Simulating adverse event spontaneous reporting systems as preferential attachment networks

Application to the Vaccine Adverse Event Reporting System

J. Scott[1]; T. Botsis[1,2]; R. Ball[1]

[1]Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, U. S. Food and Drug Administration; [2]Department of Computer Science, University of Tromsø, Tromsø, Norway

Keywords

Simulation, data mining, vaccines, safety, network analysis

Summary

Background: Spontaneous Reporting Systems [SRS] are critical tools in the post-licensure evaluation of medical product safety. Regulatory authorities use a variety of data mining techniques to detect potential safety signals in SRS databases. Assessing the performance of such signal detection procedures requires simulated SRS databases, but simulation strategies proposed to date each have limitations.

Objective: We sought to develop a novel SRS simulation strategy based on plausible mechanisms for the growth of databases over time.

Methods: We developed a simulation strategy based on the network principle of preferential attachment. We demonstrated how this strategy can be used to create simulations based on specific databases of interest, and provided an example of using such simulations to compare signal detection thresholds for a popular data mining algorithm.

Results: The preferential attachment simulations were generally structurally similar to our targeted SRS database, although they had fewer nodes of very high degree. The approach was able to generate signal-free SRS simulations, as well as mimicking specific known true signals. Explorations of different reporting thresholds for the FDA Vaccine Adverse Event Reporting System suggested that using proportional reporting ratio [PRR] > 3.0 may yield better signal detection operating characteristics than the more commonly used PRR > 2.0 threshold.

Discussion: The network analytic approach to SRS simulation based on the principle of preferential attachment provides an attractive framework for exploring the performance of safety signal detection algorithms. This approach is potentially more principled and versatile than existing simulation approaches.

Conclusion: The utility of network-based SRS simulations needs to be further explored by evaluating other types of simulated signals with a broader range of data mining approaches, and comparing network-based simulations with other simulation strategies where applicable.

Correspondence to:

John Scott
Office of Biostatistics and Epidemiology, FDA/CBER
1401 Rockville Pike, HFM-215
Rockville, MD 20852
301–827–4608 (voice)
301–827–5218 (fax)
Email: john.scott@fda.hhs.gov

Appl Clin Inform 2014; 5: 206–218
DOI: 10.4338/ACI-2013-11-RA-0097
received: November 12, 2013
accepted: 13. January 13, 2014
published: March 5, 2014
Citation: Scott J, Botsis T, Ball R. Simulating adverse event spontaneous reporting systems as preferential attachment networks: Application to the Vaccine Adverse Event Reporting System. Appl Clin Inf 2014; 5: 206–218
http://dx.doi.org/10.4338/ACI-2013-11-RA-0097

1. Background

Spontaneous Reporting Systems (SRS) are used by health authorities around the world to monitor for adverse reactions to medical products. Patients, physicians, manufacturers and other parties report adverse event experiences that they believe to be associated with use of one or more products, and these reports are collected into standardized databases for analysis. In the United States, the Food and Drug Administration (FDA) uses the FDA Adverse Event Reporting System (FAERS, formerly AERS) and the Vaccine Adverse Event Reporting System (VAERS) to help assess the safety of drugs and vaccines, respectively, post-licensure. These databases suffer from substantial and well-known limitations, including certain systematic biases and the absence of "denominators" with which to estimate adverse event incidence and relative risks [1]. Nevertheless, FAERS and VAERS are seen as valuable tools in post-marketing surveillance, potentially capable of identifying safety signals earlier than other sources [2].

Researchers and public health officials have used a wide variety of data mining methods to identify potential safety signals against the very noisy background of SRS. Roux and colleagues compared 10 such methods in 2005; more have been proposed since then [3]. The application of such data mining approaches requires a decision rule: a threshold of identified signal strength above which a potential adverse reaction would be flagged for further investigation. It is impossible to determine the operating characteristics (e.g. false positive rate, false negative rate, etc.) of any given decision rule applied to a given data mining technique analytically, due to the inherent biases and limitations of SRS databases. Consequently, decision rules used in practice are chosen primarily by convention.

1.1 SRS Simulation Strategies

A few groups have attempted to more rigorously characterize the operating characteristics of various decision rules or have compared data mining algorithms by use of simulated SRS databases, either with or without known planted signals. To our knowledge, three distinct SRS simulation strategies have been proposed. Rolka and colleagues simulated VAERS by randomly permuting vaccine and adverse event (AE) term associations as found in the VAERS dataset [4]. They used these simulations to assess the specificity of various thresholds for signal scores of the multi-item gamma Poisson shrinkage (MGPS) estimator [5]. They also added safety signals to these background databases, by explicitly adding cases of either a specific vaccine-AE association or of a vaccine-syndrome (i.e. collection of AEs) association.

