An Introduction to the Statistical Methods for Signal Detection in Pharmacovigilance

Presented By: Chris Gravel

Data Collection Methods

-Spontaneous Reporting Systems (SRS) gather data passively from health care professionals and consumers

-Some limitations of this method include
- -under/over reporting
- -multiple drugs/adverse drug reactions (ADR)
- -limited background information
  - -dose, exposure time etc.

Assumptions for this discussion

-to discuss the simplest form of these statistical methods, we will assume
- -single drug and ADR
- -no knowledge of the reporting probability and background information

The data

-The data can be gathered into a large dimensional contingency table where the *i*-th row represents a particular drug, D_i , and the *j*-th column represents a particular ADR, A_j .

-The data can then be condensed into a 2 x 2 table by summing over the columns and rows:

	Event A_j	Other Events A_j^C	Total
Drug D_i	n_{ij}	$n_{i.} - n_{ij}$	$n_{i.}$
Other Drugs D_i^C	$n_{.j} - n_{ij}$	$n_{..} - n_{i.} - n_{.j} + n_{ij}$	$n_{..} - n_{.j}$
Total	$n_{.j}$	$n_{..} - n_{.j}$	$n_{..}$

$$n_{i.} = \sum_j n_{ij} \qquad n_{.j} = \sum_i n_{ij} \qquad n_{..} = \sum_j \sum_i n_{ij}$$

-the random variable n_{ij} is an integer based, counting variable that records a count when an interaction between drug D_i and ADR A_j is observed.

Defining a Signal

-a signal, or a ‘suspiciously’ large interaction between drug and ADR, is identified when the observed interaction appears larger than would be expected due to chance alone.

Proportional Reporting Ratio (PRR) : A ratio comparing the conditional probability of observing ADR A_j given drug, D_i relative to the conditional probability of observing A_j with any other drug.

Reporting Odds Ratio (ROR) : A ratio comparing the odds (probability of success/probability of failure) of observing drug D_i and ADR A_j versus the odds of not observing the interaction.

Relative Reporting Ratio (RRR) : A ratio comparing the observed number of occurrences of drug D_i and ADR A_j to the expected number of occurrences of D_i and A_j under independence.

Sequential Probability Ratio Test (SPRT) : A comparison of the log likelihood ratio statistic to upper and lower boundaries which are functions of type 1 and type 2 error. The statistic is updated at regular time intervals until the point in which the statistic crosses a threshold.

	Estimator	Signal Detection Criteria
Proportional Reporting Ratio (PRR)	$PRR = \dfrac{\dfrac{n_{ij}}{n_{i.}}}{\dfrac{n_{.j} - n_{ij}}{n_{..} - n_{i.}}}$	$\exp\left[\ln(PRR) - 1.96\sqrt{\dfrac{1}{n_{ij}} - \dfrac{1}{n_{i.}} + \dfrac{1}{n_{.j} - n_{ij}} - \dfrac{1}{n_{..} - n_{i.}}}\right] > 1$ *OR* $PRR \geq 2,\ \chi^2_Y \geq 4,\ n_{ij} \geq 3$
Chi-Square with Yate's Correction	$\chi^2_Y = \sum_{i=1}^{2}\sum_{j=1}^{2} \dfrac{(\lvert n_{ij} - E_{ij}\rvert - 0.5)^2}{E_{ij}} \qquad E_{ij} = \dfrac{n_{i.} n_{.j}}{n_{..}}$	
Reporting Odds Ratio (ROR)	$ROR = \dfrac{n_{ij}(n_{..} - n_{.j} - n_{i.} + n_{ij})}{(n_{.j} - n_{ij})(n_{i.} - n_{ij})}$	$\exp\left[\ln(ROR) - 1.96\sqrt{\dfrac{1}{n_{ij}} + \dfrac{1}{n_{i.}} + \dfrac{1}{n_{.j} - n_{ij}} + \dfrac{1}{n_{..} - n_{i.}}}\right] > 1$
Yule's Q	$Q = \dfrac{ROR - 1}{ROR + 1}$	$\dfrac{ROR - 1}{ROR + 1} - 1.96\left(\dfrac{1}{2}(1 - Q^2)\right)\sqrt{\dfrac{1}{n_{ij}} + \dfrac{1}{n_{.j} - n_{ij}} + \dfrac{1}{n_{i.} - n_{ij}} + \dfrac{1}{n_{..} - n_{.j} - n_{i.} + n_{ij}}} > 0$

Bayesian Methods and the Relative Reporting Ratio (RRR)

