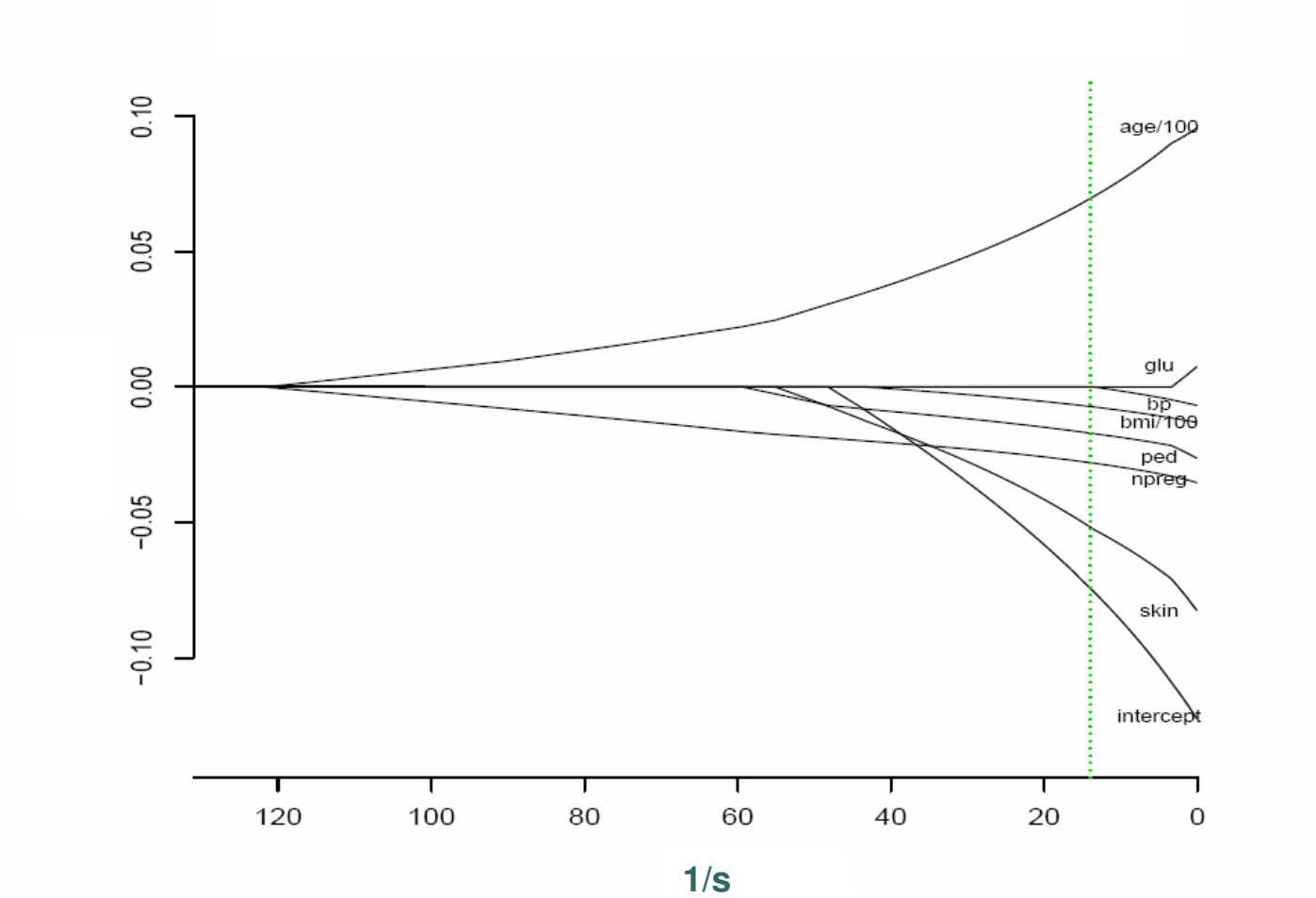
- Maximum likelihood
- and restrictions on coefficients:

$$\sum_{j=1}^{P} \left| \beta_j \right| \leq s$$

- Does subset selection by shrinking some coefficients to zero

*Tibshirani (1996)