Ahmed and colleagues followed a similar strategy of basing simulated SRS databases on an actual dataset (in their case, a French pharmacovigilance system) [6]. In their simulations, cases of drug-AE association were drawn from a multinomial distribution whose parameters were determined by first assuming independence between drugs and AEs and then adding random departures from independence according to a logistic distribution. They used this simulation strategy to evaluate operating characteristics for a Bayesian decision framework [6] and to assess data mining approaches based on false discovery rates [7].

Tubert proposed simulating SRS data by assuming a Poisson distribution for the number of cases of each possible drug-AE association [8]. The expected number of cases of each association was governed by the assumed relative risk of the association, the number of patients exposed to the drug, the background incidence of the AE and the reporting rate for the specific drug-AE association. This simulation strategy was not originally developed to evaluate data mining methods but has since been used in that way [4, 7].

Each of these simulation approaches has strengths and all are reasonable efforts to generate artificial SRS databases for the purposes of data mining methods evaluation or comparison. We believe, however, that each also has significant limitations that may affect its ability to support accurate estimates of decision rule operating characteristics. The Ahmed and Tubert approaches each simulate fundamentally pairwise drug-AE associations. In practice, associations included in SRS databases are far more complicated, with signals generally consisting of a product (or an interaction of products) associated with a constellation of AEs, both due to syndromic relationships among the symptoms in question and also to coding conventions which may lead to the same symptom being reported under different standardized terms. This may lead to overstating the effectiveness of data

mining approaches that are tailored to identify exactly the pairwise drug-AE associations upon which the simulations are based. In addition, the Tubert approach has the drawback of requiring assumptions, difficult to justify in practice, about the reporting rate of each possible drug-AE association. The Rolka approach has the advantage of allowing higher-order relationships among products and AEs to be retained but, because it only shuffles existing data, it is not capable of properly representing noise due to variability in reporting rates, and disproportionality of individual drugs or AEs is retained. That is, the most common drugs or AEs in the database are exactly the most common drugs or AEs in the simulation.

1.2 Network Analysis Framework for SRS

Ball and Botsis recently proposed a novel framework for exploring VAERS using the tools of network analysis [9]. In this framework, each vaccine or AE is a node in the network, and nodes are connected by an edge if they appear together in at least one VAERS report. The network includes vaccine-AE, vaccine-vaccine, and AE-AE connections, to allow for representing multivariate interactions among products and adverse events [9]. AEs are coded in VAERS as Medical Dictionary for Regulatory Activities [MedDRA] preferred terms [PTs]. Edges can also be weighted by the number of reports in which pairs of nodes co-occur. Botsis and Ball argued that, by representing the complex interconnections between multiple vaccines and PTs, the network analysis framework provides useful insight into the structure of VAERS and can be used as both a visualization and data mining tool for signal detection and exploration [10].

2. Objectives

We sought to develop a novel SRS simulation strategy based on plausible mechanisms for the growth of databases over time. We used the network analytic framework described above as a conceptual basis for these simulations. Our belief was that by capturing more complex interrelationships between multiple products and AEs in our simulations, we would provide a more realistic background against which to evaluate data mining methods.

3. Methods

3.1 Preferential attachment simulations

Our SRS simulation approach is based on the preferential attachment model of Barabási and Albert [11]. In their model, as new nodes enter a network, the probability of forming connections with each existing node is proportional to the number of connections the existing node already has. In other words, more highly connected nodes are more likely to make connections with new nodes, hence "preferential attachment." Rather than simulating the evolution of a network as new nodes are added one at a time, our approach simulates the evolution of the SRS database as new spontaneous reports are added one at a time. Each report contains one or more product and AE nodes, potentially including both nodes that do and that do not currently exist in the SRS database. We therefore supplement the notion of preferential attachment of new nodes with a mechanism for new connections and reinforcement of connections between existing nodes.

The mathematical details of our simulation strategy are provided in the ▶Appendix. In general terms, starting from a network representation of a single report, additional reports are simulated one at a time and added to the network iteratively. The simulation of each individual report is governed by probability distributions which describe the number of products and AEs per report, and how likely each product or AE in a given report is to be new to the network. The nodes in each simulated report are hypothetical products and AEs.

