PROBLEM 6: SIGNAL DETECTION FOR DRUG SAFETY

Viviana García[1], Peng He[2], Xuefei Jia[3], Andrea Knezevic[4], Ying Lu[5], Mahmoud Shehadeh[6], Ranye Sun[7]

Problem Presenter:
Mark Wolffe
SAS Institute, Inc.

Faculty Mentor:
Chia Ying Lee
SAMSI

Abstract

Post-market vigilance of drug safety has been legally mandated for pharmaceutical companies and regulatory agencies. However, new developments in the quantitative methodologies of what has been called the *science of safety* have been scarce in the past few decades. The main source of post-market data for potential drug induced adverse events (AE) are spontaneous reporting systems (SRS) such as the Adverse Event Reporting System (AERS) managed by the Food and Drug Administration (FDA). The goal of analysis of these reporting systems is detection of new and unexpected drug-AE relationships that may be of potential harm to the public; in the literature this is referred to as signal detection. In this paper we review existing quantitative methods for signal detection in SRS that are in widespread use, the so-called disproportionality analysis (DA) methods. We identify known drug-AE relationships using historical data on FDA labelling changes and use AERS data on these pairs as case studies. We analyze these case studies using the existing methods, employing novel approaches of signal detection over demographic strata and over time. Using our case studies, we find that these analytic approaches are potentially valuable. Furthermore, we generate simulated SRS data for the purpose of testing the sensitivity of the existing DA methods. From this exercise we conclude that more simulation should be done and we strongly advocate the development of a reference database on which to test these DA methods. We conclude that DA in the context of signal detection in SRS are an important tool for pharmacovigilance and we conclude that the development of more sophisticated statistical methods to deal with the unique and complex problems presented by analysis of SRS are valuable.

1 Introduction and Motivation

Pharmacovigilance concerns the monitoring and detection of adverse events associated with the use of medicines. This process starts with designed clinical trials, and continues throughout the drug's life cycle after approval, when its use is widespread among the population.

In the post-approval environment, the primary method of data collection for surveillance purposes comes from spontaneous reporting systems (SRS), such as the Adverse Event Reporting System (AERS) of the Food and Drug Administration (FDA). These tools produce databases which contain a collection of reports of side effects, all of which are submitted voluntarily by clinicians, patients, or product manufacturers. Each report in an SRS database typically includes limited demographic information (such as age, sex, and weight), one or more drugs, and one or more adverse events.

The objective of creating these systems was to provide data that allows for the investigation of possible safety problems associated with the use of drugs, since some of these would be impossible to detect during the limited run of a clinical trial. In addition, clinical trials are unlikely to reliably detect rare, serious adverse

[1]Harvard University
[2]University of Cincinnati
[3]Georgetown University
[4]Georgia Institute of Technology
[5]University of Minnesota
[6]Northern Illinois University
[7]Texas A&M University

events that occur in subpopulations who have not participated in studies. Furthermore, as new medical products enter the market, the potential for interactions with other drugs, biologics, medical devices, and foods increases. Detecting possible relationships between drugs and adverse events in a timely fashion could prove extremely important to public health. Signal detection can be the first indication that a certain association should be studied more closely. It could also be informative for pharmaceutical companies as a continued testing scheme, to avoid potential lawsuits and to comply with the FDA regulations regarding surveillance.

In 2007, active post-market drug safety surveillance and analysis was mandated by a law passed in Congress, the Food and Drug Administration Amendments Act. In response, the FDA put in place the Sentinel Initiative with the ultimate goal of creating and implementing a national, integrated, electronic system for monitoring medical product safety. The Sentinel Initiative represents the implementation of what the FDA calls the *science of safety* which combines medical and pharmacological data with quantitative methods, with the goal to "generate hypotheses about, and confirm the existence and causal factors of, safety problems in the populations using the products"[2].

Hence the appeal for the development of analytic methods that might help identify possible starting points of interest. In particular, the drug safety literature often uses the term *signal* to refer to early hints that point at the possibility of novel and unintended drug effects. The approach of this investigation has to do with signal detection in SRS databases.

However, as can be imagined, there are a number of limitations inherent to the type of databases obtained from SRS. These issues should be noted and kept in mind when formulating conclusions or making decisions based on the information provided by the data. We just briefly mention some of them in this document.

First of all, these datasets are incredibly large and disorganized. To give an idea of their magnitude, the FDA receives more than 400,000 of these spontaneous reports each year [1]. The lack of a standardized nomenclature for drug names (including different names for the same drugs, misspellings, or the inclusion of dosages with the name, among others) and the use of multiple terms for similar clinical conditions presents a challenge.

Several problems appear due to the voluntary nature of the reporting process. One is the serious problem of over-reporting, which could occur for example because of the influence of publicity or a warning set up on a certain drug. In addition, there could be under-reporting, which may depend on the event and its severity, for example, or the lack of knowledge of the reporting system. Finally there could be multiple submissions: for example, when a person is taking a combination of drugs the report might sent to all of the manufacturers, who in turn file separate reports to the FDA. Additional problems present themselves because many reports of events do not necessarily reflect associations to the drugs that they allude to, and because there is limited information regarding the order of exposure and condition, or even the duration of exposure. Most importantly, SRS databases don't contain information about the number of patients at risk, that is, the population that was exposed to a specific drug. In short, there is considerable bias and noise in the data that undermines its reliability.

It is extremely important to note that any conclusions obtained from these databases *cannot* establish causality. At best, the analyses might identify potential issues and associations that must be confirmed by expert epidemiologists and clinicians in follow-up studies. Actually, many signals that emerge from spontaneous report databases are mostly noise, because there are many factors that are intermingled in a report such as treatment indications, co-prescribed drugs, reporting artifacts, etc., or because the reported adverse events are already labeled, are medically trivial, or biologically implausible.

Despite the many limitations of the available datasets necessary for post-approval analyses, there is an interest in the pharmacology community to develop analytic methods to quantify and detect signals that might appear in the spontaneous report databases, since these are the only sources of information currently available about drugs once they are widespread in the market. As previously mentioned, there are thousands of drugs and thousands of adverse events (AEs) that need to be studied. The complexity of these large datasets makes drawing inferences about the extremeness of drug-event counts intractable without the help of quantitative

summaries and analysis. This difficulty, in addition to the terms of the recently approved Sentinel Initiative, has generated more interest in these methods on the part of the regulators, the health care community, and industry.

There are multiple methods of signal detection presently in use, and one of the objectives of this project is to understand the current approaches and to explore and identify potential modifications or areas of improvement for these methods. Additional objectives include analyzing a dataset for a particular drug-adverse event that has not been investigated before, and exemplifying the importance of stratification based on demographic covariates.

This document is organized as follows. In section 2 we describe the AERS database and the particular subsets that we used for analysis. In section 3 we outline the four most commonly used signal detection methods, and highlight some of their advantages and limitations. In section 4, we propose three novel sequential methods to detect possible signals from a time series of disproportionality scores. Section 5 introduces three case studies constituted by specific pairs of drug and adverse event, which will serve as examples for our new methodologies. In section 6 we present our results. We first applied the four well-known signal detection methods to a particular case study and exemplified the importance of stratification to control for demographics. Secondly, we show the results of applying our new longitudinal signal detection methods to 3 case studies. The last part in this section concerns a simulation study. Finally, we wrap up with our conclusions and future work. All the analyses performed in this paper were done using available SAS software.

2 Description of the AERS Database

The data that we have used throughout this research is a subset of the Adverse Event Reporting System (AERS) database, which contains information of medical adverse events reported to the FDA. This database is publicly available, and it is updated every 3 months, which means that the reports are grouped in quarters per year. Data is available at the FDA website[8] for the first quarter of 2004 through the first quarter of 2012, and we focused on this specific subset of data corresponding to 33 quarters.

For each quarter, the AERS database consists of six major segments, including separate files for demographic, drug, reaction, patient outcomes, report sources, and therapy dates information. These datasets are connected by a primary link key with a unique number that identifies the AERS reports, as can be seen in Figure 1.

The demographic dataset contains 231,945 unique reports. Most of them are from United States (162,336), Japan (11,199), Germany (7,313), France (7,022), United Kingdom (6,872), and Canada (6,015). Among these reports, there are $169,272$ initial reports and $62,673$ follow up reports. The number of reports by females exceeds the one for males by over 50,000 reports, and there are also unknown and unspecified genders reported. Most of these reports were issued by consumers, followed by lawyers, and then medical doctors. 33,579 unique drugs and 9,516 adverse events are included in the reports. In the outcome dataset, 161,252 unique reports were filed. As a final outcome, 68,951 of cases were hospitalized, 29,138 cases of death, and 8,044 cases reported as life-threading.

One part of our analysis was conducted only on the data pertaining to the first quarter of 2012, and will appear in Section 6.1. The longitudinal part of the analysis will consider the information in all 33 datasets, and is detailed in Section 6.2.

For the purpose of visualization, we constructed a drug-adverse event network for 1,000 combinations of drugs and adverse events from the first quarter of 2012 of the AERS database (Figure 2). The network was drawn with the open source software Cytoscape[9]. The nodes represent both drugs and adverse events, while

[8] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm
[9] http://www.cytoscape.org

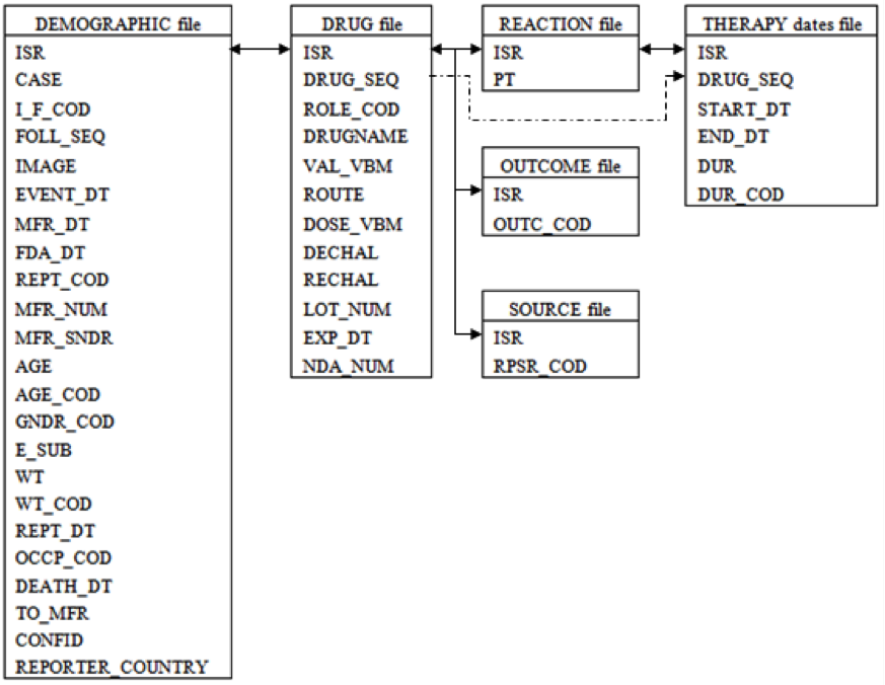

Figure 1: AERS database description.

the edges denote the relationships between them.