Profiles of the regression coefficients



Bayesian Regression

- Two shrinkage methods
  - Ridge regression - Gaussian prior

$$p(\beta_j \mid tau) \sim N(0, tau)$$

  - Lasso regression - Laplace prior

$$p(\beta_j \mid \lambda) = \lambda/2 \exp\{-\lambda |\beta_j|\}$$

- Choosing hyperparameter λ
  - Decide how much to shrink
  - Cross-validation: choose prior to fit left-out data

Bayesian Logistic Regression

- Software: Bayesian Binary Regression (BBR)*

  http://www.stat.rutgers.edu/~madigan/BBR/

  http://www.bayesianregression.org/

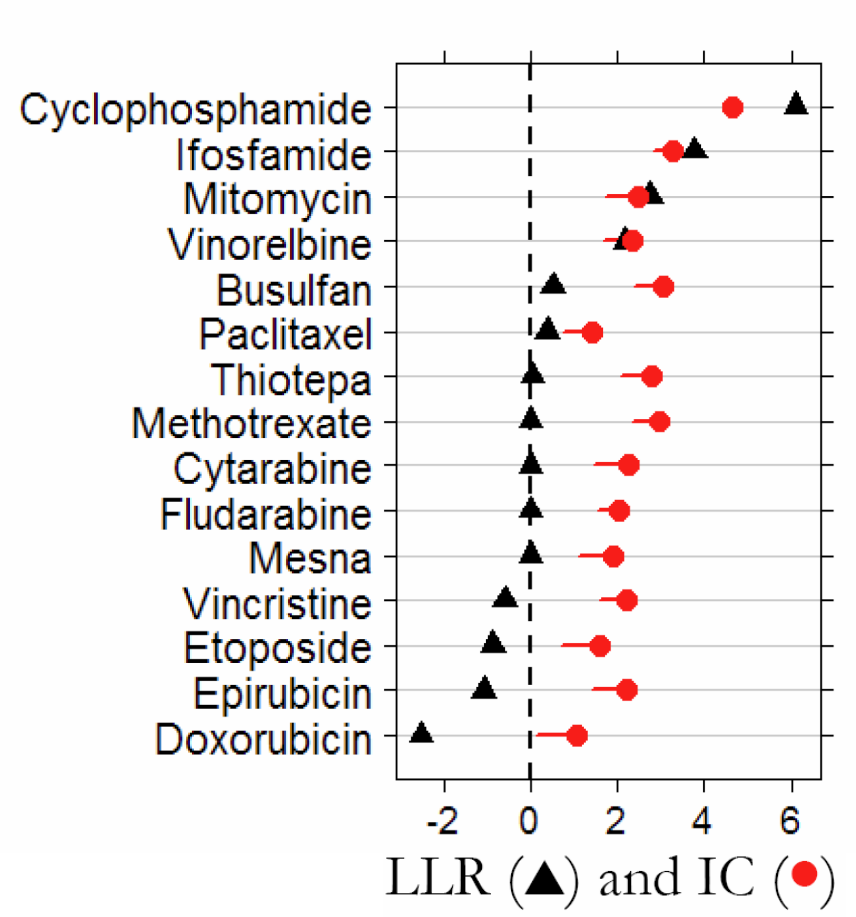
  - Two priors: Gaussian and Laplace
  - Hyperparameter choice: fixed, CV, etc.
  - Handles millions of predictors efficiently

*D. Madigan et al., (2007), Large-Scale Bayesian Logistic Regression for Text Categorization, Technometrics, vol.49, #3.

Vioxx / Transient ischemic attack

Year	N of reports	EBGM rank	BBR rank (Normal priors with CV)
1999	1	593	545
2000	26	351	70
2001	60	316	33
2002	100	431	55
2003	130	459	51

Confounding, real AERS data*



- ADR: hemorrhagic cystitis, diffuse inflammation of the bladder leading to dysuria, hematuria, and hemorrhage;
- ADR most often seen in female cancer patients as a complication of therapy;
- Drugs: anticancer drugs and mesna;
- Mesna is an adjuvant used in cancer chemotherapy involving cyclophospamide and ifosfamide.

