Signal Detection in Pharmacovigilance

Comparative Performance of Several Statistical Approaches and a James-Stein Type Shrinkage Estimation Strategy in a Logistic regression Model

Lihua An
Statistics Canada
McLaughlin Centre for Population Health Risk Assessment,
University of Ottawa

CSEB/APHEO 2009 Joint Conference, Ottawa

Contents

- Overview of the statistical approaches in signal detection
  - Relative Reporting Ratio (RRR)
  - Reporting Odds Ratio (ROR)
  - Empirical Bayes Gamma-Poisson Shrinker (GPS)
  - Information Component (IC)
  - Logistic Regression
- Comparison of the performance of these methods in detecting drug-event association using simulation
- Extension of these methods to detecting multi-item association
- Comparative performance of these methods in detecting drug-drug interaction, confounding due to co-medication and masking
- A James-Stein type shrinkage estimation strategy in a Bayesian logistic regression model

Spontaneous Reporting Systems

- Objectives: To explore for Drug-Event Associations and to detect signals
- Limitations:
  - Under reporting and multiple reporting
  - Unknown denominator
  - Unknown causality
  - Lack of control group
  - Missing and incomplete data
- Comparisons and conclusions must be made with caution

A 2 x 2 Table of Report Counts

For every pair of drug (D_i) and event (E_k) of interest,

	Reports With drug i	**Reports w/o drug i**	**total**
Reports With event k	$n_{ik} = a$	b	a+b
Reports w/o event k	c	d	c+d
total	a+c	b+d	N_{total} = a+b+c+d

Different Approaches

- $RRR_{ik} = n_{ik}/e_{ik} = a(a+b+c+d)/(a+b)(a+c)$ where $e_{ik} = N_{total} \times P_i \times Q_k$
  - P_i and Q_k are the probabilities of occurrence of a report of drug i and occurrence of a report of event k, respectively
- *Reporting Odds Ratio of event k with exposure to drug I relative to without exposure to drug I:*

  $ROR_{k,i} = ad/bc$
- GPS method:
  - Assume $n_{ij} \sim$ Poisson(μ_{ij}) and estimate $\lambda_{ij}=\mu_{ij}/e_{ij}$
  - Assume prior distribution of λ_{ij} is mixture of two Gamma distributions and estimate the 5-parameter prior from all the (n_{ik}, e_{ik}) pairs
  - EBGM (empirical bayes geometric mean of the posterior distribution of λ) is used as a measure
  - Rank cells by EB05=lower 5% point of the posterior distribution
- $IC=\log_2\{P(AE,Drug)/[P(AE)P(Drug)]\}$
  - Prior probabilities of P(AE) and P(Drug) are modeled using beta distributions
  - Prior distributions for P(AE,Drug) are represented by a Dirichlet distribution, with parameters based on the data
  - IC95, the lower 95% confidence limit is usually calculated

Different Approaches

- The logistic regression model $$z_i = \ell n \frac{P(y_i = 1)}{1 - P(y_i = 1)} = (\boldsymbol{\beta}^*)^{\mathrm{T}} \mathbf{x} \quad (i = 1, \ldots, N)$$

  where

$$y_i = \begin{cases} +1 \; if \;\; the \; event \; of \;\; interest \; is \; involved \; in \; report \; i \\ 0 \; if \;\; the \; event \; of \;\; interest \; is \; not \; involved \; in \; report \; i \end{cases}$$

  and

$$x_{ij} = \begin{cases} 1 \quad if \;\; drug \;\; j \; is \; involved \; in \; report \; i \\ 0 \quad if \;\; drug \;\; j \; is \; not \; involved \; in \; report \; i \end{cases}$$

  (i=1, . . ., N; j= 1, . . ., J)

- Two ways to use the logistic regression model for signal detection:
  - β_j is the log odds ratio of the event when x_j is coded as 0 or 1.
  - Test the significance of the slope β_j. May include other demographic variables, e.g., age, gender in the model.

Comparison of the Performance using Simulated Data and ROC Curves

- Receiver operating characteristic (ROC) curves were used as a measure of performance

(TPR=True Positive Rate=sensitivity)

vs.