3 Existing methods for signal detection

The objective of signal detection methods is to filter the dataset to try to obtain evidence of potential associations between drugs and adverse events that were not known before, by providing a measure of how rare or common a particular combination is. Disproportionality analysis methods comprise the most widely used class of analytic methods for signal detection in SRSs. These methods quantify the extent to which a given condition is "disproportionally" reported with a given drug , compared to what would be expected (a control). In other words, an observed/expected ratio of probabilities or cell counts is obtained.

Considering the limitations of the databases mentioned above, it is clear that there is no real "control group". That is, since all the reports come in a voluntary basis, it is impossible to know how many people were exposed to the drug, how many people actually experienced an event, or even how many people experienced a particular event after taking a specific drug. This gives rise to a big complication in the quantification of the rarity of an adverse event, since without the total exposures it is difficult to evaluate the importance of its occurrence. To put this in terms of the mentioned disproportionality methods, the expected counts cannot be computed directly for any drug-adverse event pair.

The existing methods try to compensate for the fact that it is impossible to quantify a drug-adverse event rate directly, by using all other drugs and all other events in the dataset as a control (or background noise) against which to compare. Therefore, they focus on low-dimensional projections of the data, particularly 2-dimensional contingency tables, of the form shown in Table 1. The difference between the methods is the way in which the expected counts are modeled.

The most commonly used methods are the proportional reporting ratios (PRR), reporting odds ratios (ROR), the multi-item gamma Poisson shrinker (MGPS), and the Bayesian Confidence Propagation Neural Network (BCPNN). PRRs and related measures based on 2x2 contingency tables are currently used in routine

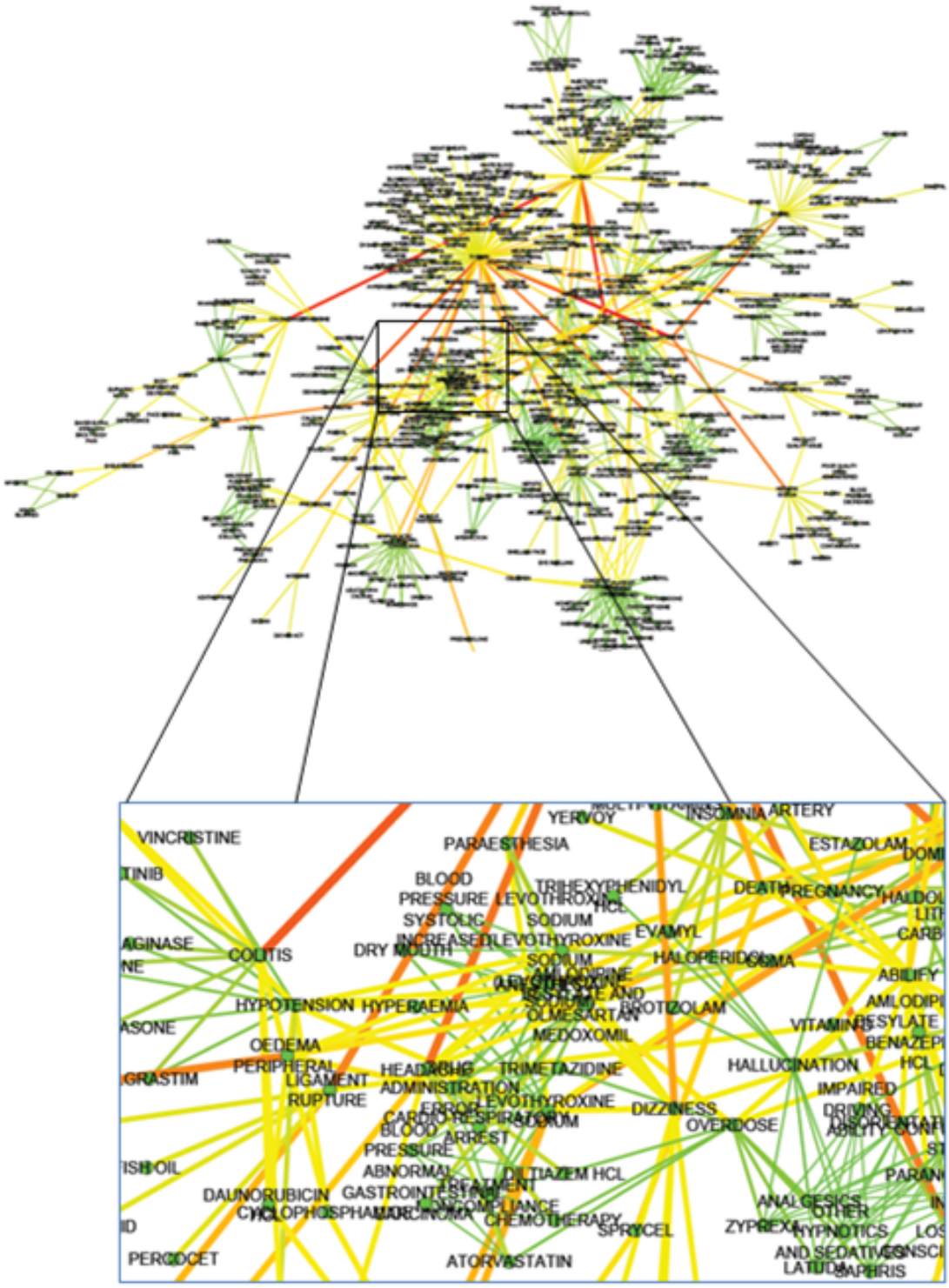

Figure 2: Network representation of drug and adverse event relationships in AERS database

pharmacovigilance activities by the Medicine Control Agency (MCA) in the UK. MGPS is currently used by the FDA, and BCPNN is employed by the World Health Organization (WHO)[7]. We give a brief description of these methods in the following sections.

	$\mathbf{AE_j = Yes}$	$\mathbf{AE_j = No}$
$\mathbf{Drug}_i$ = Yes	n_{00}	n_{01}
$\mathbf{Drug}_j$ = No	n_{10}	n_{11}

Table 1: Typical table for disproportionality analysis. AE stands for adverse event.

3.1 Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR)

The PRR is a very simple method inspired by the well-known relative risk calculation for contingency tables. By just focusing on a specific drug-adverse event combination, and pooling the counts over all other drugs and all other adverse events, it is possible to construct a 2x2 table as the one shown in Table 1. Then the PRR is computed as

$$PRR = \frac{n_{00}/n_{0\cdot}}{n_{10}/n_{1\cdot}}$$

where $n_{0\cdot} = n_{00} + n_{01}$, and analogously for $n_{1\cdot}$.

The ROR is very similar to the PRR, except for the fact that it tries to correct for certain kinds of under-reporting. It is calculated from the same 2x2 table (Table 1) as the PRR.

$$ROR = \frac{n_{00}n_{11}}{n_{01}n_{10}}$$

The interpretation of these quantities is that they measure how much more frequently the specific event is reported with the chosen drug, than with all other drugs.

It is important to keep in mind that whenever the count n_{00} is very small (which often happens in this type of datasets), this leads to substantial variability which increases the uncertainty about the true value of the measure of association to be computed. A known problem with PRR and ROR is that they do not address this issue, that is, there is no way to quantify the variability associated to this "sampling" variation. The two Bayesian methods that we will proceed to describe improve upon the methods based on relative ratios by addressing this issue, and provide solutions by considering all the reported drug-adverse event combinations at the time.

3.2 Bayesian approaches

The multi-item gamma Poisson shrinker (MGPS) and the Bayesian Confidence Propagation Neural Network (BCPNN) are Bayesian methods that aim to express possible associations between the reporting of events and drugs in terms of a function of the ratio of observed to expected frequencies mentioning drug i and adverse event j, n_{ij}/E_{ij}. That is, they look at a specific drug-event combination and try to quantify how "interestingly large" the number of reports is compared to what would be expected under the assumption of drug and event being statistically independent [7]. The expected counts E_{ij} are computed as

$$E_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

where

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{I} n_{ij}, \quad n_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

are the total counts corresponding to drug i, the total counts corresponding to adverse event j, and the total number of reports, respectively. Throughout this document we denote the total number of drugs in a particular database by I, and the total number of adverse events by J.

In particular, the measure of disproportionality for a specific drug-adverse event combination is the *information criterion* (IC), computed as

$$IC_{ij} = log_2 \left(\frac{n_{ij}}{E_{ij}} \right)$$

which is just the logarithm base 2 of the PRR.

3.2.1 Multi-item Gamma Poisson Shrinker (MGPS) - EBGM

In this method [5], each observed count for drug-adverse event pair is modeled as a draw from a Poisson distribution with varying unknown means. The means are considered to be random with a common prior distribution: a mixture of two gamma distributions (which have in total 5 parameters). In addition, an Empirical Bayes procedure is used to estimate the 5 parameters from the prior. In short, the model is

$$\begin{aligned} n_{ij} &\sim Poisson(\mu_{ij}) \\ \mu_{ij} &= \lambda_{ij} E_{ij} \\ \lambda_{ij} &\sim \text{Mixture Gamma} \end{aligned}$$

Finally, the posterior distribution of λ_{ij} is obtained, and the EBGM, defined as the geometric mean of the empirical Bayes posterior distribution of the true relative report ratio, is reported. This method is known as the Gamma Poisson Shrinker (GPS). As previously mentioned, the Bayesian methods try to account for the "sampling" variability in the reported counts, and studies have shown that the EBGM method does well even with very small n_{00} (even 1 or 2).

