Disproportionality Methods for Pharmacovigilance in Longitudinal Observational Databases

Ivan Zorych[1,2], David Madigan[1,2], Patrick Ryan[1,3,4], Andrew Bate[5,6,7]
1 Observational Medical Outcomes Partnership
2 Columbia University
3 Johnson & Johnson Pharmaceutical Research and Development
4 University of North Carolina at Chapel Hill
[5] Pfizer Inc, USA
[6] Brunel University, UK
7 NYU, New York, USA

1. Introduction

Increasing scientific, regulatory and public scrutiny focuses on the obligation of the medical community, pharmaceutical industry and health authorities to ensure that marketed drugs have acceptable benefit-risk profiles. This is an intricate and ongoing process that begins with careful pre-approval studies but continues after regulatory market authorization when the drug is in widespread clinical use. In the latter environment, surveillance schemes based on spontaneous reporting system (SRS) databases represent a cornerstone for the early detection of drug hazards that are novel by virtue of their clinical nature, severity and/or frequency. SRS databases collect voluntary reports of adverse events made directly to the regulator or to the manufacturer of the product. Such spontaneous report databases present a number of well-documented limitations such as under-reporting, over-reporting, and duplicate reporting. Furthermore, SRS databases fail to provide a denominator – the number of individuals that are actually consuming a drug – and generally have limited temporal information with regard to duration of exposure and the time order of exposure and condition (Hauben et al., 2005).

Despite the limitations inherent in SRS-based pharmacovigilance, analytic methods for spontaneous report databases have attracted considerable attention in the last decade, and several different methods have become well established, both in commercial software products and in the medical literature. A number of methods have been applied to the analysis of spontaneous reports (e.g., Praus et al., 1993, vanPuijenbroek et al., 2000 Orre et al., 2005, and Mammadov et al., 2007) including some recent work focused on Bayesian shrinkage regression (Caster et al., 2010) but all of the more widely used methods compute measures of "disproportionality" for specific drug-condition pairs (Bate and Evans, 2009). That is, the methods quantify the extent to which a given condition is "disproportionally" reported with a given drug.

Newer data sources have emerged that overcome some of the SRS limitations but present methodological and logistical challenges of their own. Longitudinal observational databases (LODs) provide time-stamped patient-level medical information. Typical examples include medical claims databases and electronic health record systems. The scale of some of these databases presents interesting computational challenges – the larger claims databases contain upwards of 50 million lives with more than 1 billion of clinical observations (prescription dispensings, diagnoses, procedures, laboratory tests) over up to 10 years per life. A nascent

literature on signal detection in LODs exists. Several papers have looked at vaccine safety in this context – see, for example, Lieu et al. (2007), McClure et al. (2008), and Walker (2010). Papers focusing on drug safety include Curtis et al. (2008), Jin et al. (2008), Kulldorff et al. (2008), Li (2009), Norén et al. (2008), and Schneeweiss et al. (2009).

In this paper we explore the application of disproportionality methods to the LOD context. The motivation for our work derives from several quarters. First, disproportionality methods have become familiar to drug safety scientists and, insofar as such methods can be applied to LODs, the learning curve should be modest. Second, many of the standard disproportionality methods have the potential to scale well to very large databases. Third, the simplicity of many of the disproportionality methods leads to transparent outputs. Finally, issues related to multiplicity and shrinkage estimation for disproportionality have attracted considerable attention and mature solutions are now available.

2. Prior work

There has been some previous work on the implementation of measures of disproportionality in the context of LODs. Jin et al. (2008) looked at the incidence of an adverse event in a fixed 6 month hazard period subsequent to the prescription of a given drug in linked pharmaceutical, hospital and medical service data from Australia. Norén et al. (2008) have chosen to adapt the Information Component (IC) measure of disproportionality to account for the occurrence, or not, of disease prior to drug prescription as done in a self- controlled case series analysis. Norén et al. (2008) refer to this measure as an IC delta and also present a lower 95% confidence limit of the IC delta. In subsequent work (Norén et al., 2010) they showed results from for a specific drug-wide screen that an IC calculated from data solely after drug exposure only highlighted quantitatively one known reaction in a top 10 listing, as opposed to the seven that the IC delta achieved. All results from this research group were based on the analysis of the UK IMS Disease Analyser data set.

Curtis et al. (2008) proposed how one of the measures of disproportionality, the Multi-item Gamma Poisson Shrinker (MGPS) might be used to screen the Medicare claims database. They only looked at a single data set and one specific established drug-AE combination that they showed they were able to highlight this established issue.

Hocine et al. (2009) combine the self-controlled case series approach with a sequential probability ratio test (SPRT) to allow for prospective repeated analysis of evolving collections of electronic patient records. As for the SPRT the authors focus on searching for predefined types of pattern rather than truly open-ended hypothesis generation.

A related approach to measures of disproportionality, is the use of a SPRT to screen healthcare data sets. Davis et al. (2005) retrospectively examined 5 years of data from 4 HMO networks to compare the rates of a small number of selected examples including intussusception after rotavirus vaccination and showed that these new vaccine effects could be highlighted early using the SPRT on this data set. The SPRT approach has subsequently been used to look at drugs and vaccines throughout the HMO Research Network (Brown, et al., 2007, Lieu, et al., 2007, and Brown et al., 2009).

3. Disproportionality Methods and Spontaneous Reports

Disproportionality analysis methods for drug safety surveillance represent the primary class of analytic methods for analyzing data from SRSs. SRSs receive reports that comprise of one or more drugs, one or more adverse events (AEs), and possibly some basic demographic information (in addition to narrative and text data). These reports are compiled into a computerized database, which can be used to standardize the identification of the co-occurrence of drugs and adverse events within each report. Table 1 below shows a conceptual representation of a typical SRS entry.

Table 1: A conceptual representation of a typical entry in an SRS database

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

Disproportionality analysis methods include the multi-item gamma-Poisson shrinker, MGPS, (DuMouchel, 1999, DuMouchel and Pregibon, 2001, Fram et al., 2003), proportional reporting ratios, PRR, (Evans et al., 2001), reporting odds ratios, ROR, (Rothman et al., 2004), and Bayesian confidence propagation neural network, BCPNN, (Bate et al., 1998, Norén et al., 2006). The methods search SRS databases for "interesting" associations and focus on low-dimensional projections of the data, specifically 2-dimensional contingency tables. Table 2 shows a typical table.

Table 2: A fictitious 2-dimensional projection of an SRS database

	AE j = Yes	*AE j = No*	*Total*
Drug *i* = Yes	w_{00}=20	w_{01}=100	120
Drug *i* = No	w_{10}=100	w_{11}=980	1080
Total	*120*	*1080*	1200

The basic task of a disproportionality method then is to rank order the tables in order of "interestingness." Different disproportionality methods focus on different statistical measures of association as their measure of "interestingness". MGPS focuses on the "reporting ratio" (RR). The observed RR for the drug *i* – adverse event *j* combination (RR_{ij}) is the number of occurrences of the combination (20 in the example above) divided by the expected number of occurrences. MGPS computes the expected value under a model of independence. Specifically, in the example above, overall, AE *j* occurs in 10% of the reports (120/1200). Thus, if drug *i* and adverse event *j* are statistically independent, 10% of the reports containing drug *i* should

include AE *j*, that is 12 reports in this case. Thus the observed RR for this example is 20/12 or 1 2/3; this combination occurred about 67% more often than expected.

Natural (though not necessarily unbiased) estimates of various probabilities emerge from tables like Table 2. For example, one might estimate the conditional probability of AE *j* given drug *i* by $w_{00}/ w_{00}+ w_{01}$ (i.e. 20/120 in the example above). That is, the observed fraction of drug *i* reports that listed AE *j*. Table 3 below lists the formulae for the various measures of association in common use, along with their probabilistic interpretation. Here "$\neg drug$" for example denotes the reports that did not list the target drug. PRR is the "Proportional Reporting Ratio", ROR is the "Reporting Odds Ratio," and IC is the "Information Component" used by BCPNN (Bate et al., 1998, Norén et al., 2006).