*Caster, Noren, Bate, Madigan, 2010

Masking

- Typical DP measures are based on

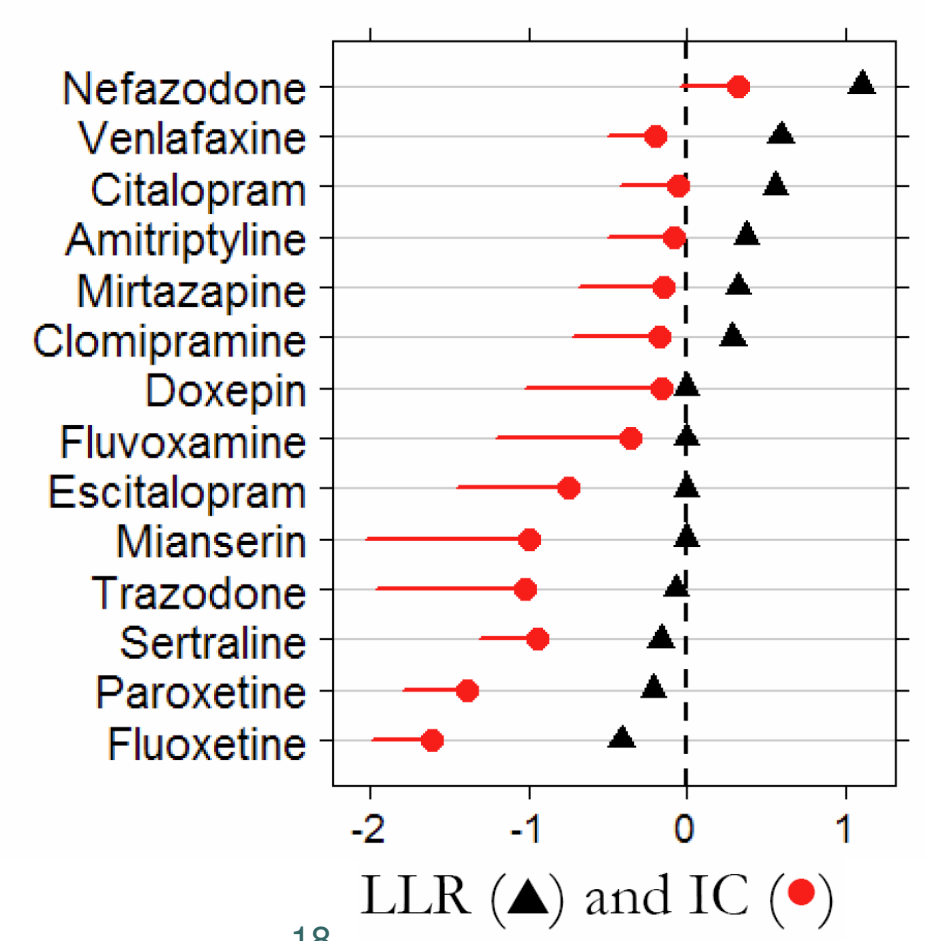
$$\frac{\Pr(AE|Drug)}{\Pr(AE)}$$

- Masking: effect when the background rate of the ADR, $\Pr(AE)$, is distorted due to massive reporting with other drug(s);

- Example: Rhabdomyolysis and Cerivastatin (Baycol, Lipobay) is a synthetic member of the class of statins;

- Cerivastatin was voluntarily withdrawn from the market worldwide in 2001 due to reports of fatal rhabdomyolysis.

Masking, real AERS data*



- ADR: Rhabdomyolysis, rapid breakdown of skeletal muscle;
- Drugs: a set of anti-depressant drugs;

*Caster, Noren, Bate, Madigan, 2010

Weakness of SRS Data

- Passive surveillance
  - Underreporting
- Lack of accurate “denominator”, only “numerator”
  - “Numerator”: No. of reports of suspected reaction
  - “Denominator”: No. of doses of administered drug
- No certainty that a reported reaction was causal
- Missing, inaccurate or duplicated data

