The Report Committee for Na Ye
Certifies that this is the approved version of the following report:

Vaccine-adverse event association analysis on the VAERS database

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Daniel A. Powers

Maytal Saar-Tsechansky

Vaccine-adverse event association analysis on the VAERS database

by

Na Ye, B.S.; M.S.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2011

Abstract

Vaccine-adverse event association analysis on the VAERS database

Na Ye, M.S. Stat

The University of Texas at Austin, 2011

Supervisor: Daniel A. Powers

The Vaccine Adverse Event Reporting System (VAERS) received thousands of reports of adverse events that occurred after vaccine administrations from the post-marketing vaccine safety surveillance. However, the causality between vaccines and reported adverse events cannot be taken for granted. In this report several data mining methods were applied to VAERS database that is coded in MedDRA terms to discover possible associations between vaccines and adverse events. Efforts were devoted to identify events that are reported more frequently after administering one vaccine than other vaccines using the following data mining techniques: relative ratio (*RR*), statistical significance (*LogP*), proportional reporting ratio (*PRR*), and screened *PRR* (*SPRR*).

The vaccine-event combinations that ranked top in each method varied substantially among the methods. *RR* and *PRR* gave excessive weight to small counts of vaccine-event pairs, but *SPRR* was able to correct this weakness. There are only 33

vaccine-event pairs that were shared among the top 1,000 ranked in each method. Evaluating the properties of these data mining methods and exploring other methods will help improve vaccine safety surveillance.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

It is accepted that the safety profile of therapeutic products (drugs, vaccines, medical devices) cannot be fully accessed before the products are approved for marketing due to the limitations of clinical trials. Therefore it is important for pharmaceutical companies and regulatory agencies to monitor and collect the side effects of products after the products are licensed for use in markets. A few spontaneous adverse event reporting systems have been initiated to collect post-marketing safety surveillance data in the United States.[1] The Vaccine Adverse Event Reporting System (VAERS) is one such system co-sponsored by the Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC) to monitor vaccine safety.[2]

VAERS receives around 30,000 reports annually from health care providers, manufacturers and vaccine recipients.[2] Each report describes one or more adverse events that occur after the administration of a certain vaccine, but VAERS does not determine its causality. Some of these events may occur just coincidentally following vaccination, while others may truly be caused by vaccination (e.g., injection site reaction). Signs and symptoms of reported adverse events on VAERS report used to be coded in the FDA's Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) from November 1990 until January 2007. Since January 2007 the VAERS coding system was converted to an international coding system Medical Dictionary for Regulatory Activities (MedDRA). There are more than 17,000 Preferred Term codes in the MedDRA system

that describe adverse events, while COSTART system is only comprised of 5,817 codes. Therefore the MedDRA system is not only standardized for international use, it is also able to code medical terms in a more exacting manner than the COSTART system design.

One interest in the analysis of VAERS data focuses on looking for patterns to detect if certain adverse events are plausibly linked to certain vaccines. However, spontaneous reporting systems such as VAERS do have some limitations including reporting bias (underreporting, differential reporting or stimulated reporting), inconsistent data quality, absence of an unvaccinated control group and inadequate denominator data (doses administrated or subjects vaccinated). The lack of control group and denominator data often limit the ability of traditional methods of signal detection and risk analysis, which require a controlled study or background rates.[3]

To address these limitations, various data mining techniques have been proposed to help detect potential safety concerns in the spontaneous reporting system. Data mining, defined as the process of looking for relationships and patterns from large datasets by combining methods from statistics and artificial intelligence with database management, is an increasingly important tool with great potential to extract previously unknown and potentially useful information, which then usually facilitates decision making[4].