Table 3: Common measures of association for 2 X 2 tables in SRS analyses

Measure of Association	*Formula*	*Probabilistic Interpretation*
RR (Reporting Ratio)	$\frac{w_{00} \times (w_{00} + w_{01} + w_{10} + w_{11})}{(w_{00} + w_{10}) \times (w_{00} + w_{01})}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae)}$
PRR (Proportional Reporting Ratio)	$\frac{w_{00} / (w_{00} + w_{01})}{w_{10} / (w_{10} + w_{11})}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae \mid \neg drug)}$
ROR (Reporting Odds Ratio)	$\frac{w_{00} / w_{10}}{w_{01} / w_{11}}$	$\frac{\Pr(ae \mid drug) / \Pr(\neg ae \mid drug)}{\Pr(ae \mid \neg drug) / \Pr(\neg ae \mid drug)}$
IC (Information Component)	$\log_2 \frac{w_{00} \times (w_{00} + w_{01} + w_{10} + w_{11})}{(w_{00} + w_{10}) \times (w_{00} + w_{01})}$	$\log_2 \frac{\Pr(ae \mid drug)}{\Pr(ae)}$

All four of these measures make sense – in each case, a particular drug that is more likely to cause a particular AE than some other drug will typically receive a higher score. Similarly, if an AE and a drug are stochastically independent, all measures will return a null value. However, all four are subject to sampling variability, i.e. a different set of AE reports from the same "population" will not give exactly the same value of the measure of association. This may be particularly the case with large sparse databases. Due to the Law of Large Numbers, this statistical variability diminishes as the sample size increases. In the SRS context, however, the count in the "w_{00}" cell is often small, leading to substantial variability (and hence uncertainty about the true value of the measure of association) despite the often large numbers of reports overall.

PRR and ROR do not address the variability issue whereas MGPS and BCPNN adopt a Bayesian approach to address the issue. MGPS places a prior distribution on RRs that encapsulates a prior belief that most RRs are close to the average value of all RR's (i.e., close to 1) whereas the BCPNN assumes a prior distribution centered around an RR of 1, based on empirical testing. Only in the face of substantial evidence from the data does BCPNN or MGPS return an RR estimate that is substantially larger than one. Thus, for example, an RR of 1,000 that derives from an observed count of w_{00}=1 might result in a MGPS RR estimate (Empirical Bayesian Geometric Mean or EBGM) of 1.5 (i.e. the crude RR is shrunk towards a value of 1) whereas an RR of 1,000 that derives from an observed count of w_{00}=100 might result in a EBGM RR estimate of close to 1,000. For the specific Bayesian setup that MGPS uses, observed counts in excess of 10 result in RR estimates that typically receive essentially no shrinkage, although in practice larger differentials have been observed depending on the thresholds used (Hauben and Zhou, 2003, Hauben et al., 2004, Hauben and Zhou, 2004). Similar properties have been observed with the BCPNN (van Puijenbroek, et. al., 2002, Bate and Evans 2009.) We note that small cell counts may be less frequent in LODs than SRS.

The EBGM and IC scores are means of the posterior distribution of the true RR. Other summaries are possible. For example, DuMouchel mentions EB05 (DuMouchel, 1999). This is the 5th percentile of the posterior distribution – meaning that there is a 95% probability that the "true" RR exceeds the EB05. Since EB05 is always smaller than EBGM this, in a sense, adds extra shrinkage and represents a more conservative choice than EBGM.

We note some analysts use the standard chi-square statistic for 2X2 tables, especially in combination with the PRR score. We include a "signed" chi square statistic in our analyses below.

4. Applying Disproportionality Analysis to Longitudinal Data

In the context of spontaneous report systems, some authors use the term "signal of disproportionate reporting" (SDR) when discussing associations highlighted by disproportionality methods (Hauben et al., 2005, Hauben and Reich, 2005). Hauben and Reich introduced the term to distinguish metric scores in SRS data from signals of suspected causality that have undergone clinical review. As in reality, most SDRs that emerge from spontaneous report databases represent noise because the reports are associated with treatment indications (i.e., confounding by indication), co-prescribing patterns, co-morbid illnesses, protopathic bias, channeling bias, or other reporting artifacts, or, the reported adverse events are already labeled or are medically trivial. In this sense, SDRs represent *generated* hypotheses about potential drug safety issues that warrant further investigation. Furthermore, spontaneous report databases present a number of well-documented limitations such as under-reporting, over-reporting, and duplicate-reporting, they fail to provide a denominator – how many individuals are actually consuming drug, and generally have limited temporal information with regard to duration of exposure and the time order of exposure and condition (Hauben et al., 2005). The richer context of longitudinal data (such as claims databases or electronic health records) affords the possibility of more refined analysis to address some of these artifacts. Nonetheless, given the wide acceptance of disproportionality

methods in pharmacovigilance, application of these approaches to longitudinal data may prove useful.

The key step in the application of disproportionality methods to any data is the mapping of the data into drug-condition two-by-two tables. With longitudinal data many choices present themselves. In this paper we consider three particular approaches, "distinct patients," "SRS," and "modified SRS."

In what follows we will illustrate the approaches using the example of Figure 1. Figure 1 shows three patients. Patient 1 consumed drug A during two separate drug eras, or spans of persistent exposure to a particular medical product. The patient experienced condition X three times during these eras, twice during the first era and once during the second. Patient 2 also had three drug eras but with three separate drugs, A, B, and C. Finally Patient 3 had two overlapping drug eras, one with drug B and one with drug C. The patient experienced condition O while taking both B and C, and conditions O and X after the drug eras.

Note we treat conditions as if they occur at distinct moments in time. In fact the data may contain condition "eras", or episode of care for a particular disease, and what we are utilizing is the timestamp of the beginning of the era. Drug eras, on the other hand, play an important role in our approach. A drug era represents a continuous period of drug usage, which can be augmented during analysis with an additional period post-exposure to capture outcomes that may still be drug-related. We refer to the optional post-exposure period as a surveillance window.

We now consider three different approaches to constructing the two-by-two table for drug A and condition X. Appendix B provide a formal mathematical description.

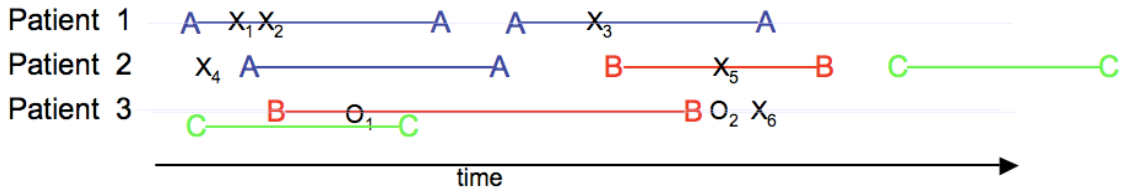

Figure 1. A longitudinal dataset with three patients, three distinct drugs (A, B, and C) and two distinct conditions (X and O).

4.1 Distinct Patients

In the "distinct patients" approach to table construction, $w_{00}+w_{01}+w_{10}+w_{11}$ (denoted w_{++}) equals the total number of patients in the database. w_{00} is the number of patients that had a drug A era and experienced condition X during a drug A era. w_{01} is the number of patients that had a drug A era and did not experience condition X during a drug A era. w_{10} is the number of patients that did not have a drug A era but experienced condition X. w_{11} is the number of patients that did not have a drug A era and never experienced condition X. Thus, for the example of Figure 1, $w_{00}=1$ (patient 1), $w_{01}=1$ (patient 2), $w_{10}=1$ (patient 3), and $w_{11}=0$. Note that $w_{++}=3$.