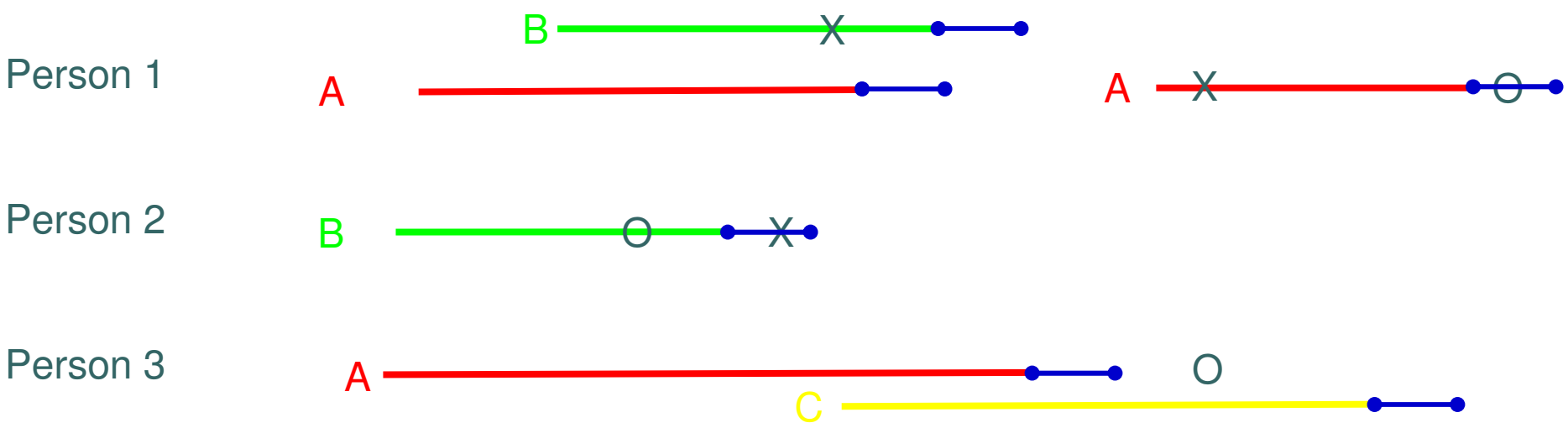
Longitudinal observational data

- Health claims records, electronic medical records
- Information on drug prescriptions, doctor visits, hospitalization

BLR and observational data

- It is relatively easy to apply Bayesian regression to AERS/SRS data.
- Situation with temporal data is more complicated. We need to create predictors for the regression analysis.

Logistic regression for observational data



prevalent X:

Y	A	B	C
1	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1

incident X:

Y	A	B	C
1	1	1	0
1	0	1	0
0	0	0	1

BLR parameter file

CONDITION_TYPE_PREVALENT_1_INCIDENT_2: 1
PRIOR_TYPE_LAPLACE_1_NORMAL_2: 1
PRIOR_VARIANCE: 1
INCLUDE_AGE_0_(NO)_OR_1_(YES): 1
INCLUDE_SEX_0_(NO)_OR_1_(YES): 1
SIZE_OF_THE_NO_CONDITION_PART: 1000000
RISK_WINDOW_IN_DAYS: 30
DRUG_PERSISTENCE_WINDOW_DAYS_0_OR_30: 30
CONDITION_PERSISTENCE_WINDOW_DAYS_0_OR_30: 30
NUMBER_OF_SPLITS: 5
DAYS_FROM_THE_DRUG_ERA_START: 1
DRUG_ERA_TABLE: OMOP_DRUG_ERA
CONDITION_ERA_TABLE: OMOP_CONDITION_ERA

Simulated Data

- 10 000 000 persons
- 5000 drugs
- 4519 conditions; 519 of conditions are ‘indications’ that don’t have any causal relationship with drugs
- 22,595,000 drug-condition pairs;

Computational side

- Each condition requires fitting of a separate regression
- Parallelization via SAS/CONNECT
- Amazon Cloud:

  68.4 GB of memory

  26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each)