3.2 Incorporation of Signals

The underlying simulation strategy is signal-free in the sense that there are no product-AE associations beyond those due to the preferential attachment mechanism and chance. These simulations can thus can be used to assess the false positive rate of a signal detection algorithm or as the basis for identifying deviations from an "expected" network under a null hypothesis of no true product-AE associations, for the purposes of signal detection. In order to incorporate safety signals for the purposes of assessing the sensitivity (true positive rate) of a data mining algorithm, we employ a strategy based on associations between a product and a syndrome of AEs. To have a simple gradient of signal strength for measuring sensitivity, we use a variant of the concept of node fitness [12]. Fitness is intended to represent the property that specific ("fit") nodes might attract future connections disproportionately from what would be expected under pure preferential attachment. That is, a fit node with a given number of connections is more likely to connect to new nodes in the network than would be a less fit node with the same number of connections.

To introduce simulated product-syndrome signals into our simulations, we choose simulated AEs for the syndrome based on strength rank in the network at the time of the product's first appearance. For example, to simulate a syndrome that consists of four AEs, two of which are relatively common and two relatively rare, we might define the syndrome to be AEs with strength rank at the 90th, 80th, 20th and 10th percentiles of all AEs at the time of the problem product's first appearance in a report. We also control the strength of association between each of the syndromic AEs and the problem product by specifying the probability of occurrence of each syndromic AE in reports that include the problem product. The time of the product's first appearance also affects the signal; due to preferential attachment, nodes that appear earlier in the network are more likely to attract connections. Therefore, a signal planted early in the growth of the simulation will tend to be easier to detect than one planted later.

3.3 Application to VAERS

As a test of this simulation approach, we attempted to create simulations based on all the reports to VAERS received in 1999. We chose 1999 because of the well-known safety signal related to intussusception following rotavirus vaccination (RV) [2, 13]. We simulated the database both with and without a planted signal based on the RV/intussusception signal. To create a simulated SRS analogous to the 1999 reports to VAERS, we derived parameters for the simulation from VAERS itself. The starting point for the network was based on the first report received in VAERS in 1999, which included two vaccines and three PTs. To simulate the number of vaccine and PT nodes for each subsequent report, we used multinomial distributions reflecting the distribution of number of nodes per report in VAERS in 1999. These distributions are shown in ▶Figure 1. The functions which govern the probability of adding novel nodes as a function of time are step functions over fixed intervals of 200 reports (approximately one week's worth of reports in VAERS in 1999). We calculated the parameters for these step functions from the VAERS database (▶Figure 2). We then let the network evolve for 12,000 reports, approximately the number of reports to VAERS in 1999.

To develop parameters for the simulated signal, one of the authors (R.B.) chose seven PTs in VAERS that were related to the rotavirus / intussusception signal: abdominal pain, dehydration, diarrhoea, hematochezia, intussusception, pyrexia and vomiting. We calculated the relative background frequency of each of these PTs in VAERS in 1999 prior to the introduction of the rotavirus vaccine, RotaShield (93rd, 88th, 95th, 16th, 40th, 99th and 97th percentile, respectively), and the probability of association of each with rotavirus vaccination (9%, 7.4%, 29.6%, 0.2%, 21.6%, 29.4% and 27.3%, respectively). We used these parameters along with a variable fitness parameter for the problem vaccine node to simulate intussusception-like safety signals of varying strengths.

3.4 Data mining example

We applied proportional reporting ratio (PRR)-based signal detection rules to a series of simulated instances of VAERS in 1999 to demonstrate the potential utility of our proposed simulation strategy [14]. We generated 1,000 simulations with no planted signal to estimate the false positive rates as-

sociated with PRR thresholds of 2.0 and 3.0 as the number of identified signals divided by the total number of vaccine-PT pairs in the simulation. We also generated simulations with planted RV/intussusception-like signals with no added fitness and with fitness parameters of 2, 3, 4, 5 and 10 (1,000 simulations each), and calculated the true positive rate for detecting any part of the RV/intussusception-like signal and for each component of the signal separately.

3.5 Software

All simulations and analyses were performed using custom code in version 2.15 of the R statistical computing environment [15]. We have also developed a Java application to create network simulations in a user-friendly environment as part of a broader Adverse Event Network Analysis application. This tool, which is available by request[1], allows users to apply various network analysis techniques to SRS data and also incorporates novel methodologies including the simulation approaches described in this paper.

4. Results

4.1 Comparison of simulated SRS networks and VAERS 1999

Although the simulation parameters were based on the ensemble of 1999 reports to VAERS, the simulations tended to have fewer PT nodes than VAERS itself. The mean (S.D.) number of vaccine nodes in our 7,000 simulations was 39.5 (6.0), and the mean (S.D.) number of PT nodes was 491.7 (22.9). These numbers were not sensitive to the presence or strength of a planted signal. In 1999, VAERS reports included 37 distinct vaccines and 781 PTs.