In the situation of lack of a true unvaccinated control, data mining methods manage to use all other vaccines as a quasi-control group to analyze the vaccine of interest. Data mining cannot remove reporting bias, but can specify different proportions

of adverse events for each vaccine, where higher safety risks for certain vaccines can be detected but absolute ratios are not calculated.[3]

In 2005 Banks et al. compared several data mining methods on the VAERS database when the adverse events were still coded in COSTART terms. Because the VAERS database had converted its data to the more accurate and comprehensive MedDRA coding system in 2007, it is of interest to see if how these data mining methods would behave in the new system of MedDRA code. The objective here is not to judge the vaccine-adverse event causality, as it's more complicated to interpret than only using data mining and statistical methods. Rather, the goal is to determine whether these methods agree with each other, as well as their advantages and disadvantages.

METHODOLOGY

The information of each report to VAERS is divided to three pieces and stored separately: the patient's demographic information, the vaccine administration record and the adverse event coded in MedDRA. For the purpose of this analysis, three datasets were assembled; only reports with complete records of vaccine and MedDRA information were kept. Then the pre-processed database was transformed to a contingency table with vaccines in rows and the MedDRAs in columns. Each cell N_{ij} contains the number of the MedDRA of that column reported after the vaccine of the corresponding row.

Though occasionally one single report would report multiple MedDRAs, the effect of this case tends to be small, thus we assume that separate events are reported independently. In this report, a few main data mining techniques that have been explored within FDA will be applied to the newest VAERS data.

RELATIVE RISK (*RR*)

The method of relative risk in evaluating N_{ij} is comprised of two steps, the first step is to define a baseline or null hypothesis frequency E_{ij}, and the second step is then to compare the observed N_{ij} to the baseline frequency E_{ij}. The null hypothesis assumes that the vaccines and the MedDRAs are independent, though it's certainly accepted that they are associated in the system; the objective is not to test their independence, but to test if

there is any N_{ij} surprisingly larger than the baseline E_{ij} This association is of interest.[6] Assuming the MedDRAs are independent of the vaccines would give:

$$E_{ij} = \frac{N_{i.} N_{.j}}{N_{..}} \quad (1)$$

This simplest criterion is defined on the ratio

$$RR_{ij} = N_{ij}/E_{ij} \quad (2)$$

The relative risk evaluation is easy to interpret, if a pair of vaccines and MedDRA has been observed thousands of times more than expected, then we can conclude that there is plausibly an association between this MedDRA and this vaccine.

STATISTICAL SIGNIFICANCE (SS)

A statistical test criterion for testing the null hypothesis that $E[N_{ij}] = E_{ij}$ is defined as

$LogP_{ij} = -\log(\Pr[X \geq N_{ij}])$, where $X \sim \text{Poisson}(E_{ij})$.

The purpose is not to test the null hypothesis, but use the test statistics or degree of the significance to rank the vaccine-MedDRA association.

In the case of $N_{ij} \gg E_{ij}$, $LogP_{ij}$ could be too large for computation[5]. Thus, the following approximation will be used when $LogP_{ij}$ becomes large:

$$
\begin{aligned}
LogP_{ij} &= -\log(\Pr[X > N_{ij}]) \\
&= -\log_{10}\left(\sum_{n\geq N}\frac{e^{-E}E^{n}}{n!}\right) \approx -\log_{10}\left(\sum_{n=N}^{N+3}\frac{e^{-E}E^{n}}{n!}\right) \\
&= \frac{E}{\log(10)} - N\log(E) + \log_{10}(N!) - \\
&\qquad \log_{10}\left[1+\frac{E}{N+1}+\frac{E^{2}}{(N+1)(N+2)}+\frac{E^{3}}{(N+1)(N+2)(N+3)}\right]
\end{aligned}
$$

(3)

PROPORTIONAL RELATIVE RISK (*PRR*)

The proportional relative risk measures the proportion of MedDRA$_j$ reported after vaccine i is divided by the proportion of MedDRA$_j$ reported after all other vaccines.[3] A large *PRR* indicates a specific vaccine - MedDRA pair has been disproportionally reported after that vaccine, compared to other vaccines.

$$PRR_{ij} = \frac{N_{ij}/N_{i.}}{(N_{.j}-N_{ij})/(N_{..}-N_{i.})}$$

(4)

PRR is also an intuitive measure in the absence of exposure data, but it has a few shortcomings. *PRR* would be infinite if the MedDRA is only reported after one vaccine but not others; in this case $N_{.j} - N_{ij} = 0$. Also if N_{ij} is small, the generated signal can have large variance.