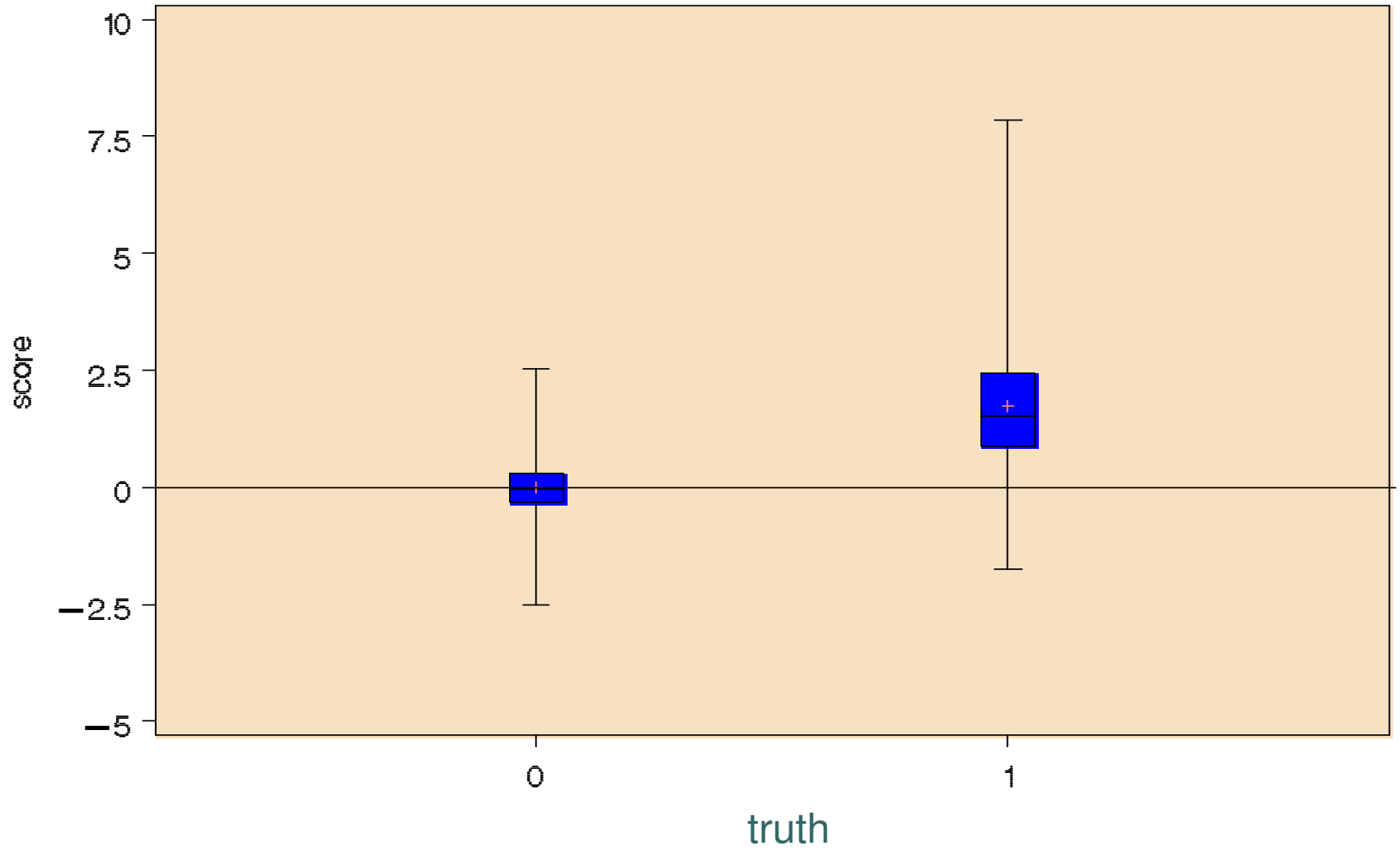
  1690 GB of local instance storage
- Typical performance (OSIM on the cloud): 12 hours on the cloud (24 parallel runs);
- OMOP stat server: Sun M5000 6X2

  2.14Ghz CPU(s)