The *degree* of a node in a network is equal to the number of distinct nodes to which it is connected, and degree distributions are often used to summarize the basic structure of a network [16]. ▶Figure 3 shows a kernel density estimate for the distribution of ln(degree) for VAERS 1999 and for five randomly selected simulations. Degree is strongly bimodal in the VAERS database. The bimodality is visible in the simulations as well, although VAERS has relatively more nodes with ln(degree) of 2 – 4 than the simulations and relatively fewer nodes with ln(degree) of 4 – 6. The randomly chosen simulations in ▶Figure 3 are signal-free; however, there is no clear visual difference in degree distribution between signal-free and signal-containing simulations (not shown).

4.2 Evaluation of PRR signal detection with simulated SRS networks

▶Table 1 shows the false positive and true positive rates of PRR with signal detection thresholds of 2.0 and 3.0, applied to 1000 simulated SRS networks with intussusception-like signals of varying strengths. The false positive rate was around 9% with a threshold of PRR >2.0, and around 3% with a threshold of PRR >3.0. These false positive rates were not sensitive to the presence of a simulated signal. The true positive rates of correctly identifying at least one vaccine-PT pair from the planted syndromic signal were 45.2% for PRR >2.0 and 40.1% for PRR >3.0 with no added "fitness" (i.e. signal strength) to the signal vaccine. The true positive rates increased with increasing signal strength and were 97.7% and 95.8% for PRR >2.0 and PRR >3.0, respectively, at a signal strength of 5.

The true positive rates (at PRR >3.0) for each component of the signal syndrome are shown in ▶Table 2 for no added fitness and fitness parameters of 5 and 10. The probability of detection of each component is related to the background frequency of the PT and to the probability that the PT will occur on each report that contains the problem vaccine. For example, the signal component based on hematochezia is almost impossible to detect; hematochezia is very uncommon in the background of VAERS, but it also has a very weak association with RV vaccination. The component cor-

1 For access to this software for research use, please contact the FDA Technology Transfer Program at techtransfer@fda.hhs.gov. Access to this software for commercial use is available through the NIH Office of Technology Transfer at http://www.ott.nih.gov/licensing_royalties/licensing_overview.aspx

responding to pyrexia is also difficult to detect even though it is strongly associated with the problem vaccine; it is too common in the background of the simulation to be seen as disproportional in its occurrence with the vaccine. On the other hand, slightly less common PTs such as those corresponding to vomiting and diarrhoea are relatively easy to detect at comparable strengths of association with the problem vaccine. And the PT corresponding to intussusception itself is generally easiest to detect, since this is a rare event in the background and is strongly associated with the problem vaccine.

5. Discussion

We have introduced a novel approach for simulating signal-free SRS databases. Our approach uses an evolutionary algorithm with a parsimonious set of assumptions. The primary assumption is that the growth of the database proceeds according to the principle of preferential attachment; that is, that more common and more highly connected products and AEs are more likely to appear in future reports. We believe that this is an intuitively plausible model for how an SRS might evolve if there were no true underlying product-AE associations and, as such, this simulation strategy provides a useful environment for assessing and comparing the false positive rate of data mining algorithms. One side effect of preferential attachment is that early "spurious" reports may disproportionately be reinforced as the network grows. This may be a realistic representation of certain features of real-world SRS databases, such as the tendency for publicized associations to lead to increased report rates of the same association.

We can derive growth parameters for these simulations from an SRS database of interest. From our comparisons between the VAERS database in 1999 and our simulations based on that database, however, it is clear that the simulated SRS has important differences from the real data. This is perhaps not surprising given that the simulation is signal-free, whereas any real SRS is likely to include a myriad of true signals of greater and lesser importance.

We also described how syndrome-like signals can be introduced into our simulated SRS databases. These signals can be tailored to provide realistic representations of specific safety events in the real world, which are rarely as clean as a single drug disproportionately associated with a single dictionary term. We believe that simulating a signal as a syndrome allows for a more realistic assessment of the sensitivity of data mining algorithms, most of which are designed to identify binary product-AE relationships. Signals of this kind could also be used to explore the operating characteristics of multidimensional data mining algorithms such as the MGPS in a nuanced fashion, although such an evaluation is beyond the scope of this article.

We presented an example in which we applied PRR-based signal detection rules to simulated VAERS data with and without a planted RV/intussusception-like signal. We showed that using a PRR >3.0 threshold decreased the false positive rate from 9% to 3% relative to a PRR >2.0 threshold, with a relatively less dramatic impact on sensitivity (e.g. a reduction from 97.7% sensitivity to 95.8% at a signal strength of 5). It should be noted, though, that even a relatively modest loss of sensitivity may be unacceptable for many adverse event signal detection applications. We were also able to investigate which components of the intussusception-like signal were easiest and hardest to detect. PTs corresponding to very common adverse events such as vomiting and diarrhoea were readily detectable when the association with a product was strong. However, even very strong relationships with the most common PT (corresponding to pyrexia in the 1999 VAERS database) were difficult to detect due to the high background rate of the event.