SCREENED PROPORTIONAL REPORTING RATIO (*SPRR*)

In the method of *PRR*, Evans et al. proposed to determine whether or not there is a signal, and its strength, is based on three pieces of information: the *PRR*, value of chi-

squared and the absolute number of reports. The chi-squared test is one degree of freedom of Yates correction to measure statistical association.[6] The Yates correction is a continuity adjustment to improve the accuracy of chi-squared approximation to the distribution of Pearson's test for independence in a contingency table.[3] They suggested defining a signal as $PRR \geq 2$, Yates-corrected chi-square ≥ 4, and $N_{ij} \geq 3$. The formula for the Yates-corrected chi-square is

$$\text{Yates-corrected } X^2 = \sum(|N_{ij} - E_{ij}| - 0.5)^2 / E_{ij} \quad (5)$$

E_{ij} has the same definition as defined in Eq. (1).

In the results that follow, all calculations were performed using a standard statistical software program (SAS, SAS Institute Inc, Cary, North Carolina).

RESULTS

Adverse events reported to VAERS system in the year of 2010 were analyzed in this report. There are 31,651 pieces of patients' demographic records, 49,736 pieces of vaccine administration records of these patients and 40,102 pieces of reported safety records in the database. Assembling these three datasets together generated 56,258 vaccine-MedDRA pairs that were reported in the system. This indicates that a single report may contain more than one vaccine and may report multiple adverse events. The data on 68 vaccines and 2,143 unique MedDRAs were reported, which can be considered to be a contingency table with 68 rows (vaccines) and 2,143 columns (MedDRAs).

A frequency plot as shown on Figure 1 indicates that 6,329 vaccine-MedDRA pairs each occurred only once in VAERS. The pairs that were reported most frequently (far right of graph) are injection site erythema that usually occurred after many vaccines. The most frequently reported vaccine-MedDRA pairs in VAERS of 2010 are listed in Table 1.

Table 1: The 10 most frequently reported vaccine-MedDRA pairs in VAERS.

MedDRA	Inj. site erythema	Inj. site erythema	Inj. site erythema	Inj. site erythema	Inj. site erythema
Vaccine	*VARCEL*	*MMR*	*Flu (10-11)*	*PPV*	*TDAP*
Count	908	563	530	496	463
MedDRA	Inj. site erythema	Erythema	Inj. site erythema	Inj. site erythema	Erythema
Vaccine	*DTAP*	*Flu (10-11)*	*HEPA*	*DTAP|PV*	*PPV*
Count	451	383	351	325	314