4.2 SRS

The second approach attempts to mimic what SRS reports the longitudinal data would generate, in that the observed co-occurrence of a drug and outcome is treated as if it were reported as a spontaneous case. w_{00} is the number of distinct X conditions the occur during drug A eras. w_{01} is the number of distinct non-X conditions that occur during drug A eras. w_{10} is the number of distinct X conditions that occur during non-A drug eras. w_{11} is the number of distinct non-X conditions that occur during non-A drug eras. Thus, for the example of Figure 1, w_{00}=3 (A+X_1, A+X_2, A+X_3), w_{01}=0, w_{10}=1 (B+X_5), and w_{11}=2 (B+O_1,C+O_1).

4.3 Modified-SRS

A third approach augments the SRS-like reports with additional denominator-based information about exposures without outcomes and outcomes that occurred without prior exposure. This approach attempts to patch obvious weaknesses of the SRS approach by taking advantage of the other information available in the LODs. Specifically, whereas the SRS systems do not contain the total number of drug exposures, only those exposures that were co-reported with an adverse event, LODs offer the potential to measure the number of event-free exposures. Similarly, the background rate of events is not well captured in SRS, since only drug-related events are recorded, whereas LOD can provide information about conditions that occur independently from drug exposures. The modified SRS approach counts "non-event" drug eras and "non-drug" conditions that can be combined with all drug-related events prior to calculating the disproportionality measures. In this approach, as with SRS, w_{00} is the number of distinct X conditions that occur during drug A eras. w_{01} however, is the number of distinct non-X conditions that occur during drug A eras plus the number of A eras in which no events occur. w_{10} is the number of distinct X conditions that occur outside drug A eras. w_{11} is the number of distinct non-X conditions that occur during non-A drug eras plus the number of non-A drug eras with no conditions plus the number of non-X conditions with no drug era. Thus, for the example of Figure 1, w_{00}=3 (A+X_1, A+X_2, A+X_3), w_{01}=1 (patient 2's A era), w_{10}=3 (X_4, B+X_5, X_6), and w_{11}=4 (patient 2's C era, B+O_1,C+O_1, O_2).

Our application of disproportionality methods to longitudinal data also makes a distinction between incident and prevalent conditions. The incident case only considers the first occurrence of each event, whereas the prevalent case (considered in the above example) considers all occurrences. Thus, for the example above, the incident analysis would proceed as above but only consider the first event of each type. Figure 2 illustrates the modified dataset used in an incident analysis. Note that our use of the term "incident" does not necessarily coincide with standard use of this term in epidemiological practice. In particular, we do not require an event-free "clean" period prior to first condition occurrence.

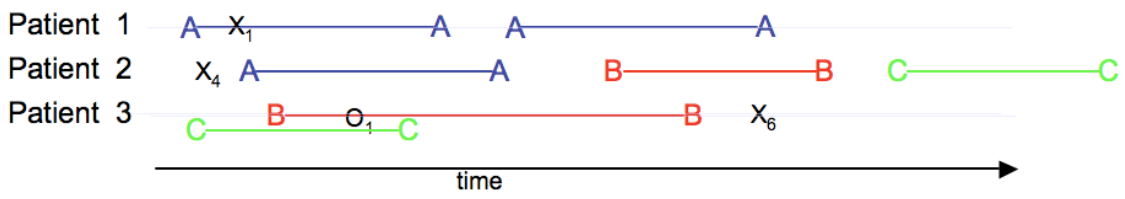

Figure 2. A longitudinal dataset with three patients, three distinct drugs (A, B, and C) and two distinct conditions (X and O). Incident conditions only.

Three mapping approaches, distinct patients, SRS, and modified SRS, together with two condition types, prevalent and incident, produce six mapping scenarios: distinct patients - prevalent, distinct patients - incident, SRS - prevalent, SRS - incident, modified SRS -prevalent, and modified SRS - incident.

In this study we use 9 disproportionality measures: PRR, PRR05 (left bound of the 90% confidence interval for PRR), ROR, ROR05 (left bound of the 90% confidence interval for ROR), IC, IC05 (lower credibility limit for 90% credibility interval for IC), EBGM, EB05, and signed Chi-square. See Appendix A for additional details.

5. Data and Evaluation Procedure

5.1 Real Data

In the numerical experiments reported below, we use de-identified Thomson Reuters MarketScan Lab database (MSLR). MSLR contains 1.5 million persons representing a largely privately-insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results. We also successfully executed all mentioned above LOD mapping approaches and calculated disproportionality metrics on several other de-identified Thomson Reuters databases: Medicaid Multi-State database (MDCD), Medicare Supplemental and Coordination of Benefits database (MDCR), and Commercial Claims and Encounters database (CCAE). These three databases contain medical records of 5 million, 11 million, and 59 million persons respectively.

5.2 Simulated Data

For performance evaluation purposes, we used a simulated database (OSIM). OSIM is modeled after real observational databases. It contains information on 10 million simulated persons, 5000 simulated drugs, and 4000 simulated conditions. Database date range is 10 years. Because of simulated nature of the data, the list of true drug-condition associations is available, enabling quantitative comparison of the different measures of disproportionality.

5.3 Mean Average Precision

To compare the performance of different disproportionality methods, we use "mean average precision" (MAP), a scoring method widely used in text retrieval. Higher values of a MAP score indicate better performance. Let $y_{dc} = 1$ if the dth drug causes the cth condition and 0 otherwise, d=1,…,D, c=1…,C. Let $M = \sum_{d,c} y_{dc}$ denote the number of causal combinations and $N = D \times C$ the total number of combinations. Note it is generally the case that $M << N$. Let z_{dc} denote the score (EBGM, IC, PRR, etc.) for the dth drug and the cth condition. For a given set of predicted values $\vec{z} = (z_{11}, \cdots, z_{dc})$, we define "precision-at-K" denoted $P^{(K)}(\vec{z})$ as the fraction of causal combinations amongst the K largest predicted values in $\vec{z}$ (Madigan et al., 2006). Specifically, let $z_{(1)} > \cdots > z_{(N)}$ denote the ordered values of $\vec{z}$. Then:

$$P^{(K)}(\vec{z}) = \frac{1}{K}\sum_{i=1}^{K} y_{(i)},$$

where $y_{(i)}$ is the causal status of the combination corresponding to $z_{(i)}$. The mean average precision or MAP score, S, is then:

$$S = \frac{1}{M}\sum_{K:y_{(K)}=1} P^{(K)}(\vec{z}).$$

If there are ties in $\vec{z}$, the results are ordered such that negative drug-condition combinations, i.e. combinations such that $y_{dc} = 0$, go first. The precision of tied positive combinations, i.e. combinations such that $y_{dc} = 1$, is calculated as if they were all encountered at the same time, i.e. if $z_{(K)} = z_{(K+1)}$, then $P^{(K)}(\vec{z}) = P^{(K+1)}(\vec{z})$.

Table 4 contains hypothetical drug-condition pairs and their scores. Suppose that there are 3 drugs, D1, D2, and D3, and 3 conditions, C1, C2, and C3. But only 5 pairs, D1-C1, D1-C2, D2-C1, D2-C2 and D3-tC3 are causal combinations and the remaining pairs are not. Table 5 illustrates the scoring process.

Table 4. Sample data: 3 drugs, 3 conditions; y_i =1 if drug-condition combination is positive, y_i =0 otherwise.

Drug	**Condition**	**Score (z)**	**Truth (y)**
D1	C1	5	1
	C2	0	1
	C3	9	0
D2	C1	8	1
	C2	5	1
	C3	0	0
D3	C1	0	0
	C2	0	0
	C3	5	1

Table 5. Illustration of the MAP methodology.