These results help to demonstrate the potential utility of our SRS simulation approach for evaluating existing data mining methods. We are also intrigued, however, by the possibility of using the simulations themselves as the basis for data mining applications. We believe that a signal-free simulation of an SRS of interest can be used as an "expected" SRS against which an "observed" (i.e. empirical) SRS can be compared. Discrepancies between the observed and simulated SRS could be indications of the presence of a safety signal.

6. Conclusion

Our network-based SRS simulations provide a novel and intuitive platform for evaluating the performance of adverse event signal detection methods. The utility of these network-based SRS simulations needs to be further explored by evaluating additional types of simulated signals with a broader range of data mining approaches, and comparing network-based simulations with other simulation strategies where applicable.

Clinical relevance statement

The interpretation of safety signal detection results from spontaneous reporting systems such as VAERS and FAERS is hampered by incomplete understanding of the operating characteristics of data mining algorithms. To adequately characterize these operating characteristics, simulated databases that capture the underlying structure of the spontaneous reporting system, with and without true known signals, are required. Our network analytic simulation strategy provides a principled means of evaluating data mining algorithms, potentially improving our understanding of medical product safety.

Conflict of Interest

The authors have no conflicts of interest to disclose.

Human Subjects Protections

This work did not involve human or animal research subjects.

Acknowledgements

The authors would like to thank four anonymous reviewers, as well as Andrea Sutherland, Ravi Goud, Pamela Toman, Vahan Grigoryan, Marek Cyran and Estelle Russek-Cohen for their insightful comments.

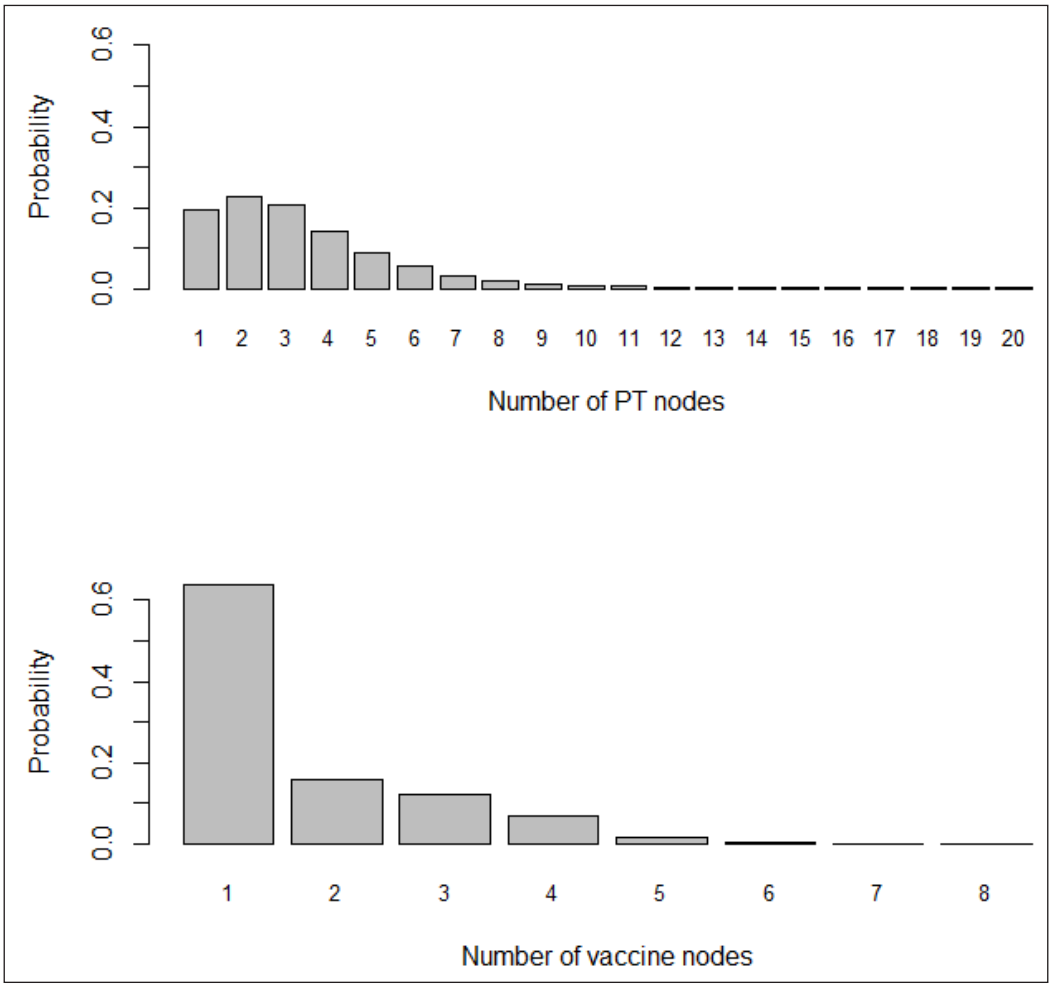

Fig. 1 Distribution of number of PTs (top panel) and vaccines (bottom panel) per report in VAERS in 1999. These distributions were used for the network simulations described in the article.