The WHO method and the Information Component (IC)

$$P(A_j \mid D_i) = \frac{P(A_j, D_i)}{P(D_i)} = \frac{P(A_j, D_i)}{P(A_j)P(D_i)} P(A_j)$$

$$If \ \frac{P(A_j, D_i)}{P(A_j)P(D_i)} > 1 \quad \Rightarrow \quad P(A_j \mid D_i) > P(A_j)$$

$IC = \log_2 \frac{P(A,D)}{P(A)P(D)}$ which can be estimated by $IC_{ij} = \log_2 \frac{n_{ij} n_{..}}{n_{i.} n_{.j}}$

Signal detection criteria

If the 2.5-th percentile of the distribution of *IC* is above zero, a signal is concluded.

$$E(IC) - 1.96 * SD(IC) > 0 \qquad IC_{.025} > 0$$

Calculating the 2.5-th percentile of *IC*

The distribution of *IC* is difficult to determine analytically however, the distributions of the components of *IC* are easily attained given that the marginal counts are considered to be binomially distributed with parameters $n_{..}$ and the respective probabilities.

These probabilities are then assigned prior distributions as

$$P(A_j) \sim Beta(\alpha_1, \alpha_0) \qquad P(D_i) \sim Beta(\beta_1, \beta_0) \qquad P(A_j, D_i) \sim Beta(\delta_1, \delta_0)$$

Which results in the posteriors

$$P(A_j \mid n_{i.}, n_{..} - n_{i.}) \sim Beta(\alpha_1 + n_{i.}, \alpha_0 + n_{..} - n_{i.}) \qquad P(D_j \mid n_{.j}, n_{..} - n_{.j}) \sim Beta(\beta_1 + n_{.j}, \beta_0 + n_{..} - n_{.j})$$

$$P(A_j, D_i \mid n_{i.}, n_{..} - n_{i..}, n_{.j}, n_{..} - n_{.j}) \sim Beta(n_{ij} + \delta_1, n_{..} - n_{ij} + \delta_0)$$

Approximations for the expectation and variance were derived using the delta method and exact derivations can be computed using the moment generating function technique.

Approximate Expectation and Variance

$$E(IC_{ij}) \approx \log_2 \frac{\delta_1 + n_{ij.}}{\delta_1 + \delta_0 + n_{..}} - \log_2 \frac{\alpha_1 + n_{i..}}{\alpha_1 + \alpha_0 + n_{..}} - \log_2 \frac{\beta_1 + n_{.j}}{\beta_1 + \beta_0 + n_{..}}$$

$$V(IC_{ij}) \approx \frac{1}{\log_2^2}\left(\frac{\alpha_0 + N - n_{i.}}{(\alpha_1 + n_{i.})(\alpha_0 + \alpha_1 + n_{..} + 1)} + \frac{\beta_0 + n_{..} - n_{.j}}{(\beta_1 + n_{.j})(\beta_0 + \beta_1 + n_{..} + 1)} + \frac{\delta_0 + n_{..} - n_{ij}}{(\delta_1 + n_{ij})(\delta_0 + \delta_1 + n_{..} + 1)} \right)$$

Exact Expectation and Variance

$$E(IC_{ij}) = \frac{1}{\ln(2)}(\Psi(\delta_1 + n_{ij}) - \Psi(\delta_0 + \delta_1 + n_{..}) - [\Psi(\alpha_1 + n_{i.}) - \Psi(\alpha_0 + \alpha_1 + n_{..}) + \Psi(\beta_1 + n_{.j}) - \Psi(\beta_1 + \beta_0 + n_{..})]$$

$$V(IC_{ij}) = \frac{1}{\ln^2(2)}(\Psi'(\delta_1 + n_{ij}) - \Psi'(\delta_0 + \delta_1 + n_{..}) + [\Psi'(\alpha_1 + n_{i.}) - \Psi'(\alpha_0 + \alpha_1 + n_{..}) + \Psi'(\beta_1 + n_{.j}) - \Psi'(\beta_1 + \beta_0 + n_{..})]$$

$$\textit{where} \quad \Psi(x) = d\ln(\Gamma(x))/dx \quad \Psi'(x) = d\Psi(x)/dx$$

Multiple authors have proposed different starting values for the parameters of the posterior distributions

More recently a method for estimating the percentiles of the distribution of *IC* has been proposed using Monte Carlo simulation

The Gamma Poisson Shinker (GPS) and the Empirical Bayes Geometric Mean (EBGM)