		Sorted Values		
Drug	**Condition**	z	y	$P^{(K)}$
D1	C3	9	1	1/1=1
D2	C1	8	1	2/2=1
D3	C3	5	0	
D1	C1	5	1	3/4=0.75
D2	C2	5	1	4/5=0.8
D2	C3	0	0	
D3	C1	0	0	
D3	C2	0	0	
D1	C1	0	1	5/9=0.55
MAP = (1 + 1 + 0.75 +0.8 +0.55) / 5 = 0.82				

In some cases, not every possible drug-condition combination may be assigned a score. Pairs without a valid score are treated as if they were given the lowest possible score value. In table 5 these pairs would be placed after all scores that were actually observed.

6. Results

6.1 Application to simulated data

Disproportionality analysis methods were computationally efficient, executing against the simulated dataset with 10 million persons for 4000 drugs and 5000 conditions and 4.9 x 10^8 drug-condition co-occurrences. All analyses were able to complete in less than 24 hours, running in a single 64-bit computer with 2.67 GHz I-7 CPU and 12 GB of RAM.

Figure 3 shows scatterplots of the various disproportionality measures (on the logarithmic scale) for a random sample of 1000 drug-condition pairs for the SRS-prevalent scenario.
PRR and ROR scores are very similar, as it is almost always the case with the LOD data that w_{00} is much smaller than w_{01}, and the same is true for w_{01} and w_{11}. Due to the same rationale, PRR05 and ROR05 scores are extremely close as well. IC and EBGM, both being shrinkage estimators, show quite similar behavior, as do IC05 and EB05. Scatterplots of PRR versus EBGM (PRR vs. IC, PRR vs. IC05, ROR vs. EBGM, PRR vs. EB05, etc.) shows greater variability because PRR gives similar to EBGM results when cell counts, w_{00}, are large but PRR is less stable for small counts (Madigan, 1999). Signed Chi-square scores are qualitatively different from other measures of disproportionality. We observed similar relationships in the other five analysis scenarios: distinct patients-incident, distinct patients – prevalent, SRS-incident, modified SRS-prevalent, and modified SRS-incident.

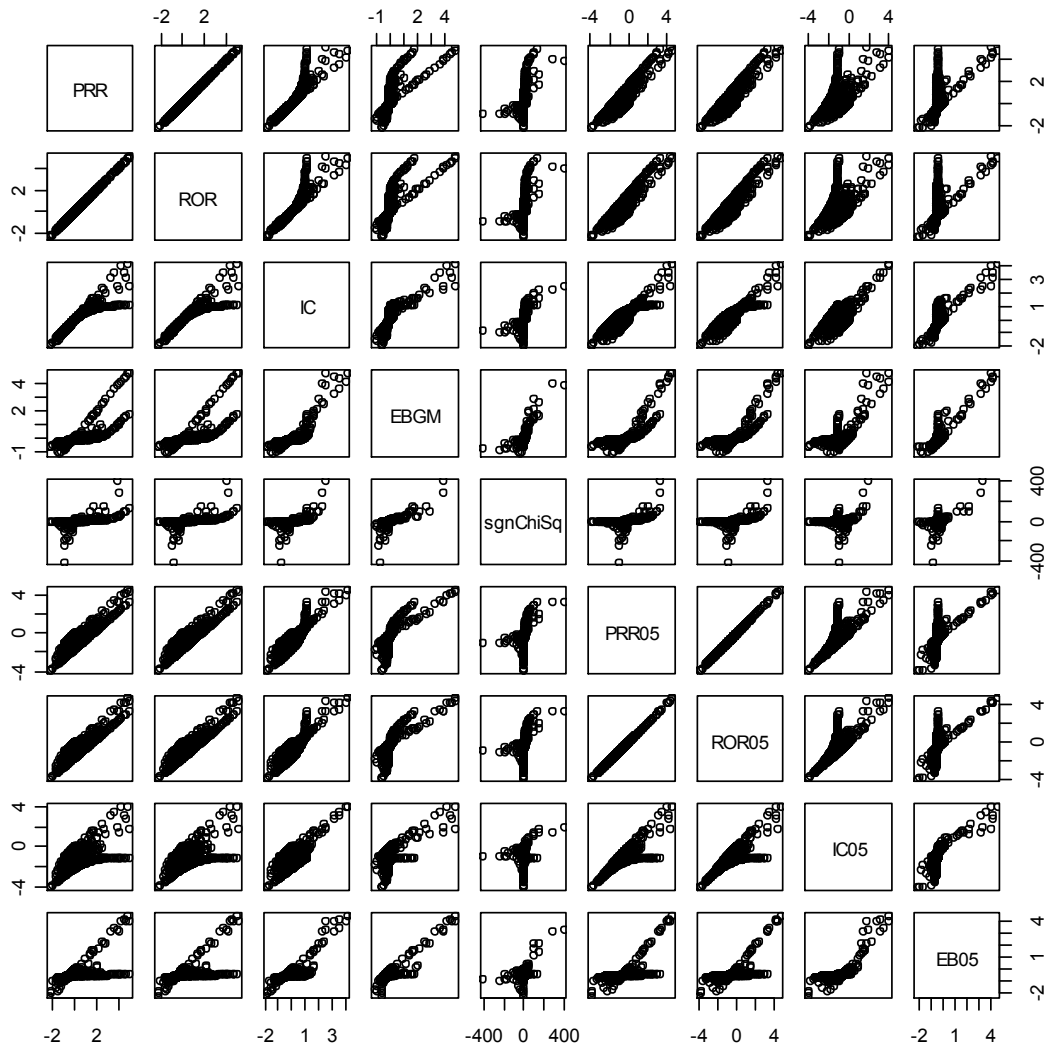

Figure 3. Prevalent events -SRS (simulated data). Scatterplots (on the logarithmic scale) of nine measures of disproportionality for the SRS – prevalent scenario on the simulated data.

Figure 4 illustrates the EBGM scores across the six different table construction scenarios. It is interesting to note that for the same counting approach, distinct patients, SRS, or modified SRS, scores are often similar across different event types, prevalent or incident. Distinct patient counting approach shows similar scores for both prevalent and incident event types. SRS counting approach shows the same. It is also worth noting that SRS -prevalent scores are similar to both SRS - incident and modified SRS - incident scores.