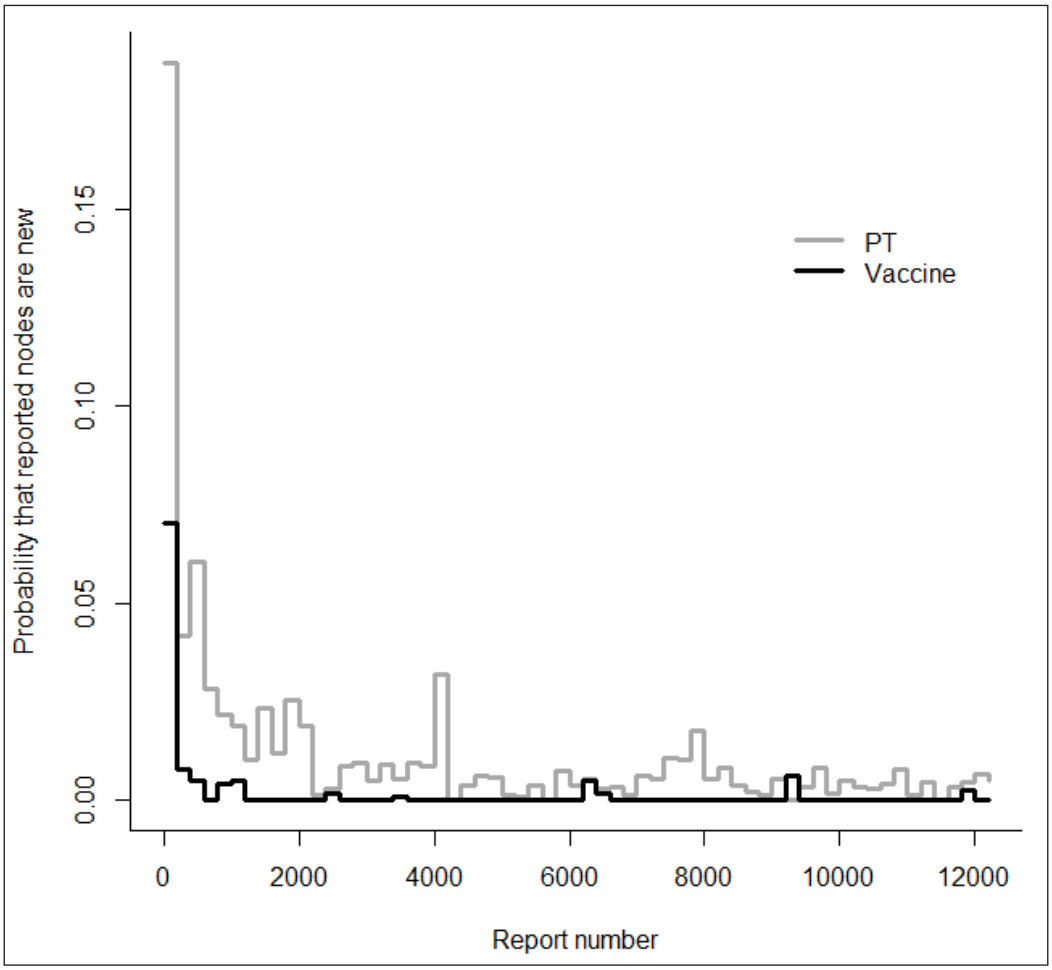

Fig. 2 Probability that each PT (gray line) and vaccine (black line) in a report is new to VAERS in 1999 as a function of report number. These probabilities were used for the network simulations described in the article.