  32G Memory

  48 Gb Swap space
- Code available from http://omop.fnih.org/MethodsLibrary

Scores for the OSIM data

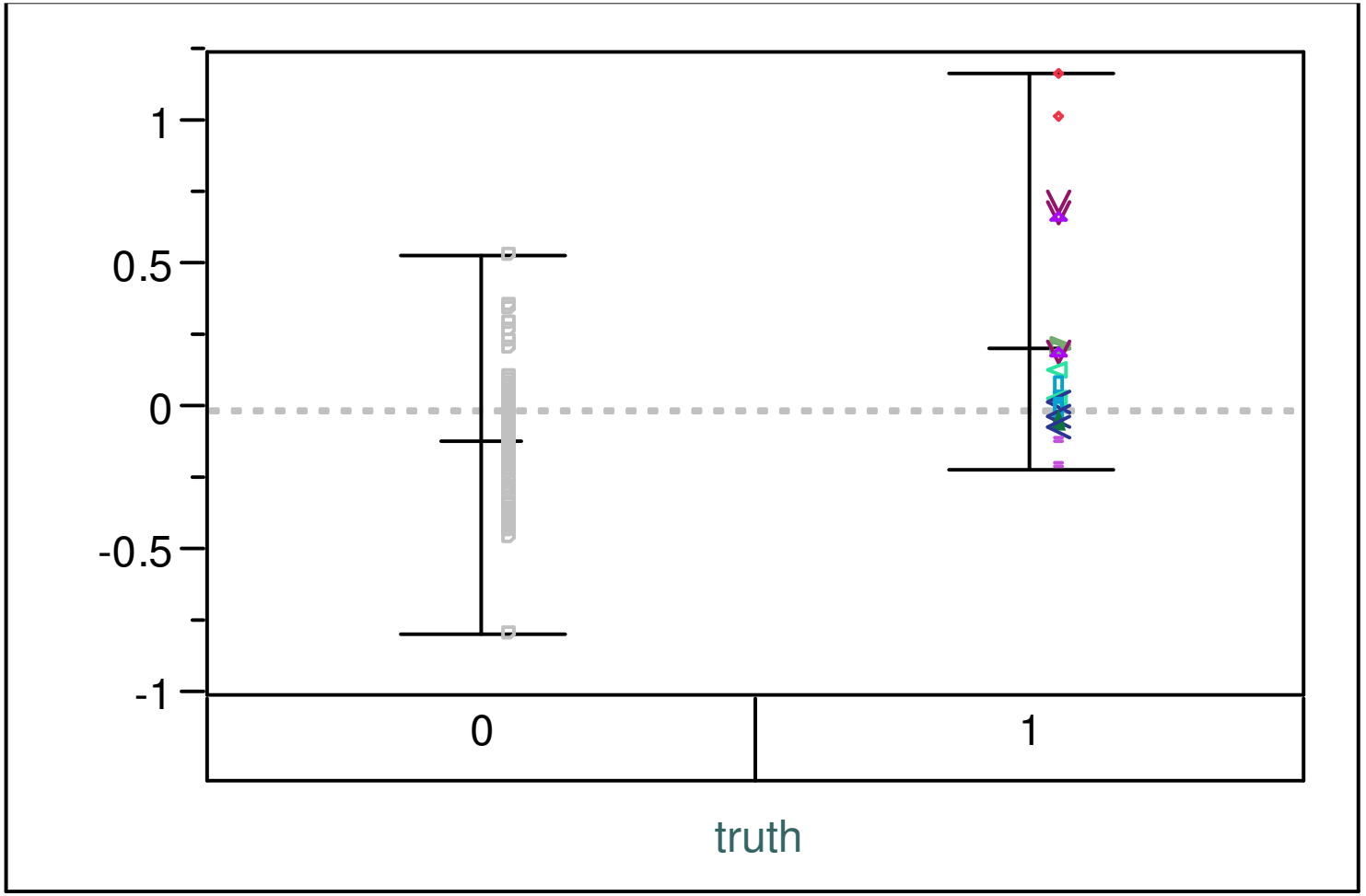


OMOP cup and BLR on OSIM (Simulated Data)