Let $n_{ij} \sim Poisson\,(\mu_{ij} \equiv \lambda_{ij} E_{ij})$ where $E_{ij} = \dfrac{n_{i.} n_{.j}}{n_{..}}$ then if $\lambda_{ij} > 1 \Rightarrow \mu_{ij} > E_{ij}$ (expectation under independence)

To 'shrink' the estimates of λ_{ij} we must first consider its prior distribution

$$f(\lambda_{ij}) = P * \Gamma(\lambda_{ij}; \alpha_1, \beta_1) + (1-P) * \Gamma(\lambda_{ij}; \alpha_2, \beta_2)$$

where Γ represents the pdf of the gamma distribution

$$g(n_{ij}) = \int h(n_{ij}, \lambda_{ij}) d\lambda_{ij} = P * NB(n_{ij}; \alpha_1, \beta_1, E_{ij}) + (1-P) * NB(n_{ij}; \alpha_2, \beta_2, E_{ij})$$

where *NB* represents the pmf of the negative binomial distribution

and *P* is the prior mixing parameter

To determine the values of the prior parameters,

$$\max_{\theta} \prod_{i,j} g(n_{ij}) \text{ where } \theta = \{\alpha_1, \alpha_2, \beta_1, \beta_2, P\}$$

Posterior Mixing Parameter

$$Q_{n_{ij}} = \frac{P * NB(n_{ij}; \alpha_1, \beta_1, E)}{P * NB(n_{ij}; \alpha_1, \beta_1, E) + (1-P) * NB(n_{ij}; \alpha_2, \beta_2, E)}$$

The Posterior Distribution

$$p(\lambda_{ij} \mid n_{ij}) = Q_{n_{ij}} \Gamma(\lambda_{ij}; \alpha_1 + n_{ij}, \beta_1 + E_{ij}) + (1 - Q_{n_{ij}}) \Gamma(\lambda_{ij}; \alpha_2 + n_{ij}, \beta_2 + E_{ij})$$

The Posterior Expectation

$$E(\lambda_{ij} \mid n_{ij}) = Q_{n_{ij}} \frac{\alpha_1 + n_{ij}}{\beta_1 + E_{ij}} + (1 - Q_{n_{ij}}) \frac{\alpha_2 + n_{ij}}{\beta_1 + E_{ij}}$$

$$E(\log(\lambda_{ij}) \mid n_{ij}) = Q_{n_{ij}} [\Psi(\alpha_1 + n_{ij}) - \log(\beta_1 + E_{ij})] + (1 - Q_{n_{ij}}) [\Psi(\alpha_2 + n_{ij}) - \log(\beta_2 + E_{ij})]$$

The Emprical Bayes Geometric Mean (EBGM)

$$\Lambda_{ij} = \exp(E(\log(\lambda_{ij}) \mid n_{ij}))$$

Signal detection criteria

The 5-th percentile of the distribution of Λ, denoted as *EB05*, can be compared to 2

Shrinkage estimates

$\log(\frac{\alpha + n}{\beta + E}) \xrightarrow[n,E]{} \log(\frac{n}{E})$ however, for small values, $\Psi(x) \le \log(x)$ for $x \ge 0 \quad \Rightarrow \quad E(\log(\lambda) \mid n) \le E(\log(\lambda))$

Sequential Probability Ratio Test

For a general case $Let\ X = (x_1,...,x_n)'$

Subdivide sample space into three regions (R) such that:

$if\ \ x_i \in R_0 \Rightarrow H_0\ accepted$

$if\ \ x_i \in R_1 \Rightarrow H_1\ accepted$

$otherwise\ \ keep\ sampling$

Define:

p_{0m} & p_{1m} Are the probability distribution functions associated with the posterior probability that *Ho* or H_1 are true respectively

α is the probability of rejecting *Ho* given that it is true (type 1 error)

β is the probability of rejecting H_1 given that it is true (type 2 error)

General Decision Rule:

$$\frac{p_{1m}}{p_{0m}} \le \frac{\beta}{1-\alpha} \Rightarrow Accept\ H_0$$

$$\frac{p_{1m}}{p_{0m}} \ge \frac{1-\beta}{\alpha} \Rightarrow Accept\ H_1$$

Otherwise, continue to sample

In Pharmacovigilance for example,

$H_0 : \mu = \phi(RR)$ $H_1 : \mu = \phi(2RR)$

If $p_{im} \sim Poisson(\mu)\ \ i = 0,1$ then $\ln(2) \cdot n_{ij} - E_{ij} \ge \ln\left(\frac{1-\beta}{\alpha}\right)$