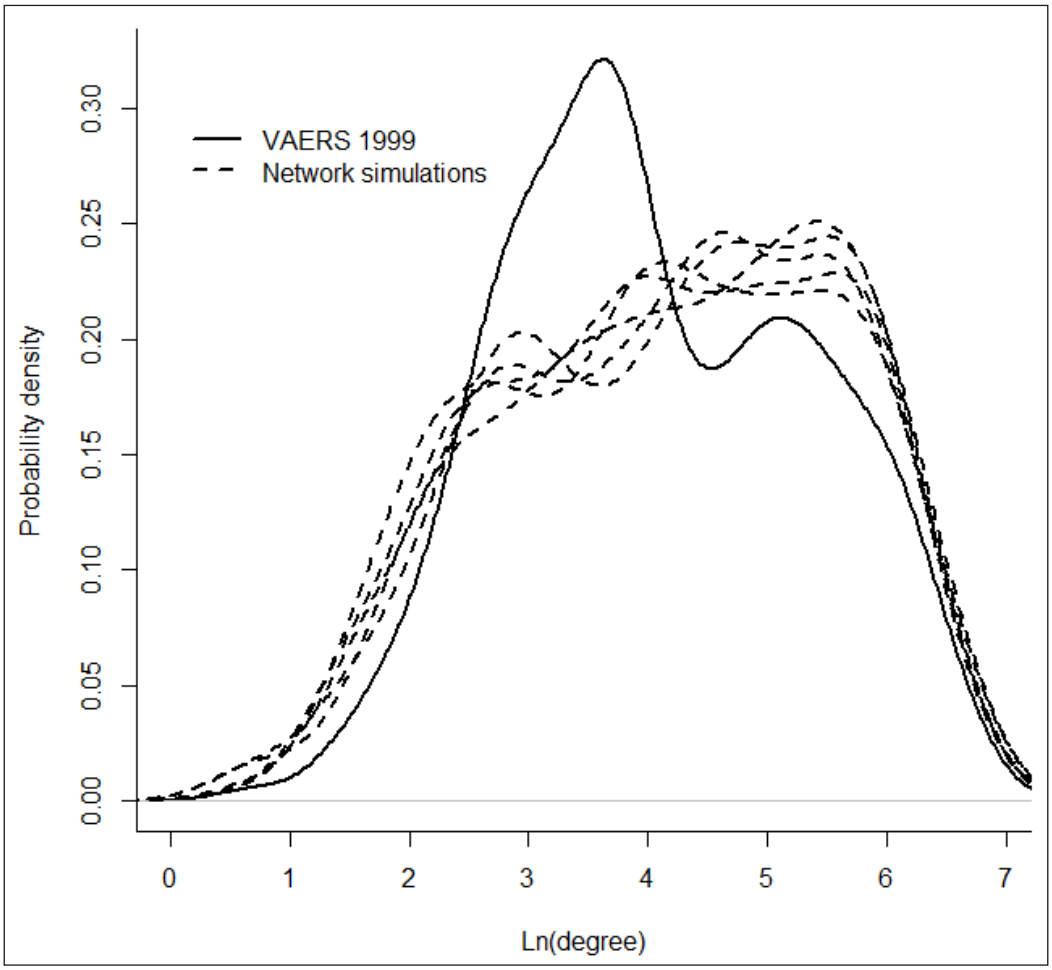

Fig. 3 Gaussian kernel density estimates of degree distribution (natural log scale) for VAERS in 1999 (solid line) and for five randomly chosen network simulations (dashed lines).

Signal strength	PRR >2.0		PRR >3.0	
	FP rate	TP rate	FP rate	TP rate
No signal	9.2%	NA	3.1%	NA
0	9.2%	45.2%	3.1%	40.1%
2	9.2%	75.9%	3.1%	71.1%
3	9.2%	87.7%	3.1%	79.4%
4	9.2%	95.6%	3.1%	92.9%
5	9.2%	97.7%	3.1%	95.8%
10	9.2%	100.0%	3.1%	99.8%

Table 1 PRR false positive and true positive signal detection rates: False positive (FP) and true positive (TP) signal detection rates of PRR applied to simulated VAERS networks with intussusception-like signals of varying strength. TP rate reflects proportion of simulations in which at least one component of the syndrome was detected. Each row represents 1,000 simulations

Table 2 Detection rates for each component of a simulated intussusception-like signal: True positive rates are based on PRR >3.0 in 1,000 simulations with signal strengths of 0, 5 and 10; signal strength of 0 corresponds to no added fitness.

PT[2]	Background frequency (%ile)	Probability of co-occurrence	True positive rate		
			0[1]	5[1]	10[1]
Abdominal pain	93	0.090	4.7%	24.6%	23.7%
Dehydration	88	0.074	2.7%	25.8%	42.7%
Diarrhoea	95	0.296	22.1%	69.0%	79.5%
Hematochezia	16	0.002	0%	0%	0.1%
Intussusception	40	0.216	10.8%	72.6%	96.8%
Pyrexia	99	0.294	11.7%	11.1%	6.7%
Vomiting	97	0.273	14.5%	33.0%	31.8%

[1] True positive rates are based on PRR >3.0 in 1,000 simulations with signal strengths of 0, 5 and 10; signal strength of 0 corresponds to no added fitness.
[2] PTs are simulated to be analogous to syndromic AEs from VAERS 1999 database in terms of background frequency and probability of co-occurrence with signal vaccine.