(FPR = False Positive Rate = 1-specificity)

- Sensitivity– proportion of actual positives which are correctly identified as such
- Specificity – proportion of negatives which are correctly identified

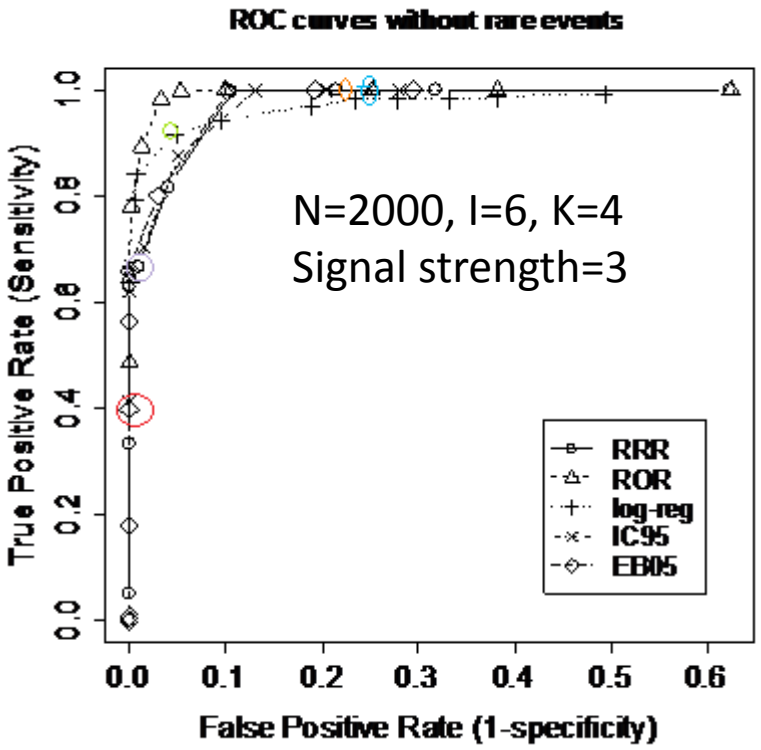
Data Generation

- Suppose we are only interested in certain drugs and certain AEs that may be associated with these drugs.
- Dropping categories and analyzing a subset of the data base is appropriate. Reduced Database is more homogeneous.
- I = # of drugs of interest, K-1=# of events of concern, Kth event=all AEs not of our concern.

  I= (6, 10), K= (4, 8)
- Number of reports N: 1,000 to 10,000
- Background probabilities: P_i : 0.01 to 0.5,

  Q_k : 0.001 to 0.25
- Signal Strength: 2 to 6 times

Comparison of different Methods in Detecting Drug-event Associations



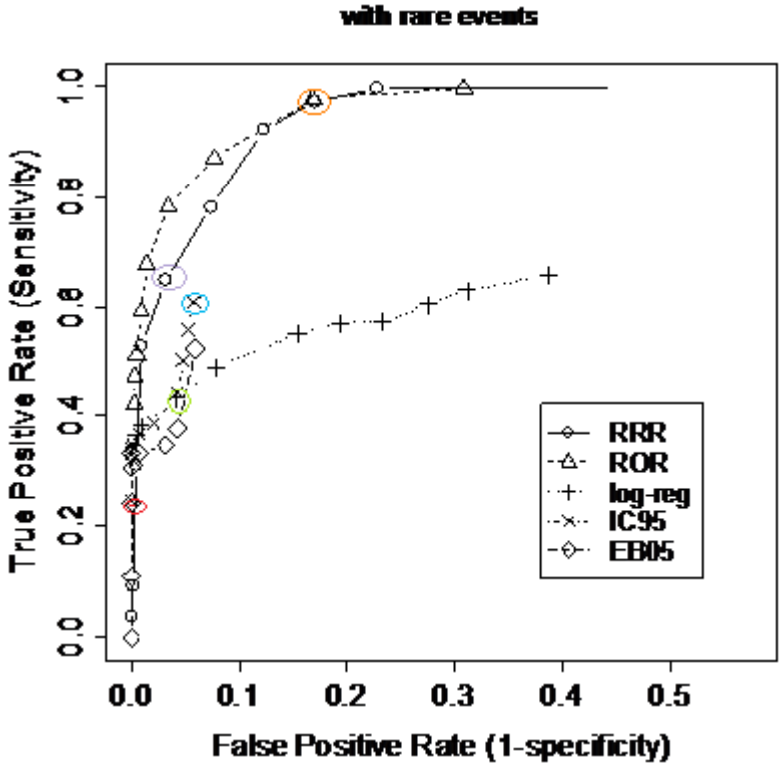
- When conventional cut-off values are used