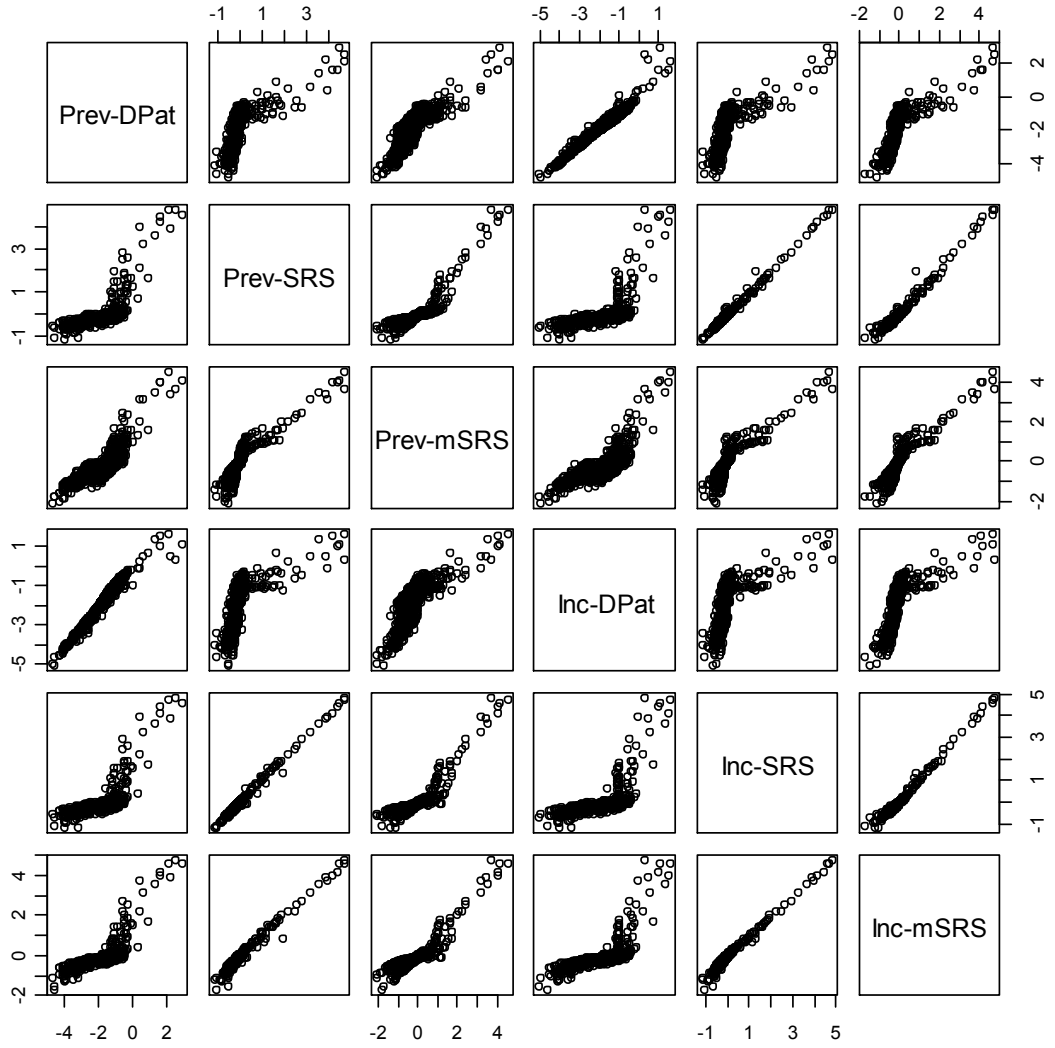

Figure 4. EBGM scores across six scenarios (simulated data). Scatterplots of EBGM scores across six scenarios: Distinct patients - prevalent, SRS - prevalent, modified SRS - prevalent, distinct patient - incident, SRS - incident, modified SRS - incident.

6.2 Performance on simulated data

Figure 5 presents MAP scores for all six approaches to table construction (distinct patients, SRS, and modified SRS for both prevalent events and incident events) and nine different scoring approaches (PRR, ROR, IC, EBGM, signed Chi-square, PRR05, ROR05, IC05, and EB05).

Figure 5 shows that

- Both the SRS and modified SRS counting approaches provide similar performance for both incident and prevalent cases;
- the distinct patients approach to table construction provides inferior performance to the other two counting approaches, SRS and modified SRS;

- the Bayesian approaches to scoring, IC and EBGM, provide the highest levels of performance, with signed Chi-square, IC05 and EB05 not far behind.
- For all three counting approaches, distinct patients, SRS, and modified SRS, ROR and PRR provide the lowest MAP scores.

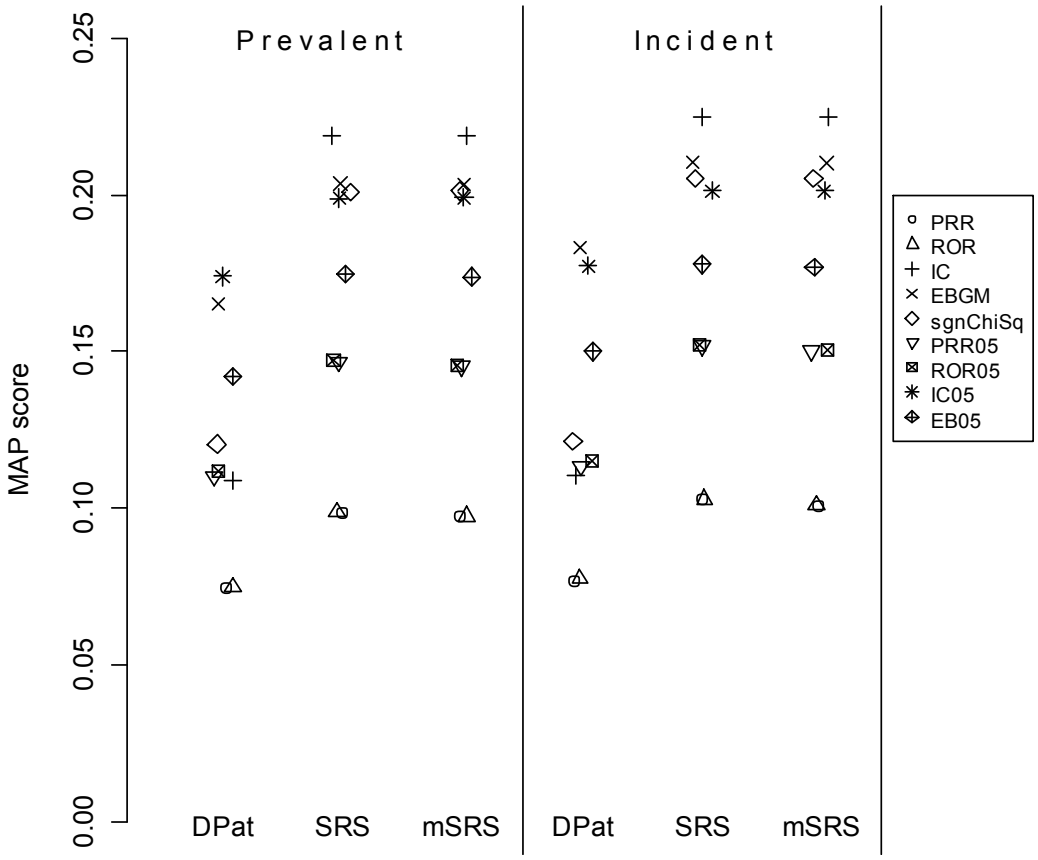

Figure 5. MAP Scores for DP Methods (simulated data). Scores for all six approaches to table construction and nine different scoring approaches.

6.3 Application to the Real Data

Whereas spontaneous reporting databases contain several million records, LODs can contain more than one order of magnitude in data size, so computational feasibility remains central to efficient analysis. In this study, all disproportionality analysis configurations were executed

against all drugs and all conditions across the 5 databases. In the largest database, CCAE, there were over 4.4 billion drug-condition co-occurrences identified over 1451 drugs and 11844 conditions. All analyses were able to complete in less than 24 hours.
Figure 6 shows various scores (on the logarithmic scale) for a random sample of 1000 drug-condition pairs for the SRS-prevalent scenario on the MSLR data. Key findings are similar to what we observed on the simulated data:

- PRR and ROR are very close to each other, as are PRR05 and ROR05;
- Bayesian shrinkage approaches, IC and EBGM, IC05 and EB05, show agreement on most of the drug/condition pairs;
- Scatterplots of PRR vs. EBGM (and other similar metrics) have several branches because amount of shrinkage depends on observed cell count;
- signed chi-square scores behave differently from other measures of disproportionality.

Similar findings apply to the SRS-incident, modified SRS-prevalent, modified SRS-incident, and distinct patient prevalent and incident scenarios.