References

1. Varricchio F, et al. Understanding vaccine safety information from the Vaccine Adverse Event Reporting System. Pediatr Infect Dis J 2004; 23(4): 287–294.
2. Niu MT, Erwin DE, Braun MM. Data mining in the US Vaccine Adverse Event Reporting System (VAERS): early detection of intussusception and other events after rotavirus vaccination. Vaccine 2001; 19(32): 4627-4634.
3. Roux E, et al. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. IEEE Trans Inf Technol Biomed 2005; 9(4): 518-527.
4. Rolka H, et al. Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. Stat Med 2005; 24(4): 551-562.
5. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In: Proceedings of the KDD, ACM, New York, 2001; 67–76.
6. Ahmed I, et al. Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. Stat Med 2009; 28(13): 1774-1792.
7. Ahmed I, et al. Pharmacovigilance data mining with methods based on false discovery rates: a comparative simulation study. Clin Pharmacol Ther 2010 ;88(4): 492–498.
8. Tubert P, et al. Power and weakness of spontaneous reporting: a probabilistic approach. J Clin Epidemiol 1992; 45(3): 283–286.
9. Ball R, Botsis T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? Clin Pharmacol Ther 2011; 90(2): 271–278.
10. Botsis T, Ball R. Network analysis of possible anaphylaxis cases reported to the US vaccine adverse event reporting system after H1N1 influenza vaccine. Stud Health Technol Inform 2011; 169: 564–568.
11. Barabási A-L, Albert R. Emergence of scaling in random networks. Science 1999; 286(5439): 509-512.
12. Bianconi G, Barabási A-L. Bose-Einstein condensation in complex networks. Phys Rev Lett 2001; 86(24): 5632-5635.
13. Haber P, et al. An analysis of rotavirus vaccine reports to the vaccine adverse event reporting system: more than intussusception alone? Pediatrics 2004; 113(4): e353-e359.
14. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001; 10(6): 483-486.
15. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2012).
16. Jeong H, et al. The large-scale organization of metabolic networks. Nature 2000; 407(6804): 651–654.

Appendix

Formal description of simulation algorithm

Suppose that a simulated SRS database based on t reports contains k nodes, m_1, …, m_k, where each node corresponds to either a product or an AE. The SRS can be represented as a k^*k strength matrix, $\mathbf{M}_t$, whose i,j^{th} entry, n_{ij}, is the number of reports in the database which contain both node m_i and node m_j, $1 \le i, j \le k$. Nodes are not considered to be connected to themselves (i.e. no loops), so the diagonal elements, n_{ii}, are all 0. The total strength of node m_i is then defined as $\sum_j n_{ij}$

Next, simulated report $t + 1$ is added to the network. Each simulated report consists of l_0 product nodes and l_1 AE nodes. These two numbers are drawn from probability distributions with density functions $f_0(x)$ and $f_1(x)$, respectively. The probability that each product and AE is new to the network is given by $p_0(t+1)$ and $p_1(t+1)$, respectively. Note that these latter probabilities are functions of report number, reflecting the fact that, as a database evolves, the proportion of reported products and AEs that are novel to the SRS decreases rapidly. The functions $f_0(x)$, $f_1(x)$, $p_0(t+1)$ and $p_1(t+1)$ can be derived from the particular database that will be simulated; we describe an example of this in Section 3.3. Assuming $0 \le l_0^{new} \le l_0$ and $0 \le l_1^{new} \le l_1$ new product and AE nodes are included in the report, the remaining nodes in the report are drawn from existing nodes m_i with probability proportional to the current strength of m_i, $\frac{\sum_j n_{ij}}{\sum_{i,j} n_{ij}}$.

The strength matrix is then updated to $\mathbf{M}_{t+1}$ to reflect the connections among the nodes, new and existing, in report $t + 1$.

When adding a signal to the simulated database, to provide a gradient of signal strength against which to evaluate data mining algorithms, we give each problem product node, m_i, a fitness score, φ_i. The probability of drawing m_i in future reports is then proportional to its strength plus fitness, rather than strength alone. That is, for each selection of a product for a simulated report, the probability that the problem product, m_i, will be selected is $\frac{\phi_i + \sum_j n_{ij}}{\phi_i + \sum_{i,j} n_{ij}}$ rather than $\frac{\sum_j n_{ij}}{\sum_{i,j} n_{ij}}$.

Note that this differs from the multiplicative definition of node fitness introduced by Bianconi [12].