	Cut-off	TPR	FPR
EB05	2	0.40	0
RRR	2	0.67	0.01
LogReg	0.05	0.92	0.05
IC95	0	1	0.28
ROR	2	1	0.20

Conventional cut-off values are not always optimal!

Comparison of different Methods in Detecting Drug-event Associations

ROC curves with Rare Events



- When conventional cut-off values are used

	Cut-off	TPR	FPR
EB05	2	0.24	0
LogReg	0.05	0.43	0.04
IC95	0	0.61	0.06
RRR	2	0.65	0.03
ROR	2	0.94	0.13

EBGM is the most conservative in both cases.

Logistic regression is affected the most by the occurrence of rare events.

Extension to Detecting Multi-Item Associations

- Consider the case of an item triplet, e.g. 2 drugs (D_i , D_j) and an event $E_{k.}$
- $RRR_{ijk} = n_{ijk}/e_{ijk}$ where e_{ijk} is based on independence model
- Suppose the item set (drug I, drug j, and event k) is unusually frequent
- Is this due to one or more of the pairs (ij, ik, jk) or
- Is this drug-drug interaction?

Detection of Drug-Drug Interaction

- MGPS:

  $EXCESS2=(EBGM_{ijk} * e_{ijk})-e_{All2F}$
  - number of cases over and above those that can be explained by the pairwise associations
  - e_{All2F} is predicted count of all-two-factor model
- Interaction Relative Reporting Ratio (counter part RRR):

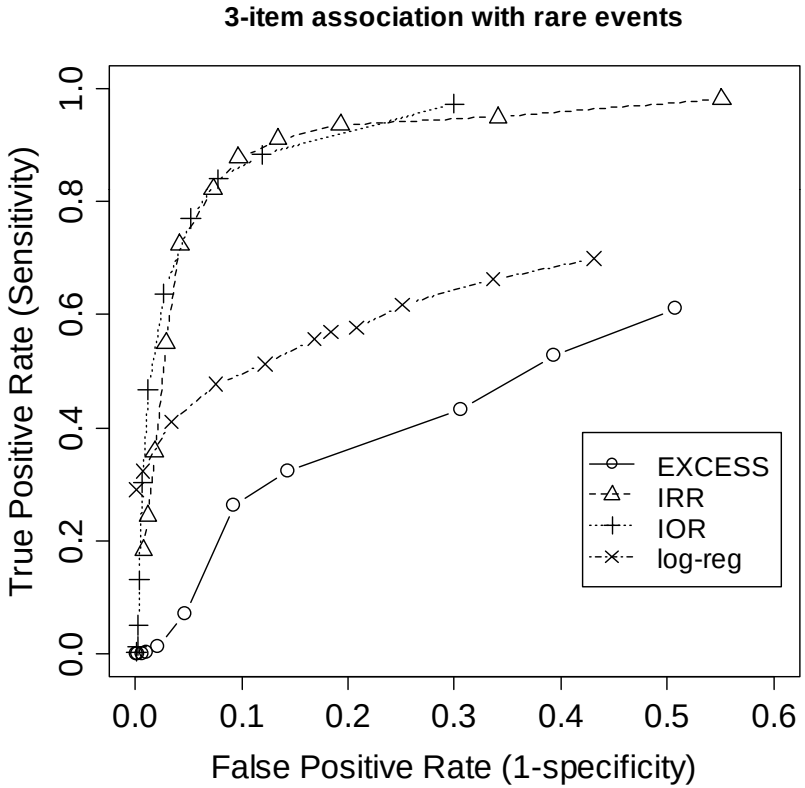
  $IRR_{ijk} =(n_{ijk} -e_{All2F})/e_{ijk}$
- **Interaction Reporting Odds Ratio** (counter part ROR):