A variant of the above method, the Multi-Item Gamma Poisson Shrinker (MGPS), allows for higher order combinations of drugs and events that are significantly more frequent than their pairwise association would suggest.

3.2.2 BCPNN

The BCPNN method [4] is similar to EBGM, but uses a multinomial model instead of a Poisson for the counts, and calculates all cell counts for all potential drug-adverse event combinations in the database (not just those that appear together in at least one report). The fact that it is embedded in a neural network gives it the ability of to handle large data sets, and is robust to missing data.

In this case there is actually a proper prior (not estimated from the data as in the empirical Bayes approach), which is taken from the family of Beta distributions. Again a Bayesian procedure is used to obtain the posterior distribution of the IC between specific drugs and events present on the same report, as well as the 95% confidence intervals. In particular, an IC with a lower 95% confidence interval bound that increases with sequential time scans establishes a criterion for signal detection.

3.3 Discussion of the methods

Although these methods have widespread use as we have previously mentioned, we wish to highlight a few issues that became apparent when studying them in detail. We hope this will help practitioners to stay aware of the advantages and limitations of the existing methods, and to take them into account for interpretation of the results or development of new techniques.

The PPR and ROR are methods that are easily interpretable by practitioners because of the analogies that can be drawn to relative risks in epidemiology, and they simplify the problem to 2×2 contingency tables. If we consider that the dataset comprises I different drugs and J different adverse events in total, one would

essentially need to construct at total $I \times J$ tables of size 2×2 in order to analyze all the possible pairwise drug-adverse event combinations with these methods. But in this case, we run into a multiple testing scenario: we are running $I \times J$ tests on the same database without adjusting the family-wise error rate to account for the multiple comparisons, which will result in a large number of spurious relationships (or false-positive signals) that appear just by chance. That is, with millions of ratios being calculated, large ratios will inevitably appear just by chance, without necessarily meaning that there might be some interesting association. It is difficult to think how to incorporate an adjustment into the analysis, because the well-known methods such as the Bonferroni correction will potentially be too conservative, since the number of comparisons is huge (I and J are very large, and for example in our Case Study 1, we had an average of 2.7 million drug-adverse event combinations to look into).

Another issue that affects PPR and ROR is a small n_{00} count in the 2×2 table, as previously discussed. Analogously, a large count of a particular kind of adverse event report can potentially inflate the denominator for a specific drug, and reduce the sensitivity in detecting other signals associated with that drug. Finally, in comparing the frequentist methods (PRR and ROR) to the Bayesian methods, the first don't take into account the variability associated to the estimation of the measures of disproportionality, whereas the Bayesian methods do, by computing the entire posterior distribution.

In terms of the BCPNN, it seems to us that there is no available way to adjust for any stratification variables. As we will exemplify in our Section 6.1, stratification is important in order to control for demographic characteristics of the subjects, that is, in order to avoid spurious associations due to bad specification of the population that takes the drug.

The MPGS method seems to be very flexible in terms of its ability to potentially include drug-drug interactions (by creating a "new" drug which combines the counts for the two drugs included in the interaction), or even higher level interactions. The limitations of this ability would come in terms of the computational challenge of increasing the number of parameters to be estimated.

As we mentioned previously, in the MPGS-EBGM method, the estimate is a summary of the empirical Bayes posterior distribution of the true relative report ratio. Shrinkage towards the mean is a nice property derived from the fact that we are using an empirical Bayes approach to estimate the parameters of the prior distribution. What this method does is that it shrinks the calculated ratios in cases where the uncertainty is large (that is, when the variance in the estimate of the ratio is large), as would happen in the case of a really small n_{00} count. When this count is in the range of say, 10 to 20, there would be only a slight shrinkage; for large counts (e.g. over 100) there would be no shrinkage. This helps mitigate the peaks that would be obtained using a method like PRR, and therefore these estimates more stable in compareison.

All the methods suffer from the fact that all the calculations that are done are very dependent on what drugs and adverse events are included in the database. This is, we believe, partially inherent to the problem at hand since the whole issue is that there is no control group to compare the specific drug-event pair to, and so the "other drugs" and "other adverse events" that are used in the relative calculations will have a deep impact on the results. For example, if there are drugs that are included in the "control group" which have very high signals for the event of interest, the denominator would be inflated, which in turn would dilute the association that is the target of the particular analysis. Maybe some effort should be put into trying to define what group of drugs or events should be included in a specific analysis. This might include some sort of grouping by defined similarities, for example, or just following the same group of drugs across the different time periods.

As we have commented in several occasions, it is important to keep in mind that none of the conclusions that we obtain with these disproportionality methods should be interpreted as causal since there is no properly controlled randomized experiment involved. These methods can be useful as detectors of possible association between specific drug-adverse event combinations, that is, signals that can be identified for further study in a medical context.

4 Longitudinal signal detection

In this section we look at the historical score for a particular drug-adverse event combination. This serves as a reference for the physicians to judge whether the current score is off the track. We analyze the three drug-adverse event combinations that will be described in more detail in Section 5, namely Avandia & myocardial infarction, Finasteride & sexual dysfunction, and Thiazolidinedione & macular edema, from the first quarter of 2004 to the first quarter of 2012 (33 quarters overall). An EBGM score is calculated for each case, for each quarter. We decided to focus on the analysis of the EBGM scores because of its shrinkage and smoothness properties discussed in Section 3. We are particularly interested in whether there was a signal of disproportionality from the trend in the past. In this section, we present the methods that were developed in order to deal with this longitudinal analysis, and in Section 6.2 we discuss the results of applying them on our three case studies.

Our starting point is a time series of scores (say EBGM) for a specific drug-adverse event combination, which we denote by $\{X_i\}_{i=1}^N$, where i corresponds to the time period and N is the total number of time periods considered. If one were to plot it, by simply looking at the curve one would be able to spot certain spikes that might be deemed as signals. To conduct a more rigorous longitudinal analysis, we propose three novel ways to quantify sudden spikes that may potentially be signals: a method based on percent changes across time, a parametric approach and a non-parametric approach.

4.1 Method 1: Percent change in disproportionality score, relative to moving average

One way to quantify a sudden spike in a time series of disproportionality scores is to look at percent change in the score relative to the past. We can compare the score to a moving average of, for example, 1 year of scores. Percent change would thus be calculated as change relative to the average of the past 4 quarters:

$$\text{Percent change} = X_i \left(\frac{X_{i-1} + X_{i-2} + X_{i-3} + X_{i-4}}{4} \right)^{-1} - 1$$

Percent changes above some value, say 100% (that is, a doubling in value) may be considered worthy of investigation.

4.2 Method 2: Parametric approach

For the parametric approach, assume that the time series data $\{X_i\}_{i=1}^N$ comes from a Gaussian distribution. If no trend occurs, $\{X_i\}_{i=1}^N$ would be independent observations from a $N(\mu, \sigma)$ distribution. If there is a upward signal at time point $\tau + 1$, then $\{X_i\}_{i=1}^{\tau} \sim N(\mu, \sigma)$ independently, and $X_{\tau+1}$ would fall in the upper tail of Gaussian distribution. To implement this method, we assume that no signal occurs for the first four time points (which are used as a baseline). Starting from the fifth time point, we decide whether the current value is within two standard deviations of the mean. If so, we include that time point into the baseline, and re-estimate the mean and variance for the Gaussian distribution for further detection. If not, we report a signal. This sequential procedure continues until a signal is reported, that is, until the current time point is outside two standard deviations of the mean, where the mean and standard deviation are both estimated based on all the previous values.

4.3 Method 3: Nonparametric approach

The third algorithm we propose is the bootstrap approach. The key idea is as follows: if no trend occurs, we would expect the slope from the ordinary least squares (OLS) fit to be close to zero. On the other hand, if an upward trend occurs, we would expect the slope to be positive. To implement the method, again assume that the first four points do not show any trend. Starting from the fifth time point, we compute the OLS slope using

all the previous points and the current point. Its value is recorded as our test statistic. Next, we bootstrap from the previous points, say 10000 times, and compute the OLS slope each time. The p-value is calculated by computing the proportion of simulated slopes that are greater than our slope statistic. If the p-value is small, say less than 0.05, evidence exists that an upward trend is highly likely to occur. On the other hand, if the p-value is relatively large, we do not have evidence of an upward trend, and we include the current time point in the baseline group and proceed to check the next time point. This sequential procedure continues until a p-value is under a pre-defined threshold. The most commonly used threshold is a 0.05 level of significance.

In the end, these three methods may serve the following two purposes. First, they can be used to examine whether there was a signal in the past (for example, in a retrospective study to decide whether there was enough information to have captured certain signals before the public health situation became more difficult). Secondly, they can be used sequentially to determine whether the current time point is a signal or not. None of these methods detect multiple signals, which is an issue that generates another direction for future research.

5 Case studies

For each of our case studies, the goal is to examine the AERS data for evidence of reporting disproportionality in the given adverse event for patients taking the drug in question. In our first two case studies, the FDA issued a warning once the drug was on the market for a significant period of time[10]. Is is possible to detect a signal in the AERS data prior to the time of the FDA warning? Using the AERS data, how early could this potential link have been recognized? In our third case study, we examine an Adverse Event/Drug combination that has been documented in a recently published clinical trial, but has not resulted in FDA action. Can we find evidence in the AERS data to support the findings of this trial? Based on what we discover, can we make a recommendation to regulators about adding a warning?

5.1 Case Study 1: Avandia and Myocardial Infarction

The diabetes drug Avandia (Rosiglitazone) went on the market after FDA approval in 1999. The drug became popular; sales of the drug from GlaxoSmithKline peaked in 2006 at $3.2 billion in the United States that year [15]. In May 2007, the FDA issued a safety alert for the drug due to potential increased risk of heart attack. In 2010, the drug was suspended from the European market and the FDA severely restricted its use [15]. Pre-market clinical trials of Avandia showed no evidence of increased risk of heart attack; however, 8 years after the drug was approved, the FDA found enough evidence to lead to a warning contraindicating high risk patients and shortly thereafter, the drug was all but taken off the mass market.