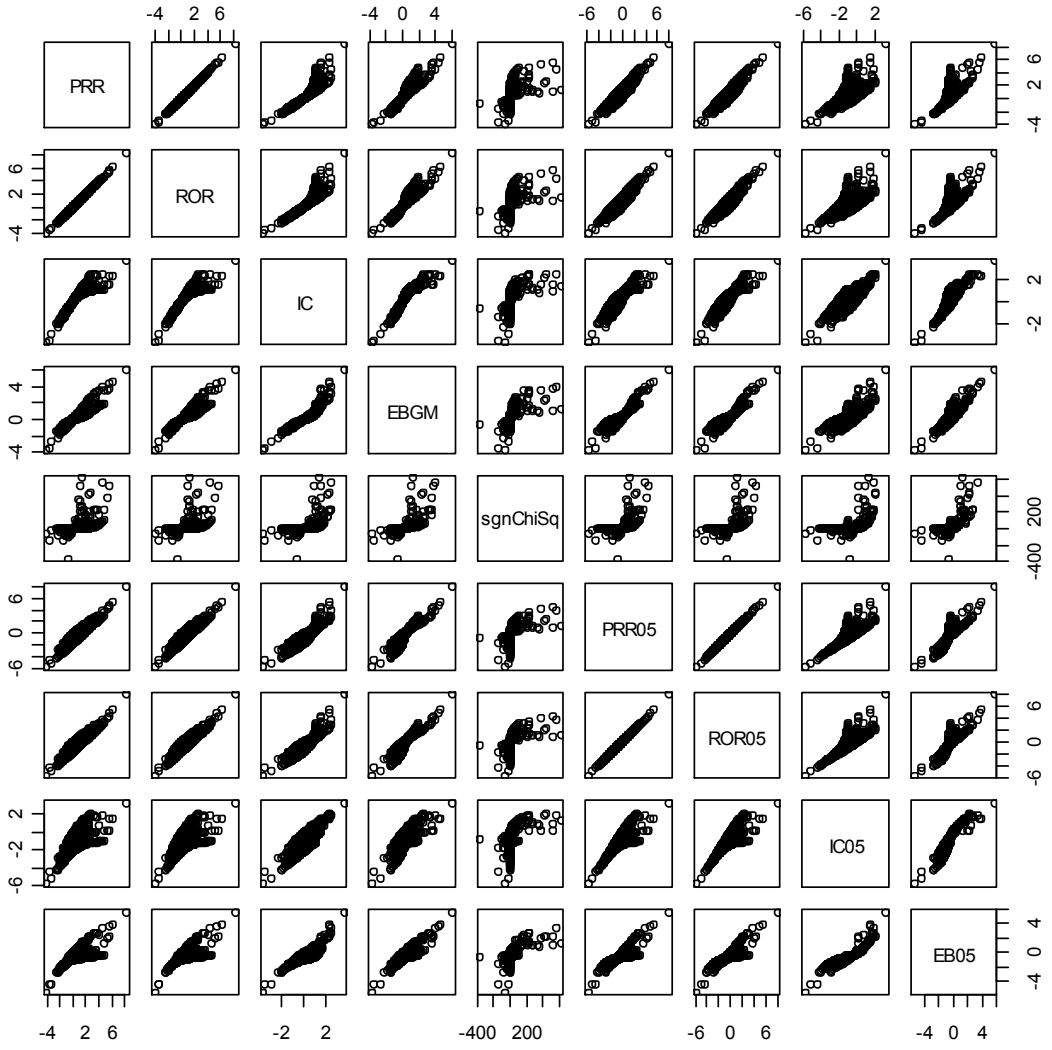

Figure 6. Prevalent events - SRS (MSLR). Scatterplots (on the logarithmic scale) of nine measures of disproportionality for the SRS – prevalent scenario on MSLR.

Figure 7 shows EBGM scores on MSLR data across all six table construction scenarios. As on the simulated data, EBGM scores on MSLR show good agreement for both event types, prevalent and incident. Distinct patient prevalent and incident scores are quite similar, as well as SRS-prevalent and SRS-incident, modified SRS - prevalent and modified SRS- incident. Within each event type, prevalent and incident, SRS and modified SRS scores are close.

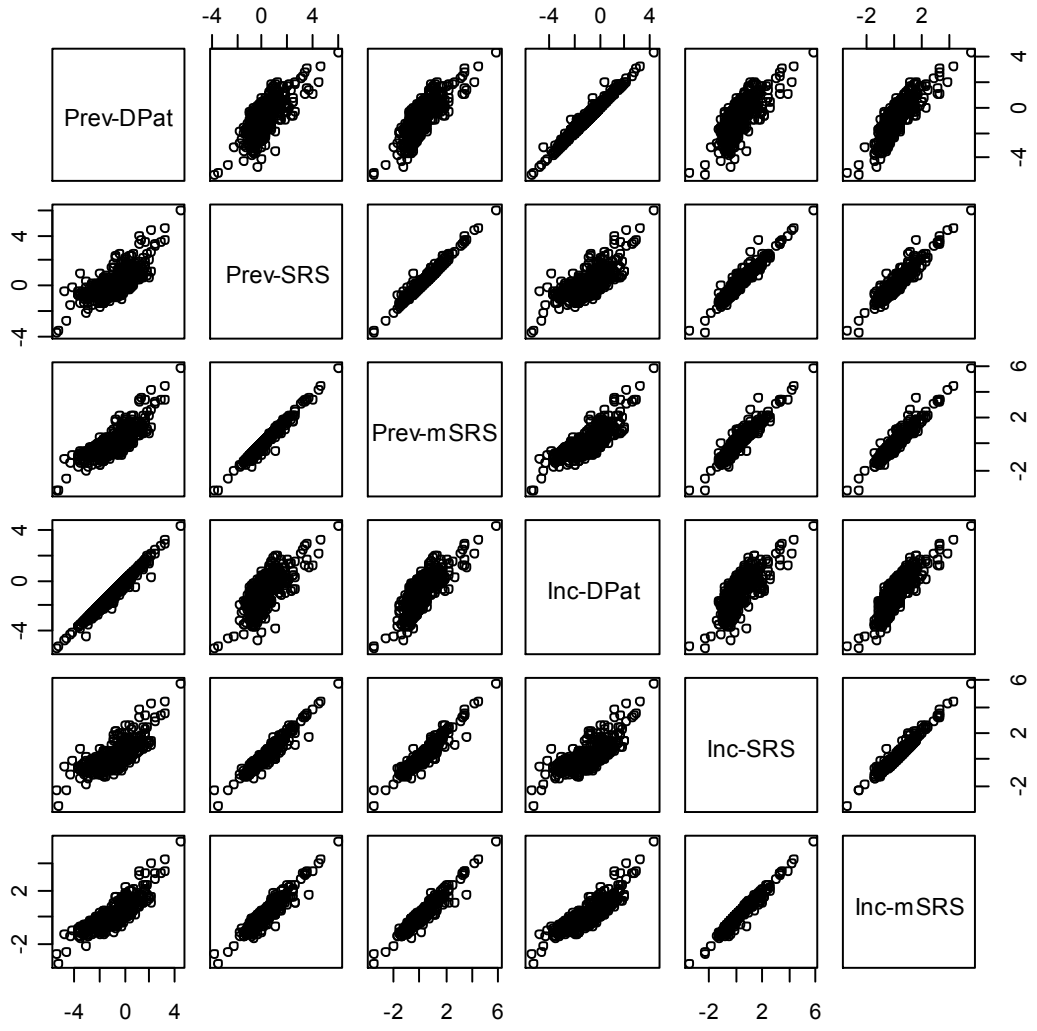

Figure 7. EBGM scores across six scenarios (MSLR data). Scatterplots of EBGM scores across six scenarios: distinct patients - prevalent, SRS - prevalent, modified SRS - prevalent, distinct patient - incident, SRS - incident, modified SRS - incident.