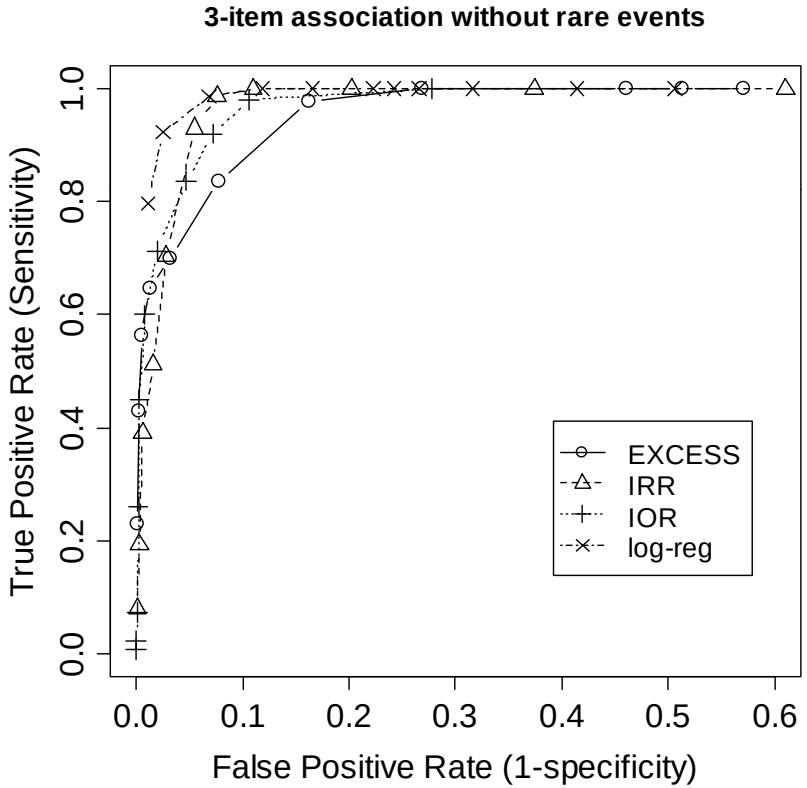
  $\mathbf{IOR_{k,ij}} = ROR_{k,ij} / \max(1, ROR_{k,i}, ROR_{k,j})$ where

  $ROR_{k,ij} = P_{k|i,j}(1-P_{k|-i,-j}) / P_{k|-i,-j}(1-P_{k|i,j})$ is the reporting odds ratio of event k with exposure to both drugs I and j relative to without exposure to both.

  $P_{k|i,j}$ ($P_{k|-i,-j}$) is the probability of event *k* when a patient is exposed (not exposed) to both drug i and drug j.
- Logistic regression: Include interaction terms in the model

Performance Comparison In Detecting Multi-Item Associations





Masking

- When studying a particular drug-event association, it is assumed that all reports except for the drug of interest constitute the general background reporting of the AE.
- If there are one or more drugs with highly frequent reports on the same AE, the background reporting becomes substantial, which increases the expected number of reports on the pair currently under study.

Confounding

- Drug A truly causes an AE, drug B does not, but B is frequently co-medicated with drug A.
- A and B both will be signalled as likely cause for the AE.
- Which of the methods can better deal with this situation?

Identifying Masking

Sig1/sig2	EB05	RRR	ROR	IC95	LogReg
2	No	No	No	No	Yes
1.5	No	No	No	Yes	Yes

Other affecting factors :

- size of the data
- cut-off values

The simulation result only shows their relative performance under certain conditions.