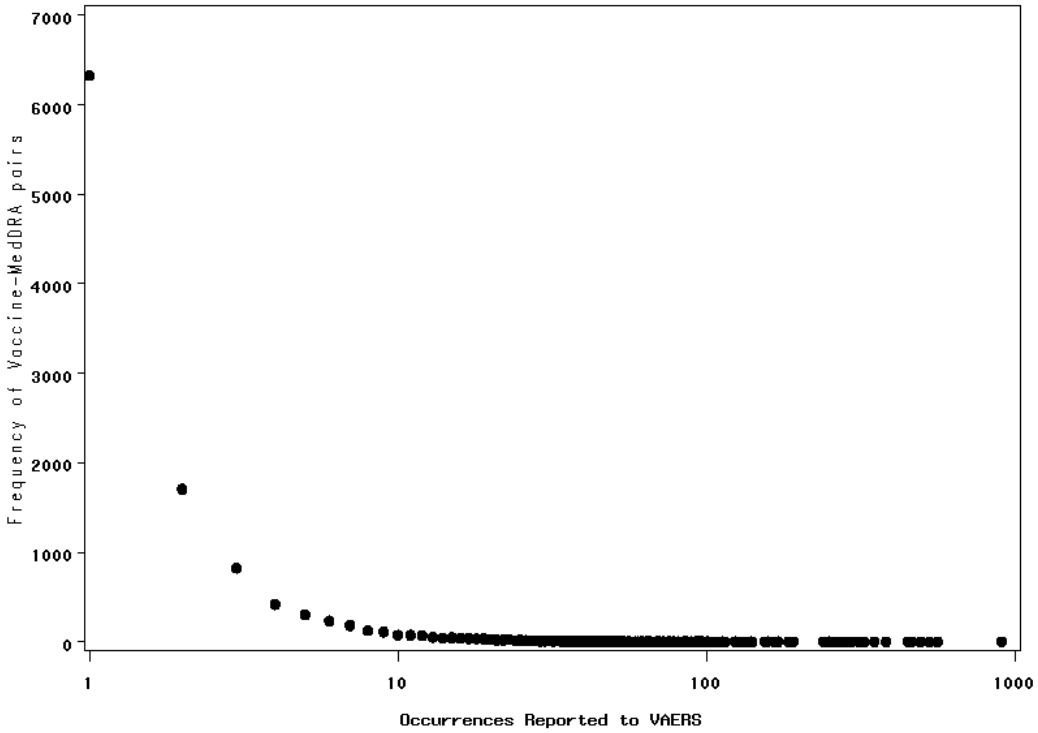

Figure 1: Frequency of occurrence of vaccine-MedDRA pairs for 56268 pairs that occurred at least once in the VAERS system. 6329 vaccine-MedDRA pairs were reported only once (far left of graph), the pairs that were reported more frequently (far right of graph) are injection site erythema that usually occurred after many vaccines.

RELATIVE RISK

The calculated relative risk ratios from formula (2) ranged from 0.039 to 7032.25. There are 10 vaccine-MedDRA pairs with a relative risk that exceeded 1000, as listed in Table 2. However, the actual occurrences N_{ij} that generate large RR scores of these 10 pairs are small. Most of them are $N_{ij} = 1$.

Table 2: Vaccine-MedDRA pairs appearing in the top 10 ranked by *RR* score

MedDRA	Pulmonary fibrosis	Drug effect decreased	Blood bicarbonate normal	Conjunctivitis infective	Heat therapy
Vaccine	*HBPV*	*RUB*	*DTPPHIB*	*DTAPH*	*DTPPHIB*
RR	7032.25	3125.44	2678.95	1654.65	1607.37
Count	1	1	1	1	1
MedDRA	Allergy to plants	Oesophagogastric fundoplasty	Thyroid disorder	Blood arsenic increased	Pulmonary fibrosis
Vaccine	*MEA*	*HBHEPB*	*HPV*	*DT*	*OPV*
RR	1406.45	1250.18	1125.16	1125.16	1081.88
Count	1	1	1	1	1

PROPORTIONAL REPORTING RATIO

According to the formula (4), a MedDRA that is reported only after one vaccine but not others would lead to an infinite *PRR* score, such pairs were omitted from the calculation of *PRR*. The rest of the pairs produced *PRR* scores ranging from 0.038 to 14,063.5. There are 13 vaccine-MedDRA pairs with a relative risk that exceed 1,000; the top 10 scores are listed in the Table 3.