There are two goals related to this case study. The first is to exemplify, via an analysis of the data for the first quarter of 2012, the consequences of stratification by age and gender. These results are presented in Section 6.1. The second goal is to analyze AERS data prior to the 2007 FDA warning for signal detection. We will apply our proposed trend analyses and the results are summarized in Section 6.2.1.

5.2 Case Study 2: Propecia and Sexual Dysfunction

Finasteride is a drug marketed as Propecia to treat male pattern baldness and Proscar to treat englarged prostate. Proscar went on the market in 1992 and Propecia in 1997. Pre-market clinical trials showed small but significant amount of sexual dysfunction [11]. The results of these trials were reported on the original label, however, in April 2011 and again in April 2012, the FDA revised the drug label to include new warnings that these drugs carry a potential risk of long-term sexual dysfunction. The goal in this case study is to look for evidence of disproportionality in reporting of sexual side effects in patients who reported taking Propecia prior to April 2011, and the results are discussed in Section 6.2.2.

[10]All specific drug label information below was retrieved from the Drugs-at-FDA website: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm. Leads to possible drug/AE combinations were found at the Pharmaceutical Drug Litigation Updates website: http://www.drug-injury.com/drug_injury.

5.3 Case Study 3: Thiazolidinediones and Macular Edema

Thiazolidinediones are a class of diabetes drugs that include Avandia (Rosiglitazone) and Actos (Pioglitazone). Macular Edema is an eye disease sometimes seen concurrently with diabetes; it is the leading cause of blindness in in diabetes patients. In July 2012, a retrospective cohort study of $103,368$ diabetic patients was published which found an increased risk of macular edema at 1-year and 10-year follow-up evaluations. Prior to this study, others that have investigated this link have found no causal evidence [10]. Post-market spontaneous reporting of Macular Edema is listed on the Avandia label; however, the FDA has not issued any warning regarding this particular drug/AE combination. The goal for this case study is to look for evidence of disproportionality in reporting of macular edema in patients who reported taking Avandia or Actos, and the results can be found in Section 6.2.3.

6 Results

6.1 Case Study 1: Avandia vs. Myocardial Infarction.

6.1.1 Description of datasets

Avandia has been brought under the scanner of the United States Food and Drug Administration (US FDA) in the context of adverse events related to Common Cardio Problem. FDA suspected that this drug yielded an unexpectedly large amount of heart related problems, thus issued a warning in the fourth quarter of 2007.

Is Avandia really that risky? By analyzing the retrospective data of the first quarter of 2012 that is publicly available in the US FDA Adverse Events database, it seems that Avandia deserves tremendous attention from physicians and drug manufacturers.

We begin by analyzing Avandia and each of the adverse events recorded in the database. There are 1136 Avandia-related adverse events. For each of the Avandia-related adverse event, we essentially compute its disproportionality, defined as observed counts/expected counts. Intuitively, if the observed count for a particular Avandia-adversed event pair is way higher than its expected count, it sends out alarms for investigation.

To serve this purpose, four methods are being used, namely PRR, ROR, BCPNN and EBGM. All of the four methods are based on analyzing the 2×2 contingency table(Table1) for each Avandia-adverse event pair.

After each pair is scored using those four methods, we rank the scores from highest to lowest. Results show that the four methods are in general consistent in the sense that four methods give similar ranking. Below are the top 10 Avandia-adverse related events rankings according to EBGM.

Drug	Event	EBGM	IC	ROR	PRR
AVANDIA	CORONARY ARTERY BYPASS	52.10	6.54	180.35	126.75
AVANDIA	CARDIOVASCULAR DISORDER	42.66	5.64	62.37	54.62
AVANDIA	STENT PLACEMENT	40.05	5.98	96.26	78.59
AVANDIA	MYOCARDIAL INFARCTION	30.10	4.97	37.37	34.62
AVANDIA	CORONARY ARTERY DISEASE	25.00	4.74	30.35	28.44
AVANDIA	CARDIAC FAILURE CONGESTIVE	20.89	4.44	24.60	23.37
AVANDIA	CEREBROVASCULAR ACCIDENT	17.44	4.18	20.12	19.28
AVANDIA	CARDIAC DISORDER	17.16	4.120	20.11	19.25
AVANDIA	CARDIAC PACEMAKER INSERTION	16.37	4.91	45.49	41.14
AVANDIA	IMPLANTABLE DEFIBRILLATOR INSERTION	15.65	5.11	64.02	55.67

Surprisingly, all of these adverse events are cardio related problems, indicating a strong warning that investigation be taken.
Moreover, we analyze Avandia-related adverse events by age and gender. We bracket age by classifying people with age 65 as old people, and age below 65 as young people (though most people in that group are mid-aged). This stratification is instructive because it may be that certain adverse events needs to be alerted for a specific

subgroup, while it is not a concern for another subgroup. In this study, we create subgroups of old people, young people, female and male. We also dig further by looking at old female, old male, young female and young male. All four scores are calculated and compared. We define it as our ungrouped dataset.
We also conduct another interesting study by subsetting all the drugs that were reported common cardio adverse events. This excludes irrelevant drugs in the study, and compares Avandia with all other drugs associated with common cardio adverse events. The group of common cardio adverse events is defined as the top 10 cardio related adverse events. By repeating the same stratification in the last paragraph, we create subgroups of old people, young people, female and male, old female, old male, young female and young male. All four scores are calculated and compared. And we define this as our grouped dataset.

The following tables and plots will give us more details and results.

6.1.2 Results of the analysis for the first quarter dataset of 2012 from AERS

- Combination of Avandia and cardiac disorder adverse event in the ungrouped dataset.

The four association measures - EBGM, IC, ROR, and PRR - of the combination of Avandia drug and cardiac disorder adverse event are computed using data from the first quarter of 2012. First of all, we deleted the pairs which have the counts that are less than 10. For the specific drug Avandia, there are 106 related adverse events. After stratifying by gender, there are only 39 related adverse events for female and 56 related adverse events for male. Only stratifying by age, there are 26 related adverse events left for old people while 40 for young people. Finally, we also stratified the dataset by gender and age at the same time, obtaining just 13 related adverse events for females under the age of 65, 12 for females over the age of 65, 13 for males over the age of 65, and 25 for males under the age of 65. We choose one cardiac related AEs (cardiac disorder) to show how the AE acts differently in different strata.

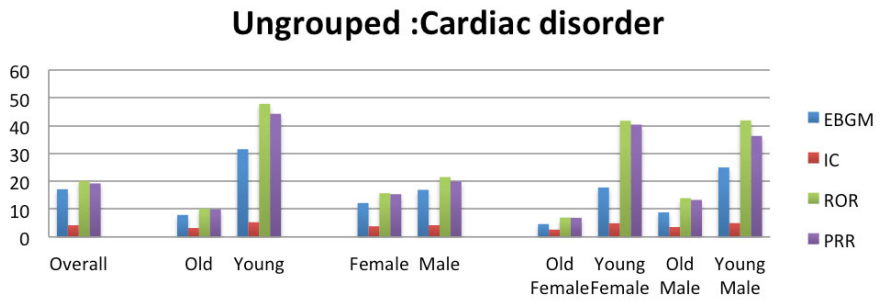

Figure 3: Avandia Use and Common cardiac disorder

Figure 3 shows that the scores calculated for people under 65 years of age are higher than those for people over 65 years old, and there is a slight difference between the male and female. The last group which is stratified by gender and age also shows that females under 65 years of age and males under 65 years of age have much larger scores than the females and males over 65 years old. The picture indicates that young people may be more prone to getting a cardiac disorder event than the older people when they take Avandia at the same time and same dose. This difference is not very clear between males and females. This stratification shows that probably an association with younger age groups is the main reason for the increase the total score as obtained from the non stratified database.

- Combination of Avandia and cardiac disorder adverse event in the grouped dataset.

For the result of the ungrouped dataset, we found out that 7 out of the top 10 high score Avandia-AE pairs are cardiac related, hence we are assuming that Avandia is more likely related to cardiac adverse events. It is reasonable for us to see how the score be changed in different strata. Therefore we chose the top 10

high Avandia related cardiac adverse events, and take them as a group named "COMMON CARDIO", and we keep the rest AEs as the same. Again, we deleted the pairs which have the counts that are less than 10. So in the grouped dataset, there are 98 adverse events in total. And stratified by gender, there are 23 related adverse events for female and 51 related adverse events for male. And stratified by age, there are 24 related adverse events for old while 25 for young. And we also stratified the dataset by gender and old, so there are 13 AEs for young female, 12 for old female, and 13 for old male and 23 related AEs for young male.

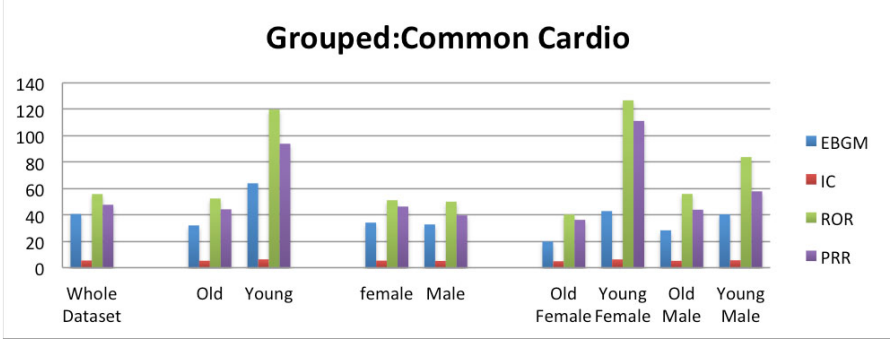

Figure 4: Avandia Use and Common cardiac disorder

For each group, we used four methods to get the score (EBGM IC ROR and PRR). We choose the "COMMON CARDIO" AE to see how the AE acts differently in different strata. Figure 4 shows that the young people have a higher score then the old people; and there is a slight difference between the gender groups; the last group which is stratified by gender and age also shows that young female and young male have much larger scores then the old female and old male. And RPR and ROR score of young female are much larger than the score for young male. The plot indicates that young people may more likely to have a cardiac disorder than the old people when they take Avandia. But the difference is not very obverse between male and female. So it probably the young people, especially young female, that increase the total score.