Dealing with Confounding

Co-medication	EB05	RRR	ROR	IC95	LogReg
90%	No	Yes	No	No	Yes
50%	No	Yes	No	Yes	Yes
25%	No	Yes	Yes	Yes	Yes

Other affecting factors :

- size of the data
- strength of the signal
- cut-off values

A Bayesian Logistic Regression Model

- Genkin et al. considered a Bayesian Approach to logistic regression to deal with sparseness.
- It assumes Laplace priors for the distribution of β_j^* s:

$$p(\beta_j^* \mid \lambda_j) = \frac{\lambda_j}{2} \exp(-\lambda_j \mid \beta_j^* \mid)$$

- The log posterior density for **β*** is given by:

$$\ell(\boldsymbol{\beta}^*) = -\sum_{i=1}^{N} \ell n(1 + \exp(-(\boldsymbol{\beta}^*)^{\mathrm{T}} \mathbf{x_i} y_i)) - \sum_{j=1}^{d} (\ell n 2 - \ell n \lambda_j + \lambda_j \mid \beta_j^* \mid)$$

- The **β*** that maximizes $\ell(\boldsymbol{\beta}^*)$ is called the maximum a posteriori (MAP) estimate.

A James-Stein Type Shrinkage Estimation Strategy

- Many similar medical events exist simultaneously, but analyzed separately in SRS databases. (E.g., heart block, heart disorder, heart attack, …)
- It may be true that certain drugs increase the risk of these AE according to a similar biological mechanism.
- If this prior belief is true, we may pool these events to "borrow strength" from the different events to obtain improved estimates.

Shrinkage Estimation and Bayesian Logistic Model

- Suppose we are interested in the simultaneous associations between E adverse events and a fixed combination of k drugs.
- The logistic regression model

$$z_i = \ell n \frac{P(y_i = 1)}{1 - P(y_i = 1)} = (\boldsymbol{\beta}^*)^{\mathrm{T}} \mathbf{x} \quad (i = 1, \ldots, n) \quad (1)$$

  can be used to model each of E events separately.
- "Stack" the E models to obtain a general form

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

- Our prior belief can be expressed in terms of

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad (3)$$

  where the matrix **C** is an r x pE matrix of rank r, p=k+1.

Shrinkage Estimation and Logistic Model

- Let $\hat{\boldsymbol{\beta}}^{un}$ and $\tilde{\boldsymbol{\beta}}^{re}$ be the MAP estimates of $\boldsymbol{\beta}$ under the unrestricted model and restricted models, separately.
- The James-Stein type shrinkage estimator is

$$\hat{\boldsymbol{\beta}}^{JS} = \tilde{\boldsymbol{\beta}}^{re} + (1 - \frac{c}{F}) I(F \geq c)(\hat{\boldsymbol{\beta}}^{un} - \tilde{\boldsymbol{\beta}}^{re}) \quad (4)$$

- where c=(k-3)m/((k-1)(m+2)), m=n-2E,

$F_{(r,n-k)} = \frac{[\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}]'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}[\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}]}{rs^2}$ is a test statistic, and $I(F \geq c)$ is an indicator function

- James-Stein is a weighted average of $\hat{\boldsymbol{\beta}}^{un}$ and $\tilde{\boldsymbol{\beta}}^{re}$.
- When H_0 is likely to be true, the restricted estimate is given more weight.

A Simulation Study of the Relative Performance of the Estimators Using MSE as a measure

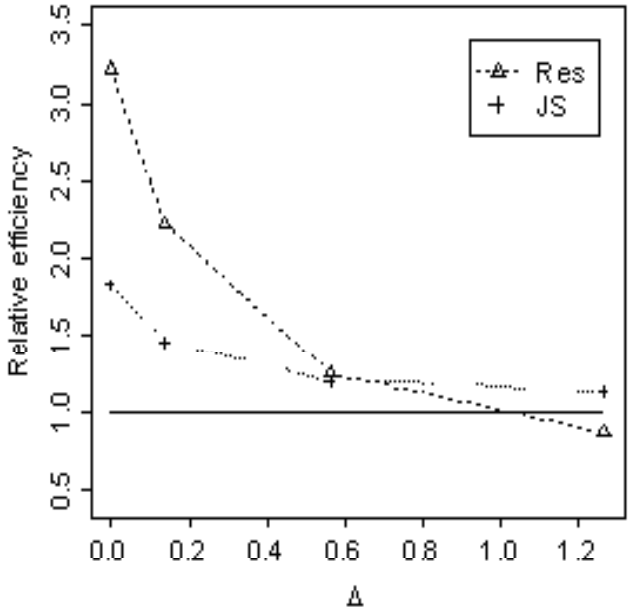
- In an *E*-sample setup, the mean squared error (MSE) of an estimator $\hat{\boldsymbol{\beta}}$ of β is defined as