Table 3: Vaccine-MedDRA pairs appearing in the top 10 ranked by *PRR* score

MedDRA	Pulmonary fibrosis	Blood bicarbonate normal	Conjunctivitis infective	Pulmonary fibrosis	Heat therapy
Vaccine	*HBPV*	*DTPPHIB*	*DTAPH*	*OPV*	*DTPPHIB*
PRR	14063.5	4017.93	3308.29	2162.77	2008.96
Count	1	1	1	1	1
MedDRA	Oesophagogastric fundoplasty	Change of bowel habbit	Conversion disorder	Raynaud's phenomenon	Blood lead increased
Vaccine	*HBHEPB*	*6VAX-F*	*HBHEPB*	*HPV*	*DT*
PRR	1874.77	1278.09	1249.84	1125.16	1081.88
Count	1	1	1	1	1

Comparing Table 3 to Table 2, most of the pairs that ranked in the top 10 by *RR* method also appear in the top 10 list of *PRR* method, and all of them have a count of 1.

Figure 2 displays the natural logarithm of the *RR* signal versus the natural logarithm of the *PRR* signal. The two evaluation methods are in good agreement for most of the vaccine-MedDRA pairs, except for a few pairs that the red arrow points to. These are the pairs that have large *PRR* scores for $N_{ij} = 1$, confirming that *PRR* gives excessive weight to singleton pairs, and thus it is highly subject to sampling variance[3].

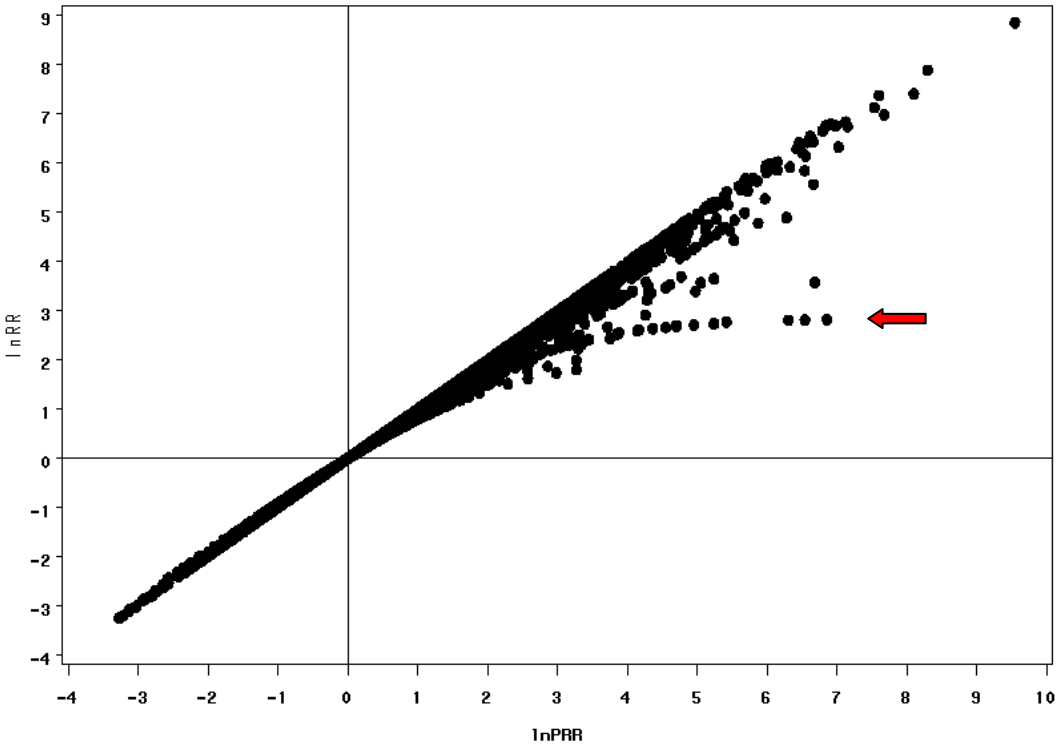

Figure 2: Scatterplot of log *RR* vs log *PRR*. The two methods are in good agreement except for a few points (arrow) that consist of vaccine-MedDRA pairs for which only one report was received. For these singleton cases, *PRR* gives excessive weight compared to *RR*.