- Summary. From the two groups of dataset, the scores of the grouped AE dataset are higher than the scores of the ungrouped AE dataset. we get pretty much the same result, that is young people are more likely to have a cardio related adverse event compare to old people while there is little difference between male and female. Also, by comparison, the difference between EBGM, ROR, and PRR for the whole dataset group is less than the difference for the "young" group which suggests that examining the data by different demographic factors, such as age and gender, would lead to better results of detecting signals of drugs.
- Limitations and future directions: This study suffers from several limitations. First of all, we do not have the BMI for each patient. BMI may be an important factor to stratify. Second, the data is a collection of spontaneous response from physicians and patients, which may suffer from sampling bias. Third, it would be better if drugs are classified, for example, by biological component and chemical component. Studying how these two types of drugs relate adverse events would be an interesting and meaning topic. Lastly, it might be worthwhile to include all the drug-adverse event pairs in the ranking.

6.2 Time trend analysis

6.2.1 Case Study 1: Avandia and Myocardial Infarction

The quarterly AERS data allows us to analyze disproportionality measures longitudinally. Here we plot disproportionality measures for all time points preceding the issue of the FDA warning on Avandia regarding heart attacks.

The four methods produce very similar results. The only noticeable difference is that he EBGM does not spike as high in 2006 Q4, perhaps because this measure is less sensitive to small changes in the number of

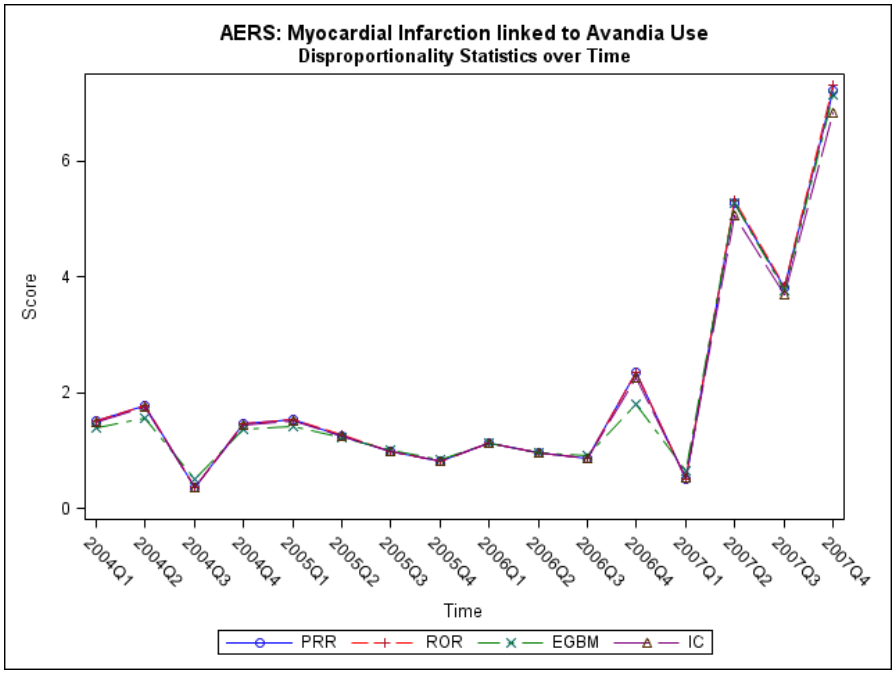

Figure 5: Myocardial Infarction linked to Avandia Use

drug/AE pairs reported (i.e. small changes in the number n_{00}).

The FDA warning was issued after the first quarter of 2007. With the benefit of hindsight we can interpret the spike in the fourth quarter of 2006 as a signal. Indeed this is consistent with what we see in the plot; before 2006 Q4, the disproportionality rates hover around 1 (no disproportionality) and then in 2006 Q4, they suddenly double.

For reasons discussed earlier, we pick the EBGM score to use in our longitudinal analysis. We employ our methods of longitudinal data analysis on EBGM score with the results in 2. The percent change and deviation from the mean methods successfully identify 2006 Q4 as the first time point to produce a signal; the non-parametric trend method does not identify a signal until 2007 Q2. Notice that all three methods produce results starting in 2005, that is, only after 4 initial time points. At least 4 time points are needed to calculate baseline statistics from which measures of change are determined.

6.2.2 Case Study 2: Propecia and Sexual Dysfunction

In the Finasteride/Sexual Dysfunction case study, the 4 disproportionality methods produce different results. The PRR and ROR scores overlap almost precisely, and they are both more sensitive to changes in number of reported drug/AE pairs, hence they produce dramatic spikes. The EBGM and IC methods follow the same trend as the frequentist methods, but produce smoother curves.

Again, we choose EBGM scores for further investigation. The plot of EBGM score over time shows a few spikes that may potentially be signals: 2008 Q3, 2010 Q1 and 2011 Q3 stand out to the naked eye. The FDA added sexual dysfunction to the Finasteride label after 2011 Q1 so we are most interested in detecting signals at 2008 Q3 and 2010 Q1.

The percent change method detected two signals prior to the one we expect to see at 2008 Q3 (see 3. The percent change is 169% at 2006 Q2 and 220% at 2006 Q4. This algorithm successfully detected 2008 Q3 (249%) and 2010 Q1 (144%), as well as the last three time points on the plot (283%, 190%, 179%).

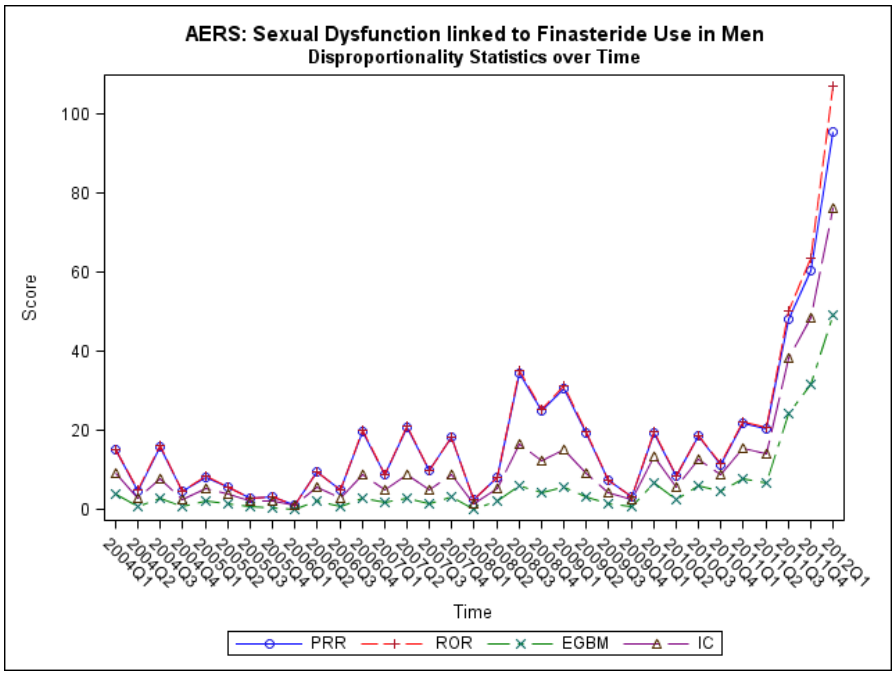

Figure 6: Sexual Dysfunction linked to Finsteride Use

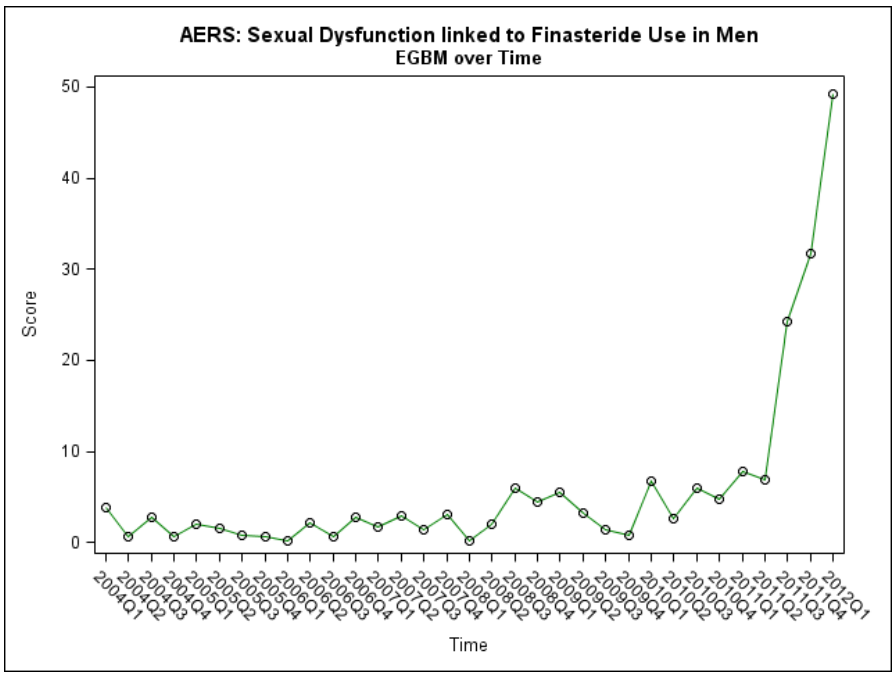

Figure 7: Sexual Dysfunction linked to Finasteride Use

Time	Method 1 (% change)	Method 2 (p-value)	Method 3 (p-value)
2005Q1	17%	0.66	0.98
2005Q2	0%	0.94	0.98
2005Q3	-11%	0.52	0.76
2005Q4	-32%	0.31	0.55
2006Q1	0%	0.91	0.52
2006Q2	-7%	0.56	0.42
2006Q3	-8%	0.46	0.23
2006Q4	**88%**	**0.02**	0.88
2007Q1	-47%	0.13	0.48
2007Q2	386%	<0.001	<**0.001**

Table 2: Avandia/Myocardial Infarction: Longitudinal signal detection results (EBGM)

The parametric method successfully detected 2008 Q3 and 2010 Q1 as well as the last three time points ($p<0.001$). It did not detect any other time points. Finally, the non-parametric trend method only detected a signal in 2009 Q1, two time points after the initial spike in 2008.