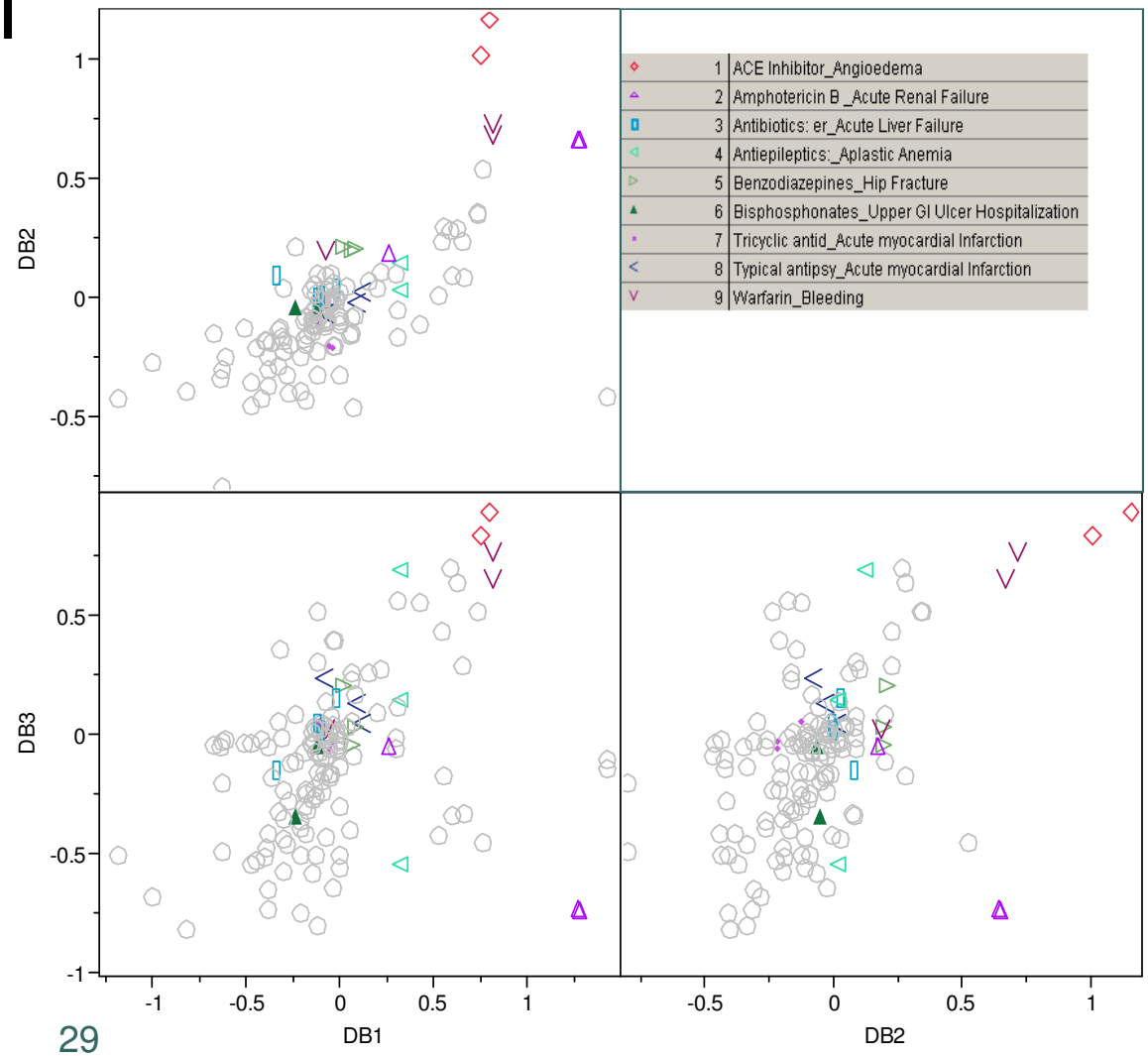
Participant 1	0.2662359
Participant 2	0.2570616
Participant 3	0.2569417
Participant 4	0.2569404
Participant 5	0.2568678
BLR	0.2557354 (same run, exclude ‘indications’)
Participant 6	0.2483813
Participant 7	0.2483137
Participant 8	0.2358521
BLR	0.2356831 (run: c2p2v1b1s0dw0cw0a0sx0)
…	
Random	0.0157622

MAP score

Real DB (preliminary data)



3 Real Databases



Discussion of Logistic Method

- Advantages over low-dimensional tables
  - Correct confounding and mask effect
  - Analyze multiple drugs/vaccines simultaneously
- Limitations
  - Build separate model for each AE
    - Ignore dependencies between AEs
  - Fail to adjust for unmeasured/unrecorded factors
    - health status, unreported drugs, etc.