STATISTICAL SIGNIFICANCE

LogP of 25 vaccine-MedDRA pairs could not be obtained in the computation of *LogP*, indicating that $N \gg E$ for these pairs. So the approximation method as described in formula (3) was applied when $N - E > 5$. Table 4 lists the top 10 vaccine-MedDRA pairs that were ranked by *LogP*.

Table 4: Vaccine-MedDRA pairs appearing in the top 10 ranked by *LogP* score

MedDRA	Herpes Zoster	Blood amylase increased	Crying	Syncope	Cellulitis
Vaccine	*VARZOS*	*MMR*	*DTAP|PVHIB*	*HPV4*	*PPV*
LogP	82.38	71.52	56.24	53.39	52.19
Count	94	86	95	124	158
MedDRA	Parotitis	Drug administration error	Expired drug administered	Chills	Varicella post vaccine
Vaccine	*MMR*	*FLUN(H1N1)*	*FLUN*	*FLUHD(10-11)*	*VARCEL*
LogP	49.56	46.03	44.15	39.67	37.40
Count	59	43	52	87	62

In contrast with the *RR* and *PRR* methods, the top pairs chosen by the *LogP* method were reported frequently in the system.

SCREENED PROPORTIONAL REPORTING RATIO

The screened version of *PRR* tends to repair the shortcoming of the *PRR* method mentioned above. The *SPRR* dropped the pairs with $N_{ij} < 3$, and also required the pairs to be statistically significant and have a *PRR* score ≥ 2. There are 808 vaccine-MedDRA pairs by *SPRR* definition; the top 10 scores are listed in Table 5.

Table 5:Vaccine-MedDRA pairs appearing in the top 10 ranked by *SPRR* score

MedDRA	Mumps	Uveitis	Blood amylase increased	Blood amylase	Myopericarditis
Vaccine	*MMR*	*HEP*	*MMR*	*MRR*	*SMALL*
PRR	953.57	800.92	694.98	549.52	534.50
Count	59	22	86	34	3
MedDRA	Myocarditis	Vaccine breakthrough infection	Echocardiogram abnormal	Alpha 1 foetoprotrein normal	Haemophilus test positive
Vaccine	*SMALL*	*MMR*	*SMALL*	*HPV4*	*HIBV*
PRR	237.55	226.27	195.98	189.58	189.49
Count	4	14	11	12	4

Figure 3 shows a plot of natural logarithm of the *RR* score against the natural logarithm of the screened *PRR* score. In comparison with Figure 2, all the points have a logarithm score ≥ 0.693 (ln2), as *SPRR* only keep pairs with *PRR* score ≥ 2. The points with extremely large *RR* scores and *PRR* scores disappear from the figure, as they correspond to cells with $N_{ij} = 1$ or where Yates-corrected X^2 are not statistically significant.