To summarize, in this case we see that, while the percent change method successfully detected the signals we expected, it was too sensitive to small changes in score and detected two signals erroneously. The non-parametric trend test lagged behind and only detected a signal two time points after one occurred and the parametric method successfully detected the signals we expected and no others.

6.2.3 Case Study 3: Thiazolidinediones and Macular Edema

In the Thiazolindinedione/Macular Edema case study, we see again that while the 4 disproportionality methods follow the same trends, they produce different results that are due to the PRR and ROR scores being more sensitive to small changes. Overall, we see more variability over time in this drug/AE pair, with scores suddenly spiking, even on the smoothest curve (EBGM). This drug/AE pair presents the most challenging scenario of signal detection of the three cases.

Recall that evidence of the link between Thiazolidinediones and Macular Edema was published very recently and the FDA has not made a recommendation nor added a warning regarding this link. Any signals detected in this data could potentially be worth investigation. To the naked eye, 2006 Q1, 2010 Q1 and 2010 Q3 stand out as potential signals. 2004 Q2 is also a spike, but will not be detected by our methods as it in the set of first 4 measures and must be used to calculate baseline statistics.

2006 Q1 is detected as a signal by the percent change method (678%) and the parametric method ($p<0.001$), but not by the non-parametric trend method. The non-parametric trend method picks up 2006 Q4 as a signal, again a few time points behind.

We note here that the non-parametric trend method, while not good at picking up single time point spikes, works well to detect signals where several time points in a row have an elevated score compared to the past, as in 2006 Q1-Q4 here.

2010 Q3 is detected by both the percent change method (235%) and the parametric method ($p<0.001$). 2009 Q4 is only detected by the percent change method (166%). For this time point, we cannot judge whether the sensitivities of percent change and deviation from the mean are too high or too low, since we can only speculate as to whether this time point represents a 'real' signal.

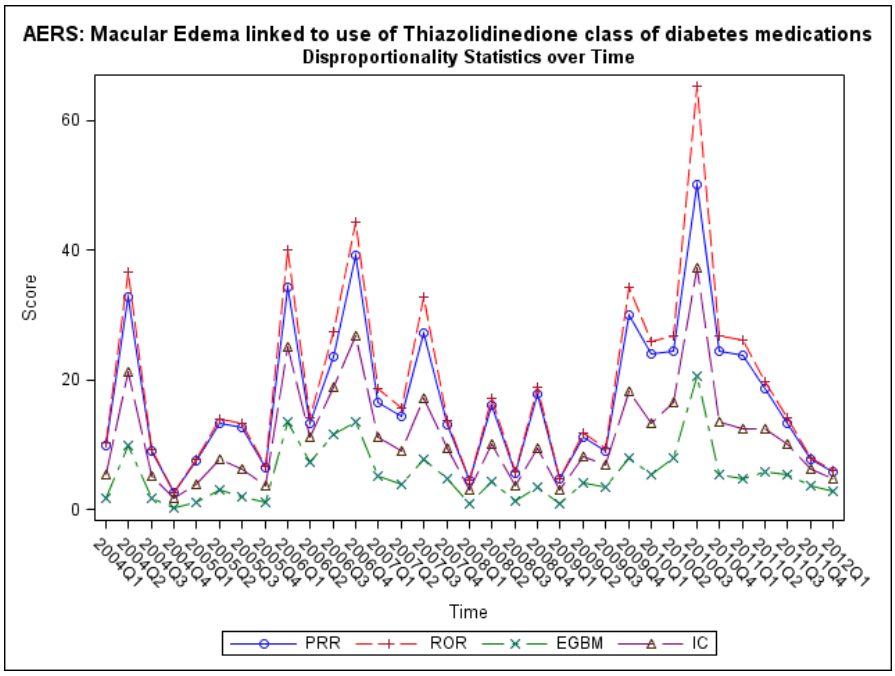

Figure 8: Macular Edema linked to Thiazolidinedione Use

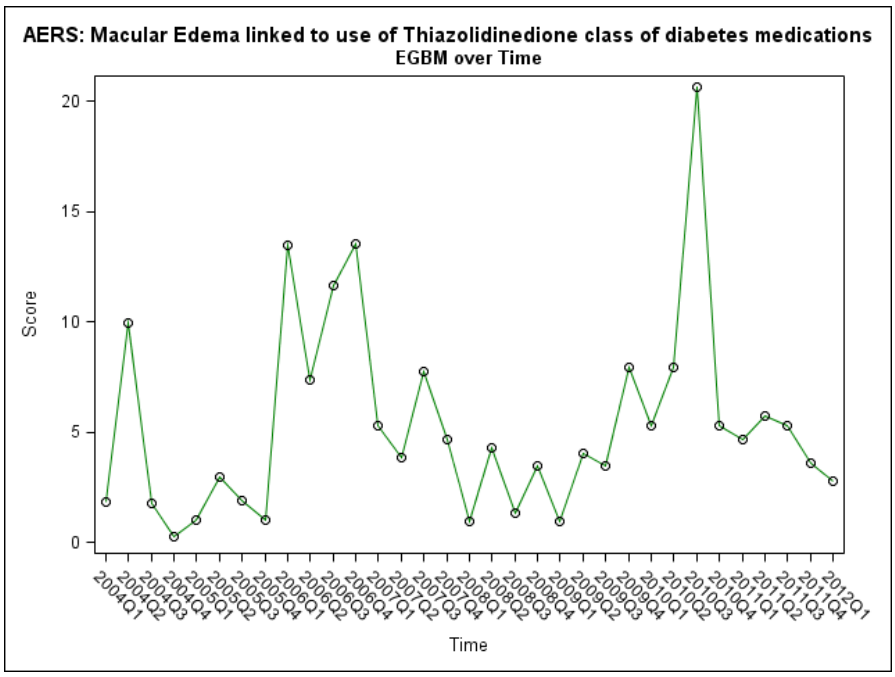

Figure 9: Macular Edema linked to Thiazolidinedione Use

Time	Method 1 (% change)	Method 2 (p-value)	Method 3 (p-value)
2005Q1	5%	0.95	0.46
2005Q2	6%	0.78	0.58
2005Q3	-53%	0.39	0.43
2005Q4	-56%	0.33	0.27
2006Q1	84%	0.25	0.14
2006Q2	**169%**	0.58	0.27
2006Q3	33%	0.45	0.24
2006Q4	**220%**	0.23	0.53
2007Q1	23%	0.85	0.62
2007Q2	55%	0.26	0.86
2007Q3	-27%	0.86	0.91
2007Q4	41%	0.17	0.69
2008Q1	-90%	0.18	0.88
2008Q2	6%	0.73	0.95
2008Q3	**249%**	**<0.001**	0.17
2008Q4	53%	0.09	0.11
2009Q1	75%	0.02	**0.02**
2009Q2	-29%	0.56	
2009Q3	-71%	0.61	
2009Q4	-79%	0.41	
2010Q1	**144%**	**<0.001**	
2010Q2	-15%	0.89	
2010Q3	106%	0.04	
2010Q4	19%	0.22	
2011Q1	56%	0.01	
2011Q2	29%	0.05	
2011Q3	**283%**	**<0.001**	
2011Q4	**190%**	**<0.001**	
2012Q1	**179%**	**<0.001**	

Table 3: Finasteride/Sexual Dysfunction: Longitudinal signal detection results (EBGM)

By carefully observing the time point at which a signal was detected, we find interestingly that the bootstrap approach does not detect a signal as fast as the parametric approach. This is mainly due to the nature of non-parametric bootstrap: by having fewer assumptions, we lose efficiency. On the other hand, this conservativeness gives more reliable results. The parametric approach, however, detects a signal quickly, but it suffers when the assumption may not be valid and is more likely to give false signals than the bootstrap approach.

6.3 Simulation

In this section, we show attempt numerical experiments of the approaches listed above in Section 3 using simulated data. Because the uncertainty and complexity of the real data described previously is very challenging, a more clear insight could potentially be obtained through simulations where the truth is known. This is especially important because there is no gold standard for signal detection techniques in the literature. In addition, little work has been conducted regarding simulation in the pharmacovigilance field. Three recent papers proposed simulation of data generation processes but with very different philosophies [13], [14], [3]. In our case, we address this issue in a different way with many interesting outcomes.

Time	Method 1 (% change)	Method 2 (p-value)	Method 3 (p-value)
2005Q1	-70%	0.58	0.66
2005Q2	-9%	1.00	0.75
2005Q3	28%	0.77	0.6
2005Q4	-34%	0.58	0.55
2006Q1	**678%**	**<0.001**	0.52
2006Q2	53%	0.44	0.42
2006Q3	96%	0.10	0.15
2006Q4	61%	0.07	**0.04**
2007Q1	-54%	0.96	
2007Q2	-59%	0.74	
2007Q3	10%	0.64	
2007Q4	-39%	0.85	
2008Q1	-83%	0.31	
2008Q2	-1%	0.83	
2008Q3	-70%	0.38	
2008Q4	24%	0.73	
2009Q1	-62%	0.35	
2009Q2	61%	0.87	
2009Q3	44%	0.77	
2009Q4	**166%**	0.42	
2010Q1	30%	0.90	
2010Q2	53%	0.43	
2010Q3	**235%**	**< 0.001**	
2010Q4	-49%	0.97	
2011Q1	-52%	0.86	
2011Q2	-41%	0.96	
2011Q3	-42%	0.97	
2011Q4	-32%	0.68	
2012Q1	-43%	0.56	

Table 4: Thiazolidinedione/Macular Edema: Longitudinal signal detection results (EBGM)

6.3.1 Data generation

Unlike previous work by Ahmed, et al. [3], which suggests generating data in terms of number of events, our starting point is to generate data in terms of the patient's reporting mechanism.

We assume the distribution of one adverse event $AE_j,\ j \in \{1, \cdots, J\}$ from a patient's report follows a Bernoulli distribution with success probability p_{AE_j}. The probability p_{AE_j} in the Bernoulli distribution for patient N is determined by the following equation:

$$p_{AE_j}(N) = Prob(AE_j|patient\ N) = \frac{1}{e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_I x_I)}+1} \tag{1}$$

where $x_i \in \{0, 1\},\ i = 1, \ldots, I$ is an indicator for whether the patient is using drug i. Let $\beta_i, i \in \{1, \cdots, I\}$ denote the coefficients of the drug effect to a particular AE. The larger value of β_i, the higher the effect drug i has on AE_j. Let β_0 denote the constant, which could be viewed as the background noise in the simulation.