7. Conclusions

There is a significant interest in utilization of observational health care data in drug safety research. Statistical methods based on different measures of disproportionality were among the first employed in pharmacovigilance (Finney, 1971). Therefore it seems natural to extend these methods to the analysis of longitudinal data. This article is the first systematic attempt of such an extension.

We propose three counting approaches, distinct patients, SRS, and modified SRS, which in conjunction with two event types, prevalent or incident, lead to six different mappings of LOD into the form appropriate for disproportionality analysis.

In the numerical experiment, we considered nine measures of disproportionality. Among those nine, shrinkage approaches, IC and EBGM, showed the best performance on the simulated data with respect to MAP, closely followed by the derivative shrinkage measures, EB05 and IC05, and signed Chi-square test. No metric performed perfectly suggesting that further analysis of outputs of any dispropiortionality metric would be needed to prevent false positive and false negative results. Qualitative comparison of the results from the simulated dataset and MSLR data suggests that the same set of disproportionality measures and similar mapping approaches, SRS and modified SRS, may achieve the best performance on real data. Further research will help to test these hypotheses.

While spontaneous adverse drug reaction databases may contain up to several million reports, disproportionality analysis of observational health databases involves much larger amount of data because each person may contribute numerous reports over time. Nevertheless we are able to complete all calculations for disproportionality analysis even on a database that contains information on almost sixty million people. The feasibility of such large scale screening of LOD is promising. This opens new possibilities in pharmacovigilance research and, eventually, will contribute to improved drug safety. The results presented here will also serve as a benchmark for comparison when in the future other statistical methods are applied to the same data sets. Software for all of the methods described in this paper is freely available at http://omop.fnih.org.

References

Bate, A., Lindquist, M., Edwards, I.R, et al. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*, 54(4):315-321.

Bate, A., Evans, S.J., (2009). Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety,* 18 (6): 427-436.

Brown, J.S., Kulldorff, M., Chan, K.A., Davis, R.L., Graham, D., Pettus, P.T., Andrade, S.E., Raebel, M.A., Herrinton, L., Roblin, D., Boudreau, D., Smith, D., Gurwitz, J.H., Gunter, M.J., Platt, R., (2007). Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and Drug Safety*, 12:1275-1284.

Brown , J. S., Kulldorff, M., Petronis, K. R., Reynolds, R., Chan, K. A., Davis, R. L., Graham, D., Andrade, S. E., Raebel, M. A., Herrinton, L., Roblin, D., Boudreau, D., Smith, D., Gurwitz, J. H., Gunter, M. J., Platt, R., (2009). Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. *Pharmacoepidemiology and Drug Safety*, 18(3): 226–234.

Caster O., Norén G.N., Madigan D., and Bate A. (2010). Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. *Statistical Analysis and Data Mining*, 3(4):197-208.

Curtis, J.R., Cheng, H., Delzell, E., Fram, D., Kilgore, M., Saag, K., Yun, H., and DuMouchel, W. (2008). Adaptation of Bayesian data mining algorithms to longitudinal claims data. *Medical Care*, 46: 969-975.

Davis, R.L., Kolczak M., Lewis E.,Nordin, J., Goodman, M., Shay, D.K., Platt, R., Black, S., Shinefield, H., and Chen, R.T. (2005). Active Surveillance of Vaccine Safety Data for Early Signal Detection. *Epidemiology,* 16: 336-341.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat*, 53(3):170-190.

DuMouchel, W., Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. *Proceedings of the seventh ACM SIGKDD*, San Francisco, California, 67 - 76.

Evans, S.J., Waller, P.C., Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety,* 10(6):483-486.

Finney, D. J. (1971). Statistical logic in the monitoring of adverse reactions to therapeutic drugs. *Methods of Information in Medicine*, **10**, 237-242.

Fram, D., Almenoff, J., DuMouchel, W. (2003). Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Hauben, M. and Zhou, X. (2003). A brief primer on automated signal detection Quantitative methods in pharmacovigilance: focus on signal detection. *The Annals of Pharmacotherapy*, 37(7-8):1117-1123.

Hauben, M., Reich, L., Zhou, X. (2004). Safety related drug-labelling changes: findings from two data mining algorithms. *Drug Safety*, 27(10):735-744.

Hauben, M., Zhou, X. (2004). Trimethoprim-induced hyperkalaemia - lessons in data mining. *British Journal of Clinical Pharmacology*, 58(3):338-339.
Hauben, M., Madigan, D., Gerrits, C. M., Walsh, L., Van Puijenbroek, E. P., (2005),
The role of data mining in pharmacovigilance. *Expert Opionion on Drug Safety*, 5: 929-948.

Hauben, M., Reich, L., (2005). Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: A retrospective evaluation. *Journal of Clinical Pharmacology,* 45 (4): 378-384.

Hocine, M. N., Musonda, P., Andrews, N. J., Farrington, C., (2009), Sequential case series analysis for pharmacovigilance, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1): 213–236.
Jin, H., Chen, J., He, H., Williams, G.J., Kelman, C., and O'Keefe, C.M. (2008). Mining unexpected temporal associations: Applications in detecting adverse drug reactions. *IEEE Transactions on Information Technology in Biomedicine*, 12: 488-500.

Kulldorff, M., Davis, R.L., Kolczak, M., Lewis, E., Lieu, T., and Platt, R. (2008). A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Preprint*.

Li, L. (2009). A conditional sequential sampling procedure for drug safety surveillance. *Statistics in Medicine*, 28(25):3124-38.

Lieu, T.A., Kulldorff, M., Davis, R.L., Lewis, E.M., Weintraub, E., Yih, K., Yin, R., Brown, J.S., and Platt, R. (2007). Real-time vaccine safety surveillance for the early detection of adverse events. *Medical Care*, 45, S89-S95.

Madigan, D., (1999), Discussion of "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system", *Am Stat*, 53(3): 198 - 200.

Madigan, D., Vardi, Y., and Weissman, I. (2006). Extreme value theory applied to document retrieval from large collections. *Information Retrieval*, 9, 273-294.

Mammadov, M.A., Rubinov,A.M., and Yearwood, J. (2007). The study of drug–reaction relationships using global optimization techniques. *Optimization Methods and Software,* **22,** 99-126.

McClure, D.L., Glanz, J.M., Xu, S., Hambidge, S.J., Mullooly, J.P., and Baggs, J., (2008). Comparison of epidemiologic methods for active surveillance of vaccine safety. *Vaccine*, 26: 3341-3345.

Norén, N.G., Bate, A., Orre, R. and Edwards, I.R. (2006). 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*, **25**, 3740-3757.

Norén, G.N., Bate, A., Hopstadius, J., Star, K., and Edwards, I.R. (2008). Temporal pattern discovery for trends and transient effects: its application to patient records. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* - KDD '08.

Norén, G.N., Hopstadius J., Bate A., Star K., Edwards I.R., (2010). Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3):361-387.

Orre, R., Bate, A., Norén, G.N., Swahn, E., Arnborg, S., and Edwards, I.R. (2005). A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets *International Journal of Neural Systems*, **15**, 207-222.

Praus, M., Schindel, F., Fescharek, R., and Schwarz, S. (1993) Alert systems for post-marketing surveillance of adverse drug reactions. *Statistics in Medicine* **12**, 2382-2393.

Rothman, K.J., Lanes, S., Sacks, S.T., (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and Drug Safety,* 13(8):519-523.

Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H., and Brookhart, M.A. (2009). High-dimensional propensity scoring adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20: 512-522.

Szarfman, A., Machado, S.G., O'Neill, R.T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Safety, 25(6):381-392.

vanPuijenbroek, E.P., Egberts, A.C., Heerdink, E.R., and Leufkens, H.G. (2000). Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs European journal of Clinical Pharmacology. 56, 733-738.

van Puijenbroek, E. P., Bate, A., Leufkens, H.G.M., Lindquist, M., Orre, R., and Egberts, A.C.G. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf*, **11**, 3-10.