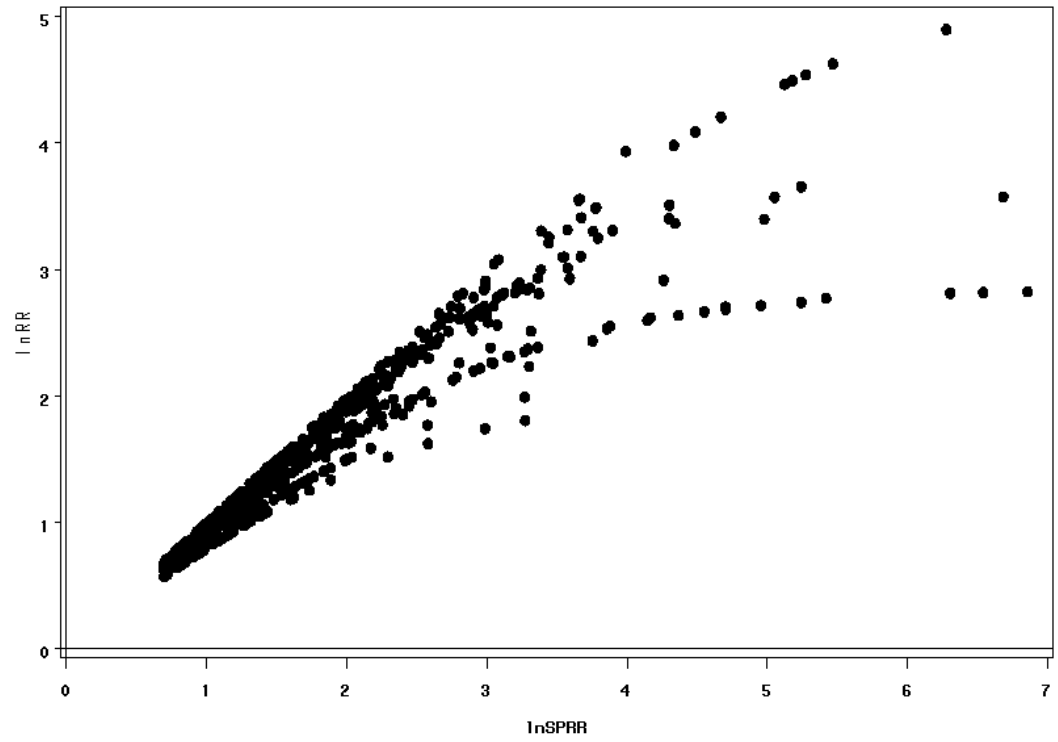

Figure 3: Scatterplot of ln *RR* vs ln *SPRR*. The points with extremely small scores and extremely large scores disappear from the figure in comparison with Figure 2, since *SPRR* only include cells for which $N_{ij} \geq 3$, $PRR \geq 2$ and Yates-corrected X^2 is statistically significant.

Table 6 shows some statistics for the 33 pairs that ranked in the top 1,000 pairs by three criteria: *SS*, *LogP* and *SPRR*. The fact that three sets of 1,000 pairs only have 33 in common indicates that there are real differences among those three selections. Table 4 shows the frequency of each pair and the rankings of this pair by these selection methods, the ranking by the *PRR* score are also listed for reference. It is interesting to see that these pairs also ranked among top 1,000 by the *PRR* method, and the rankings by the method of *RR* are closer to the method of *PRR*, while the rankings produced by *SPRR* are closer to

the ones of *LogP*. In addition, the intersections of pairs selected by these three criteria ranked in the top 100 by the *SPRR*, implying *SPRR* might be a more powerful selection method.