Instead of directly generating counts for each drug-event combination, we first generate the patient reports, each with a number of drugs and AEs. The counts of each drug-event combination will then be calculated

from the patient's report.

For all simulations, we consider the number of drugs to be $J = 4$, and $I = 4$ the number of AEs. We will generate information for $10,000$ patients. Our simulation study consists of the following scenarios:

CASE 1. The reporting of adverse events are completely random, and independent with any possible factors. Without loss of generality, we assume $AE_j,\ j \in \{1, \cdots, J\}$ has 50% probability to appear in a patient's report.

CASE 2. The chance of appearance of $AE_j, j \in \{1, \cdots, J\}$ depends on the Bernoulli distribution described in (1). In CASE 2, we fix $\beta_0 = -2.2$, and the other coefficients are given in Table 5. In this setup, AE_1 is mainly dominated by $drug_1$, AE_2 is mainly dominated by $drug_2$, AE_3 is mainly dominated by $drug_1, drug_2$ and $drug_3$, and AE_4 is dominated by $drug_1, drug_2, drug_3$ and $drug_4$.

	β_1	β_2	β_3	β_4
AE_1	2.00	0.00	0.00	0.00
AE_2	0.00	2.00	0.00	0.00
AE_3	1.50	1.50	2.00	0.00
AE_4	1.50	1.50	1.50	2.00

Table 5: Coeffients in simulation CASE II.

CASE 3. In this case, we increase the first cell coefficient in the above table, and others remain the same (Table 6). In this case, the influence of AE_1 by $drug_1$ increases from 37.75% to 78.58%. Compared to CASE

	β_1	β_2	β_3	β_4
AE_1	3.80	0.00	0.00	0.00
AE_2	0.00	2.00	0.00	0.00
AE_3	1.50	1.50	2.00	0.00
AE_4	1.50	1.50	1.50	2.00

Table 6: Coeffients in simulation CASE III.

II, in CASE III the probability of the drug-event combination $drug_1 - AE_1$ is increased.

CASE 4. In this case, we increase every cell from Table 6, which means the counts of every drug-event combination increases significantly (Table 7). In CASE IV, the probabilities of all counts have a very large

	β_1	β_2	β_3	β_4
AE_1	3.80	0.00	0.00	0.00
AE_2	0.00	3.80	0.00	0.00
AE_3	2.50	2.50	3.80	0.00
AE_4	2.50	2.50	2.50	3.80

Table 7: Coeffients in simulation CASE IV.

jump.

For each individual case from the setup above, we generated 50 datasets. Then for each dataset we used the four disproportionality methods described in Section 3 to obtain the PRR, ROR, EBGM, and BCPNN scores for each drug-AE combination, and finally we calculated the mean and variance across the 50 datasets (to account for simulation variability).

6.3.2 Simulation results

Tables 8, 9, 10 and 11 in the Appendix show the results for CASE I, II, III, and IV, respectively. Every drug-event combination is listed in the results table for the four drugs and four AEs under consideration. Columns N00 through N11 are the corresponding cells in the 2×2 contingency table pictured in Table 1. The columns EBGM, IC, PRR, and ROR correspond to the measures of disproportionality calculated for each drug-adverse event combination through each of those four methods, and are shown here together with their standard deviation.

In Table 8, since the assumption for the simulation was that adverse events and drugs are reported completely at random and independently, the values of EBGM all equal one for every drug-event combination. Similar results occur for the other three methodologies: IC, PRR, and ROR. There is no sign of disproportionality in our simulation CASE I, as expected.

In Table 9, the probability of $drug1 - AE1$ is around 30%. This is reflected from N00 in the first row. However, the EBGM is still very close to 1, and PRR and ROR do not change significantly. In Table 10, the EBGM, PRR and ROR are still close to 1, which indicates that all existing algorithms have difficulty detecting the signal change, even when the probability changes dramatically. In Table 11, we obtained similar results for EBGM, PRR and ROR, although they have a slightly higher volitality than the results before. Although the counts in N00 have increased significantly compared to Table 9, the signal detection approaches have failed to find these changes.

Through this simulation, we found some potential issues to look into regarding those existing approaches.

- None of these approaches take drug-drug interactions into consideration. The only difference between CASE II and CASE III is that the probabilty of AE_1 caused by $drug1$ changes from 37.75% to 78.58%. In Table 10 and 11, the N00 value for drug1-AE1 changes from 1378.54 to 2273.06, which is an expected increase since we increase the probability of occurrence. However, the N10 value for drug1-AE1 also changes from 4072.8 to 7051.20, which is a big change and dilutes the signal. The reason behind this phenomenon is that some patients take other drugs besides drug1; however, once they have an AE1 event, this event will also be counted into the "other drugs" effect even if the real reason is only drug1.

- The pool of all drugs and all AEs will affect the results greatly. In our simulation study, we only considered 4 drugs and 4 AEs. In this case, we have the same scale of values from N00 to N11. However, in the AERS data, the total number of drugs and AEs are extremely large, and the consequence of this is that the N00 number can be very small compared to N10, N01, and N11. For example, N00 could be just 100, whereas N10, N01, and N11 are all in the millions. This case actually raises a very important question, which is how to choose the pool of drugs and AEs against which to make the comparisons. Is it more reasonable to try to include as many drugs and AEs as possible, or should we only select those drugs and AEs that are known to be related (say because of their chemical content or because of the known classification of events by system-organ class)?

- The probability of AEs appearing as a consequence of taking a specific drug in our simulation ranges from 10% to 90%. However, in the AERS dataset the probability of AEs appearing is very low, say 1%. All the existing approaches we studied fail to detect the disproportion in our simulation when the probability is high. We should direct more attention on how to assess the performance of those approaches and how they compare to each other. The high chance of certain events appearing should be reflected by those methods even in special situations, such as those fabricated in our simulation.

6.3.3 Conclusions from simulation

The main purpose of our simulation was to try to compare and contrast the results that are provided by each of the different methods. Since all the methods have different approaches and modeling techniques, we were hoping that the simulation would shed some light into what those differences are, and maybe some of the advantages and limitations of each. This could provide guidelines on what could be improved or modified to get more significant results.

A comparison of all the disproportionality methods currently in use can be a valuable exercise, given some sort of gold standard.We propose that in order to validate any results that may be obtained using data mining techniques, a reference database with known drug-adverse event combinations is needed. The information needed to construct this database could be obtained from studies published in the literature and from information obtained through clinical trials. To the best of our knowledge, only a few examples exist in the literature of databases constructed for this purpose (e.g. see [9]), and we strongly feel that this approach should have a more widespread use.

7 Conclusions

In this project, we investigated the four most popular signal detection methods in the current literature: PRR, ROR, MGPS, and BCPNN, and highlighted some of their advantages and limitations, as well as points for improvement.

We also developed three novel algorithms for signal detection that incorporate the time factor into the analysis, allowing for a sequential determination of "importance of association" between specific drug-adverse event pairs: one based in percent changes, a parametric approach, and a non-parametric approach. We conducted analyses for three drug-adverse event combinations, namely Avandia & myocardial infarction, Finasteride & sexual dysfunction, and Thiazolidinedione & macular edema. We found that the parametric approach is the fastest to identify a potential signal, but it might be likely to produce more false positives. On the other hand, the nonparametric approach seems to be more conservative (more evidence needs to accumulate in order for it to detect a signal) which may make it more reliable, but it is slower than the parametric approach.

Additionally, we analyzed the Avandia & myocardial infarction pair for a specific period of time, highlighting the importance of stratification on the demographic characteristics of the individuals reporting adverse events. Finally, we did a simulation study to gain insight into the existing methods and to try to provide a starting point for future studies, since we believe that validation of the performance of any data mining algorithm is essential.

8 Future work and recommendations

There are many prospective lines of future work related to the signal detection problem. Some of them we have already outlined in the paper, but here we make some additional recommendations.

The reliability of the dataset is a very important issue to keep in mind. To this respect, we believe that much work can be put into homogenizing the names of the drugs and adverse events, and this can potentially be addressed via language processing techniques. In addition, due to the importance of conducting stratified analyses, more demographic covariates should be collected, as well as information related to the drug doses and exposure times, which could be very valuable in dismissing spurious signals. We also propose including other sources of information, such as results of clinical trials, epidemiological tracking information from the CDC, or drug labels, in order to introduce some level of validation to the voluntary reporting system.

Other directions for future work lie in how to model interactions: both drug-drug interactions, and adverse events with others. Some efforts have been made to introduce drug-drug interactions into existing methods

by expanding the definition of "drug" to drug combinations [5]. There have also been attempts to use logistic regression to discover associations between drugs [12]. But these seem to suffer from the issue of high dimensionality, so maybe clustering methods could be explored to this respect to reduce the dimensionality of the data. In terms of the event associations, a possibility would be to include system organ class (SOC) information into a hierarchical model as prior information.

A promising area of exploration lies in how to incorporate historical information into the modeling, since a lot is learned in each time period and could be potentially used to estimate the background noise for future time points.

Two final thoughts include taking into consideration the way in which the reports are submitted. First, it could be promising to try to model the reporting mechanism and include it as prior information in a Bayesian model, since in this way we take into account the uncertainty regarding the number of exposed individuals. Secondly, weights could be added depending on the reliability of the agent submitting the report (clinicians, patients, manufacturers), since we could have more confidence on the information provided by different sources.

In our opinion, there are several issues that remain problematic. One of them is the absence of a gold standard against which to evaluate the performance of data mining and signal detection methods. Another point for concern is the lack of validation and comparisons of the different methods.

Since there are no best practices or golden standard for signal detection, a great area of opportunity arises for the development of analytical tools, but also for the misinterpretation of their results. It is important to always keep in mind that no algorithm can replace the role of trained physicians, since signal detection requires clinical judgement and knowledge of thresholds, but the methods can serve as initial indicators of the possibility of associations between drugs and adverse events.