Discussion: Frequentist methods versus Bayesian methods

Frequentist

Poor choice of the complement set of drugs and ADRs will bias the results

These estimators heavily rely on the data and as a result the distributions of these estimators can be highly skewed

The calculation of the estimators at a specific cell is subject to the ‘complement’ set which may contain extremely large observations influencing their value

Small expected frequencies may tend to produce false positives (Type 1 error)

Bayesian

The distribution of the Relative Reporting Ratio will be pulled towards the chosen prior with a minimal data requirement

Over time, the updating of the parameters will cause the posterior to move towards the ‘true’ distribution

In other words, in the short term, the results of signal detection algorithms will be dependent on the appropriateness of the prior distribution

Bayesian methods lower impact of random fluctuations of the relative reporting ratio (shrinkage)

This may cause false negatives (Type 2 error)

Empirical Bayes is a computationally intensive algorithm

Further topics in Passive Pharmacovigilance Research

Higher Dimensional Associations

Multiple Drug/ADR associations

Confounding Factors

Covariate Stratification: Using pre-existing knowledge (eg. stratify on high risk demographic groups) to lower the effect of confounding variables

Multiple Logistic Regression: Modelling a predictor variable with respect to all possible covariates (ie. the set of all relevant drugs)

Unsupervised Pattern Recognition: Using data visualization techniques to identify how the data is clustered (organized)

References

P. Tubert et. al., “Power and weakness of spontaneous reporting: a probabilistic approach,” *Journal of Clinical Epidemiology,* vol. 45 no. 3 pp. 283-286, 1992.

D. Spiegelhalter et al. “Risk-adjusted sequential probability ratio tests: applications to bristol, shipman and adultcardiac surgery,” *Int. J. for Quality in Health care,* vol.15, no. 1, pp.7-13, 2003.

W. duMouchel, “Bayesian data mining in large frequency table, with an application to the FDA spontaneous reporting system,” *American Statistician*, col. 53, no. 3 pp.177-190, 1999.

E. Roux et al., “Evaluation of Statistical Association Measures for the Automatic Signal Generation in Pharmacovigilance,” *IEEE Transactions on Information Technology in Biomedicine,* vol. 9, no. 4, Dec. 2005.

A. Bate et al, “A Bayesian neural network method for adverse drug reaction signal generation,” Eur. J. Clin. Pharmacol, vol. 54, pp. 315-321. 1998.” *Pharmacoepidemiology and Drug Safety*, vol. 12, no.7, pp. 559-573, 2003.

S.J. Evans, P.C. Waller, and S. Davis, “Use of Proportional Reporting Ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports,” *Pharmacoepidemiology and Drug Safety*, vol. 10, no. 6 pp. 483-486, 2001.

K. J. Rothman, S. Lanes and S. T. Sacks, “The reporting odds ratio and its advantages over the Proportional Reporting Ratio,” *Pharmacoepidemiology and Drug Safety*, vol. 13, pp. 519-523, 2004.

Eugene P. van Puijenbroek, et.al, “A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions,” *Pharmacoepidemiology and Drug Safety,* vol. 11, pp. 3-10, 2002.

Gould, A.L. “Practical Pharmacovigilance analysis strategies,” *Pharmacoepidemiology and Drug Safety*, vol. 12, pp.559-574, 2003.

Orre, R. et al. “Bayesian neural networks with confidence estimations applied to data mining,” *Comput. Stat. Data Anal.* 34, 2000.

van Puijenbrook, E.P. et al. “Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. *Eur. J. Clin. Pharmacol.* 56, 2000.

Hocine, M.N, et al. “Sequential case series analysis for pharmacovigilance,” *J. R. Statist. Soc.* 172 part 1 pp. 213-236, 2009.

Ahmed, I et al. “Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting,” *Statist. Med.*, 28(13), 2009.

Orre, R, et al. “A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets,” *International Journal of Neural Systems*. Vol. 15 No. 3, 2005.

Norén, G. et al. “Case based imprecision estimates for Bayes classifiers with the Bayesian Bootstrap,” *Machine Learning*. 58, 2005.

Norén, G. et al. “Extending the methods used to Screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events,” *Statistics in Medicine*. vol. 25, 2006.

DuMouchel, W. and Pregibon D. “Empirical Bayes screening for multi-item associations,” *Proc. of the Seventh AACM SIGKDD International conference on Knowledge Discovery and Data Mining*, pp. 67-76, 2001.

Gould, A.L. “Accounting for multiplicity in the evaluation of ‘signals’ obtained by data mining from spontaneous report adverse event databases,” *Biometrical Journal*, vol. 49, pp. 151-165, 2007.