$$MSE(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

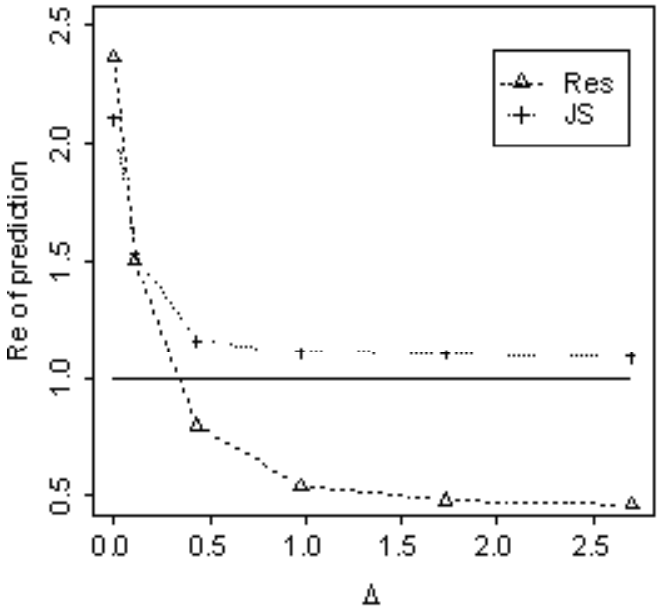
- MSE of the estimation and prediction based on different estimators were computed based on 5,000 Monte Carlo simulations with parameter setting as follows:
  - N=200, E=8, k=4
- $H_0 : \beta_{i1} = \beta_{i2} = \cdots = \beta_{iE}, \quad i = 1, \ldots, k$

A Simulation Study of the Relative Performance of the Estimators Using MSE as a measure

Parameter estimation



Prediction



Y-axis: Efficiencies of β^{re} and $\hat{\beta}^{JS}$ relative to $\hat{\beta}^{un}$

X-axis: A distance measure between the true parameter value and the value under null hypothesis. E.g., Δ=0 implies H0 is true.

A James-Stein Shrinkage Estimation Strategy

- Simulation shows that the maximum efficiency of both $\tilde{\boldsymbol{\beta}}^{re}$ and $\hat{\boldsymbol{\beta}}^{JS}$ occur at Δ=0.
- As Δ increases, the relative efficiency (RE) of both decreases.
- While the RE of drops rapidly below one, the efficiency of $\tilde{\boldsymbol{\beta}}^{JS}$ approaches but always stays above one, indicating that it dominates the unrestricted estimator throughout the entire parameter range.
- This same pattern is observed in the relative efficiencies of prediction.
- James-Stein estimator is an improved estimator over the unrestricted estimator when estimating several parameters simultaneously.

Conclusions and Discussion

- A simulation-based testing environment for evaluating signal detection methods is used in this study. The sensitivities and specificities of the core statistical measures in five popularly used approaches are assessed under different situations based on simulated data.
- Simulation shows that the conventional cut-offs are often not optimal.
- The performances of the measures in detecting multi-item association and identifying masking and confounding are also evaluated and some general guides are provided.
- The validation of these data mining tools in an absolute sense is an unreachable goal due to the limitations of the SRS. Also, the simulated data may not contain some features in the real SRS databases. The objective of this study is to compare the methods and not to draw substantive conclusions.
- A James-Stein type shrinkage estimation strategy is proposed to borrow strength across medically related events to improve the estimation and prediction in a logistic regression model.
- Our initial results suggest that this new approach holds considerable promise for use in signal detection.

Questions?