Walker, A.M., (2010). Signal detection for vaccine side effects that have not been specified in advance. *Pharmacoepidemiology and Drug Safety*, 19(3):311-7.

APPENDIX A. Formulae for measures of disproportionality

Let us assume that for each stratum i we calculated 2-dimesional summary that is shown in table 6. Let N_i be the sum of all cell counts, $N_i = W_{i00} + W_{i01} + W_{i10} + W_{i11}$.

Table 6. Counts for the ith stratum.

	AE Yes	*AE No*
Drug Yes	W_{i00}	W_{i01}
Drug No	W_{i10}	W_{i11}

Using these counts we can define stratified measures of disproportionality.

PRR, proportional reporting ratio*:*

$$PRR = \frac{\sum_i W_{i00} * (W_{i10} + W_{i11}) / N_i}{\sum_i W_{i10} * (W_{i00} + W_{i01}) / N_i}.$$

PRR05, left bound of the 90% confidence interval for PRR*:*

$$PRR05 = PRR \cdot \exp\left(-1.645 \cdot \left(\frac{\sum_i ((W_{i00} + W_{i01})(W_{i10} + W_{i11})(W_{i00} + W_{i10}) - W_{i00} * W_{i10} * N_i) / N_i^2}{\sum_i W_{i00}(W_{i10} + W_{i11}) / N_i \sum_i W_{i10}(W_{i00} + W_{i10}) / N_i} \right)^{1/2} \right).$$

ROR, reporting odds ratio*:*

$$ROR = \frac{\sum_i W_{i00} W_{i11} / N_i}{\sum_i W_{i10} W_{i01} / N_i}.$$

ROR05, left bound of the 90% confidence interval for ROR:

$$ROR05 = ROR \cdot \exp(-1.645 \cdot \sigma),$$

$$\sigma^2 = \frac{\sum_i (W_{i00} + W_{i11}) W_{i00} W_{i11} / N_i^2}{2(\sum_i W_{i00} W_{i11} / N_i)^2} + \frac{\sum_i (W_{i10} + W_{i11}) W_{i01} W_{i10} + (W_{i01} + W_{i10}) W_{i00} W_{i11} / N_i^2}{2 \sum_i W_{i00} W_{i11} / N_i \sum_i W_{i01} W_{i10} / N_i} + \frac{\sum_i (W_{i01} + W_{i10}) W_{i01} W_{i10} / N_i^2}{2(\sum_i W_{i01} W_{i10} / N_i)^2}.$$

IC, Information component*:*

$$IC = \log_2 \left(\frac{\sum_i W_{i00} + 1/2}{\sum_i \frac{(W_{i00} + W_{i01})(W_{i00} + W_{i10})}{N_i} + 1/2} \right).$$

IC05, lower credibility limit for 90% credibility interval for IC*:*

$IC05 = \log_2(z)$, z is the solution to equation $\int_0^z \frac{(e+1/2)^{n+1/2}}{\Gamma(n+1/2)} \lambda^{n+1/2-1} e^{-(n+1/2)\lambda} d\lambda = 0.05,$

where $e = \frac{(W_{i00} + W_{i01})(W_{i00} + W_{i10})}{N_i}$ and $n = \sum_i W_{i00}$.

For more details regarding IC and IC05 see (Norén et al. 2008, Bate et al. 1998).

Signed Shi-square*:*

$$\operatorname{sgn} ChiSquare = sign\left(\sum_i W_{i00} - \sum_i (W_{i00}+W_{i01})(W_{i00}+W_{i10})/N_i\right)\frac{\left(\sum_i W_{i00} - \sum_i (W_{i00}+W_{i01})(W_{i00}+W_{i10})/N_i\right)^2}{\sum_i \left((W_{i00}+W_{i01})(W_{i10}+W_{i11})(W_{i00}+W_{i10})(W_{i01}+W_{i11})/(N_i^2(N_i-1))\right)}.$$

Details regarding EBGM, geometric mean of the empirical Bayes estimate of the posterior distribution of the reporting ratio, and EB05, its 5th percentile, can be found in reference (DuMouchel, 1999, DuMouchel and Pregibon, 2001).

APPENDIX B. Mapping approaches for longitudinal data

This appendix provides a formal mathematical definition of the alternative methods of constructing two-by-two tables from longitudinal data for disproportionality analysis.

Notation

Let $y_{ict} = 1$ if patient *i* has condition *c* at time *t*, and $y_{ict} = 0$ otherwise, *i*=1,…,*I*, *c*=1,…*C*, and *t*=1,…,*T*. Let $x_{idt} = 1$ if patient *i* "takes" drug *d* at time *t*, and $x_{idt} = 0$ otherwise, *i*=1,…,*I*, *d*=1,…*D*, and *t*=1,…,*T*. This may include the user-defined off-drug "surveillance" window. Let $y^*_{ict} = 1$ if $y_{ict} = 1$ and $y_{ics} = 0$ for all $s < t$, 0 otherwise. Let $z_{it} = 1$ if patient *i* has coverage at time *t*, 0 otherwise.

Let D_{id} be the set of all ordered pairs (*r*,*s*), *r*,*s* ∈ {1,…,*T*}, where $x_{idr} = 1$ and (*r*=1 or $x_{id(r-1)}$ =0), $x_{ids} = 1$ and (*s*=*T* or $x_{id(s+1)}$=0), and $y_{ict} = 0$ for all *c* and all *t* ∈ [*r*,…,*s*], and $z_{it} = 1$ for all *t* ∈ [*r*,…,*s*]. This defines condition-free periods of continuous drug exposure.

Define *I*(*x*) = 1 if *x*>0, 0 otherwise.

Prevalent Conditions, Distinct Patients

$$w_{00} = \sum_i I(\sum_t x_{idt} y_{ict})$$

$$w_{01} = \sum_i \left\{ I(\sum_t z_{it} x_{idt}) - I(\sum_t x_{idt} y_{ict}) \right\}$$

$$w_{10} = \sum_i \left\{ (1 - I(\sum_t z_{it} x_{idt})) \times I(\sum_t z_{it} y_{ict}) \right\}$$

$$w_{11} = \sum_i \left\{ (1 - I(\sum_t z_{it} x_{idt})) \times (1 - I(\sum_t z_{it} y_{ict})) \right\}$$

Prevalent Conditions, SRS

$$w_{00} = \sum_i \sum_t x_{idt} y_{ict}$$

$$w_{01} = \sum_i \sum_t \sum_{c' \neq c} x_{idt} y_{ic't}$$

$$w_{10} = \sum_i \sum_t \sum_{d' \neq d} x_{id't} y_{ict}$$

$$w_{11} = \sum_i \sum_t \sum_{d' \neq d} \sum_{c' \neq c} x_{id't} y_{ic't}$$

Prevalent Conditions, Modified-SRS

$$w_{00} = \sum_i \sum_t x_{idt} y_{ict}$$

$$w_{01} = \sum_i \left\{ \sum_t \sum_{c' \neq c} x_{idt} y_{ic't} + |D_{id}| \right\}$$

$$w_{10} = \sum_i \left\{ \sum_t \left\{ \sum_{d' \neq d} x_{id't} y_{ict} + y_{ict} z_{it} (1 - I(\sum_d x_{idt})) \right\} \right\}$$

$$w_{11} = \sum_i \left\{ \sum_t \sum_{d' \neq d} \sum_{c' \neq c} x_{id't} y_{ic't} + \sum_{d' \neq d} |D_{id'}| + \sum_t \sum_{c' \neq c} z_{it} y_{ict} (1 - I(\sum_d x_{idt})) \right\}$$

For incident conditions replace y_{ict} in the above definitions by y^*_{ict}.