9 Appendix

Drug	Event	N00	N10	N01	N11	EBGM	std	IC	std	PRR	std	ROR	std
drug1	AE1	2512.92	7505.46	7560.38	22496.18	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug1	AE2	2503.66	7507.38	7522.04	22541.86	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug1	AE3	2504.28	7509.60	7521.42	22539.64	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug1	AE4	2506.58	7525.06	7519.12	22524.18	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE1	2500.60	7517.78	7471.34	22585.22	1.00	0.01	0.00	0.01	1.00	0.01	1.01	0.02
drug2	AE2	2488.56	7522.48	7483.38	22580.52	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE3	2487.06	7526.82	7484.88	22576.18	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE4	2495.72	7535.92	7476.22	22567.08	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE1	2516.66	7494.38	7556.64	22507.26	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE2	2521.66	7492.22	7551.64	22509.42	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE3	2522.06	7509.58	7551.24	22492.06	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE4	2493.68	7524.70	7510.32	22546.24	1.00	0.01	-0.00	0.01	1.00	0.01	0.99	0.01
drug4	AE1	2502.16	7508.88	7501.84	22562.06	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug4	AE2	2500.88	7513.00	7503.12	22557.94	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug4	AE3	2507.28	7524.36	7496.72	22546.58	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug4	AE4	2511.18	7507.20	7514.52	22542.04	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01

Table 8: Simulation results for CASE I.

Drug	Event	N00	N10	N01	N11	EBGM	std	IC	std	PRR	std	ROR	std
drug1	AE1	1378.54	4072.80	7474.06	22023.86	1.00	0.01	-0.00	0.02	1.00	0.01	1.00	0.02
drug1	AE2	1357.76	4088.02	7349.08	22154.40	1.00	0.01	0.00	0.02	1.00	0.01	1.00	0.02
drug1	AE3	2688.20	8148.98	6018.64	18093.44	0.99	0.01	-0.01	0.01	0.99	0.01	0.99	0.01
drug1	AE4	3299.66	9915.30	5407.18	16327.12	1.00	0.00	0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE1	1352.06	4099.28	7331.90	22166.02	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.02
drug2	AE2	1344.94	4100.84	7339.02	22164.46	0.99	0.01	-0.01	0.02	0.99	0.01	0.99	0.02
drug2	AE3	2698.84	8138.34	5985.12	18126.96	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE4	3288.12	9926.84	5395.84	16338.46	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE1	1381.24	4064.54	7471.36	22032.12	1.00	0.01	0.00	0.02	1.00	0.01	1.00	0.02
drug3	AE2	2744.24	8092.94	6108.36	18003.72	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE3	3348.58	9866.38	5504.02	16230.28	1.00	0.00	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE4	1359.52	4091.82	7346.34	22151.58	1.00	0.01	0.00	0.02	1.00	0.01	1.00	0.02
drug4	AE1	1361.84	4083.94	7344.02	22159.46	1.00	0.01	0.01	0.02	1.00	0.01	1.01	0.02
drug4	AE2	2705.90	8131.28	5999.96	18112.12	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug4	AE3	3278.60	9936.36	5427.26	16307.04	1.00	0.00	-0.01	0.01	0.99	0.01	0.99	0.01
drug4	AE4	1361.22	4090.12	7345.62	22152.30	1.00	0.01	0.00	0.02	1.00	0.01	1.00	0.02

Table 9: Simulation results for CASE II.

Drug	Event	N00	N10	N01	N11	EBGM	std	IC	std	PRR	std	ROR	std
drug1	AE1	2273.06	7051.20	7292.50	22568.32	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug1	AE2	1379.28	4105.66	8489.42	25210.72	1.00	0.01	-0.00	0.02	1.00	0.02	1.00	0.02
drug1	AE3	2765.98	8230.90	7102.72	21085.48	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug1	AE4	3359.40	10019.60	6509.30	19296.78	1.00	0.00	-0.00	0.01	1.00	0.01	0.99	0.01
drug2	AE1	2373.54	6950.72	7566.10	22294.72	1.00	0.01	0.01	0.01	1.00	0.01	1.01	0.01
drug2	AE2	1398.18	4086.76	8541.46	25158.68	1.00	0.01	0.01	0.02	1.01	0.01	1.01	0.02
drug2	AE3	2784.62	8212.26	7155.02	21033.18	1.00	0.00	-0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE4	3383.30	9995.70	6556.34	19249.74	1.00	0.00	-0.00	0.01	1.00	0.01	0.99	0.01
drug3	AE1	1332.50	4152.44	8233.06	25467.08	0.99	0.01	-0.01	0.02	0.99	0.01	0.99	0.02
drug3	AE2	2690.52	8306.36	6875.04	21313.16	1.00	0.01	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE3	3269.48	10109.52	6296.08	19510.00	1.00	0.00	0.00	0.01	1.00	0.01	1.00	0.01
drug3	AE4	2313.62	7010.64	7497.56	22363.26	0.99	0.01	-0.01	0.01	0.99	0.01	0.98	0.01
drug4	AE1	1374.98	4109.96	8436.20	25263.94	1.00	0.01	0.00	0.02	1.00	0.01	1.00	0.02
drug4	AE2	2755.76	8241.12	7055.42	21132.78	1.00	0.00	0.00	0.01	1.00	0.01	1.00	0.01
drug4	AE3	3366.82	10012.18	6444.36	19361.72	1.00	0.00	0.01	0.01	1.01	0.01	1.01	0.01
drug4	AE4	2364.04	6960.22	7504.66	22356.16	1.01	0.01	0.01	0.01	1.01	0.01	1.01	0.01

Table 10: Simulation results for CASE III.

References

[1] Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation. Emerging Safety Science: Workshop Summary. Washington (DC): National Academies Press (US); 2008. 8, Pharmacovigilance. Available from: http://www.ncbi.nlm.nih.gov/books/NBK4056/

[2] FDA, The Sentinel Initiative: National Strategy for Monitoring Medical Product Safety.

[3] Ahmed, I., et al. (2009). *Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting*. Statistics in Medicine **28**, 17741792.

[4] Bate A. (2007). *Bayesian confidence propagation neural network*. Drug Safety **30**, 623625.

Drug	Event	N00	N10	N01	N11	EBGM	std	IC	std	PRR	std	ROR	std
drug1	AE1	2258.06	6999.36	10031.94	30693.88	0.99	0.01	-0.01	0.01	0.99	0.01	0.99	0.01
drug1	AE2	2261.30	6972.52	10076.18	30673.24	0.99	0.01	-0.01	0.01	0.99	0.01	0.99	0.01
drug1	AE3	3650.80	11154.56	8686.68	26491.20	1.00	0.00	-0.00	0.01	1.00	0.01	1.00	0.01
drug1	AE4	4131.48	12555.16	8206.00	25090.60	1.00	0.00	0.00	0.00	1.00	0.00	1.01	0.01
drug2	AE1	2327.12	6930.30	10274.32	30451.50	1.00	0.01	-0.00	0.01	1.00	0.01	1.00	0.01
drug2	AE2	2341.78	6892.04	10259.66	30489.76	1.00	0.01	0.01	0.01	1.01	0.01	1.01	0.01
drug2	AE3	3738.36	11067.00	8863.08	26314.80	1.00	0.00	0.00	0.01	1.00	0.00	1.00	0.01
drug2	AE4	4194.18	12492.46	8407.26	24889.34	1.00	0.00	-0.00	0.00	1.00	0.00	0.99	0.01
drug3	AE1	2284.50	6949.32	10005.50	30743.92	1.00	0.01	0.01	0.01	1.01	0.01	1.01	0.01
drug3	AE2	3627.80	11177.56	8662.20	26515.68	1.00	0.00	-0.00	0.00	1.00	0.00	0.99	0.01
drug3	AE3	4119.64	12567.00	8170.36	25126.24	1.00	0.00	0.01	0.00	1.01	0.00	1.01	0.01
drug3	AE4	2378.34	6879.08	10375.98	30349.84	1.01	0.01	0.01	0.01	1.01	0.01	1.01	0.01
drug4	AE1	2346.24	6887.58	10408.08	30341.34	0.99	0.01	-0.01	0.01	0.99	0.01	0.99	0.01
drug4	AE2	3788.40	11016.96	8965.92	26211.96	1.00	0.00	0.00	0.00	1.00	0.00	1.01	0.01
drug4	AE3	4241.34	12445.30	8512.98	24783.62	1.00	0.00	-0.01	0.00	0.99	0.00	0.99	0.00
drug4	AE4	2293.90	6963.52	10043.58	30682.24	1.00	0.01	0.01	0.01	1.00	0.01	1.01	0.01

Table 11: Simulation results for CASE IV.

[5] DuMouchel W. (1999). *Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System*. Am. Stat. **53**, 17790.

[6] Gould, A.L. (2003). *Practical pharmacovigilance analysis strategies*. Pharmacoepidemiology and Drug Safety **12**, 559-574.

[7] Hauben, M. (2003). *A Brief Primer on Automated Signal Detection*. The Annals of Pharmacotherapy **37(7/8)**, 1117-1123.

[8] Hauben, M., Zhou, X. (2003). *Quantitative Methods in Pharmacovigilance*. Drug Safety **26-3**, 159-186.

[9] Hochberg, A.M., et al. (2009). *An Evaluation of Three Signal-Detection Algorithms Using a Highly Inclusive Reference Event Database*. Drug Safety **32 (6)**, 509-525.

[10] Idris, I., et al. (2012). *Association Between Thiazolidinedione Treatment and Risk of Macular Edema Among Patients With Type 2 Diabetes*. Arch Intern Med. **172 (13)**, 1005-1011.

[11] Irwig, M. S. (2012), *Persistent Sexual Side Effects of Finasteride: Could They Be Permanent?* Journal of Sexual Medicine **9 (7)**.

[12] Madigan, D., et al. (2010). *Bayesian methods in pharmacovigilance*. In: J. M. Bernardo, et al. (eds), Bayesian Statistics **9**, Oxford University Press.

[13] Rolka H., et al. (2005). *Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data*. Statistics in Medicine **24(4)**, 551562.

[14] Roux E., et al. (2005). *Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance*. IEEE Transactions on Information Technology in Biomedicine **9(4)**, 518527.

[15] *Avandia (Drug)*, New York Times 'Health' Section article. From http://topics.nytimes.com/top/news/health/diseasesconditionsandhealthtopics/avandiadrug/index.html.