Table 6. Pairs appearing in the top 1,000 when ranked by all four methods

Vaccine	MedDRA	N	Rank by *RR*	Rank by *PRR*	Rank by *SPRR*	Rank by *LogP*
6VAX-F	Crying	**5**	995	824	87	153
ANTH	Echocardiogram abnormal	**5**	697	469	42	118
ANTH	Echocardiogram normal	**9**	664	436	39	65
DPP	Crying	**4**	887	818	84	222
DTAPHEPBIP	Abnormal faeces	**3**	769	594	51	300
DTAPIPVHIB	Culture urine	**3**	884	546	49	349
FLUHD(10-11	Blood creatinine increased	**3**	761	439	40	286
FLUN(10-11)	Epistaxis	**7**	807	595	52	117
FLUN(10-11)	Cheilitis	**4**	754	529	48	175
FLUN(H1N1)	Underdose	**14**	635	118	13	37
FLUN(H1N1)	Drug administration error	**43**	732	242	23	7
FLUN	Contraindication to vaccination	**15**	661	433	38	34
HEP	Drug ineffective	**4**	709	123	14	161
HEP	Uveitis	**22**	634	17	2	20
HIBV	Haemophilus test positive	**4**	556	87	10	139
PNC	Pneumococcal bacteraemia	**3**	708	246	26	263
PNC	Chronic obstructive pulmonary	**4**	882	470	43	217
ROT	Somnolence	**3**	421	361	31	174
ROTH1	Incorrect route of drug administration	**9**	760	614	53	76
ROTH1	Culture stool negative	**3**	643	474	45	235
ROTH1	Intussusception	**5**	639	471	44	106
ROTHB5	Regurgitation	**5**	657	245	25	112
ROTHB5	Infantile spitting up	**5**	759	385	32	128
ROTHB5	Rotavirus infection	**12**	770	432	37	46
ROTHB5	Culture stool positive	**7**	883	555	50	129
SMALL	Electrocardiogram abnormal	**5**	355	220	20	82
SMALL	Pericarditis	**3**	303	189	18	145
SMALL	Myopericarditis	**3**	141	33	5	105
SMALL	Myocarditis	**4**	216	67	6	86
SMALL	Acute myocardial infarction	**4**	230	91	11	89
SMALL	Blood creatine phosphokinase M	**3**	394	244	24	165
SMALL	Echocardiogram abnormal	**11**	226	84	8	31
SMALL	Echocardiogram normal	**17**	242	109	12	17

DISCUSSION

Significant effort has been devoted to improving the efficiency of adverse event reports using data mining methods in recent years. An Empirical Bayes geometric mean (*EBGM*) has been proposed by DuMouchel[5], and has been proposed to be a metric to measure the drug and adverse event association in the spontaneous reporting system. This method is currently adopted by FDA and has been incorporated in a software package called *MGPS* (Multi-Item Gamma Poisson Shrinker) in collaboration by DuMouchel, Lincoln Technologies and GlaxoSmithKline[1]. The *EBGM* method assumes that the counts N_{ij} in each cell are random variables from Poisson distribution with unknown means μ_{ij} where μ_{ij} are random variables from a common distribution. This common distribution is usually taken to be a mixture of two gamma distributions. The *EBGM* method has been shown to shrink the estimates of reporting ratio parameters in the Poisson distributions towards each other, thus reducing the effect of sampling variation in the data; this method also preserves the interpretability of the parameters and the estimates. But this approach is computationally intensive, thus due to the limited computational resources available in this report, this method was not studied.

Several data mining methods have been applied to the VAERS database when the system was coded in COSTART terms. In this report these data mining methods were applied to the new VAERS data where accurate adverse event descriptions in MedDRA terms were used. The qualitative features of the comparisons are as follows: Metrics based on the *RR* seem to be too favorable for cells with small counts. This may be improved by requiring a minimum value of *N* to be included into the analysis, such as $N \geq 5$, but the cutoff is arbitrary.[5] The raw *PRR* has the same problem with small counts, such that singleton cases tend to generate extremely large *PRR* scores. In

comparison with *PRR*, the *SPRR* has a significant improvement than PRR; moreover, previous studies has pointed out that *SPRR* is competitive with the EBGM method.[3] Determining which method is best will need additional analysis that focuses on true alarms and false alarms where known vaccine-event associations may help.

References

1. Fram DM, Almenoff JS, DuMouchel W. Empirical Bayesian Data Mining for Discovering Patterns in Post-Marketing Drug Safety. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining 2003.
2. http://vaers.hhs.gov.
3. Banks D, Woo EJ, Burwen DR, Perucci P, Braun MM, Ball R. Comparing data mining methods on the VAERS database. Pharmacoepidemiology Drug Safety 2005; **14**: 601-609.
4. Witten IH, Eibe F, Hall, MA. Data mining: practical machine learning tools and techniques. 2011. 3rd ed. Elsevier Inc.
5. DuMouchel W. Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System. The American Statistician 1999; **53**:177-190.
6. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology Drug Safety 2001; **10**: 483-486.