A Novel Approach for Detecting Rare Drug-Drug Interaction using Data Mining

Bazila Banu, Appavu alias Balamurugan and Shyni Beaulah.M

ABSTRACT

Discovering unknown Drug-Drug interaction (DDI) as early as possible is highly advantageous. Substantial effort has been taken by Food and Drug Adverse Event Reporting System (FAERS) to identify unknown Drug-Drug interaction. Interactions can direct to safety measures in prescribing, absolute contraindications for combination use, or even drug withdrawal. In particular, understanding drug interactions between commonly prescribed drugs is of great clinical importance. Drug safety depends greatly on postmarketing surveillance - the systematic detection and evaluation of medicines once they have been marketed. Current postmarketing methods mostly rely on Spontaneous Reporting System (SRS). Currently, there is a lack of literature describing DDI. Multi-Item Gamma Poisson Shrinker (MGPS) is a well known statistical algorithm used to detect frequent combination of multiple drugs and adverse events. The limitation of MGPS is its lack of ability to detect links between rare adverse events. Traditional signal detection algorithms are capable of exposing frequent drugs which are insufficient since people consuming multiple drugs are common nowadays. Hence, a novel algorithm has been proposed for the identification of known and unknown DDIs from adverse event reports in SRS. The challenge of discovering drug interactions is exacerbated because large clinical trials regularly focus on establishing the effects of single drugs. Changeable combinational effects can be identified only through postmarketing surveillance and signal detection. The proposed Prior-Weighted algorithm can be utilized by Pharmacovigilance cell (PhV), pharmaceutical industry and the physicians to gain knowledge about rare drug interactions.

INTRODUCTION

Medicines are designed to cure or prevent diseases but there is also risk in taking any medicine which may result in short term or long term Adverse Drug Reactions (ADR) which can cause serious harm to patients (Mei Liu et al., 2012) (Elisabetta et al., 2012). ADR monitoring is essential for each drug throughout its development phase including drug design. Different phases of clinical trials and postmarketing surveillance plays a significant role in drug safety. Nowadays, analysis about Drug-Drug interactions (DDIs) are required (Marc et al., 2012) (Elisabetta et al., 2012). Discovery of unknown interaction as early as possible is highly advantageous. The interactions between drugs are difficult to study, and there are few predictive methods for discovering novel DDIs. Unpredictable adverse events due to DDIs, can be identified only through postmarketing surveillance. Even though drugs are subjected to clinical trials they are monitored through postmarketing surveillance with large number of patients with different medical conditions. Pharmacovigilance (PhV), also known as drug safety surveillance, defined by World Health Organization (WHO) as the process of detection and evaluation of drug-related problems (Elisabetta et al., 2012). It is broadly classified into two categories premarketing surveillance and postmarketing surveillance. Pre-marketing surveillance deals with information regarding ADR collected from pre-clinical screening and clinical trials (Heba et al., 2014). Postmarketing surveillance deals with the post approval of drug life. The major aim of PhV is the timely detection of new ADR and to monitor the severity of ADR.

Spontaneous Reporting Systems (SRS) serves as the foundation for data-collection system and it is used by post-marketing surveillance since 1960. Spontaneous reports are gathered from regional, national ,international level through different databases (Heba et al., 2014) (DuMouchel & Darly, 2013). The Uppsala Monitoring
Centre in Sweden is responsible for the worldwide gathering of all serious ADR received by regulatory authorities and companies (http://www.who-umc.org). The Food and Drug Adverse Event Reporting System (FAERS) collects ADR from the US as well as rare and severe events from other countries. Some commonly used SRS are the Adverse Event Reporting System (AERS) maintained by the US FAERS and the VigiBase managed by the World Health Organization (WHO) (http://www.who-umc.org). Information in spontaneous reporting system usually include the drugs suspected to cause the ADR, concomitant drugs, indications, suspected events, and limited demographic information. Many post-marketing surveillance analysis relay on the reports submitted to the SRS, which uses disproportionality analysis and data mining algorithms (Mei Liu et al., 2012).

Computer-assisted procedure processes the SRS dataset by techniques like pruning, classification, association and clustering known to be Data Mining Algorithms (DMAs). Most of the DMAs are used in PhV is for quantitative signal detection (Balakin et al., 2009), (Hauben et al., 2005). The purpose of quantitative signal detection are many-fold and may vary depending on the local habit of PhV experts (Stephenson & Hauben M., 2007). For instance, DMAs can be used as an aid to the traditional case-by-case assessment; as a screening tool to periodically generate a list of signals requiring in depth investigation (i.e., to prioritize signals) (Wilson et al., 2004). Complex data dependencies are difficult to detect by manual process (e.g., drug-drug interactions or drug-related syndromes) (Bate et al., 2002). Also, the accuracy of data mining techniques need to be tested to determine the known safety issues (Evans S. J et al., 2001). However, new drugs and its interactions should be monitored on regular basis to provide a complete health care system to the society. (Balakin et al., 2009) (Hauben et al., 2005).

**Literature Review:**

Several methods are proposed to estimate the disproportionality measure and its confidence interval. DPA includes Proportional Reporting Ratio (PRR) (Evans S. J et al., 2001), (Evans S. J., 2000), Reporting Odds Ratio (ROR) (Ma et al., 2003), Information Component (IC) (Bates et al., 1997), (Bates et al., 2002) and Multi-item Gamma Poisson Shrinker (GPS or MGPS) (DuMouchel, 1999) are widely used, and currently employed by the Medicines and Healthcare products Regulatory Agency (MHRA), UK, the Netherlands Pharmacovigilance Centre Lareb, the World Health Organization (WHO), and the US FAERS, respectively. There are some more complex algorithms based on Bayesian statistics were developed such as the gamma-Poisson shrinker (GPS) (Ahmed et al., 2009), the multi-item gamma- Poisson shrinker (MGPS) (DuMouchel, 1999), (Almenoff et al., 2005) and empirical Bayesian geometric means (EBGMs) (DuMouchel et al., 2004), (Gould, 2007). The MGPS method is in use by FAERS and Bayesian Confidenc Propagation Neural Network (BCPNN) (Bate et al., 1998) (Andrew et al., 2002) is based on Bayesian logic where the relation between the prior and posterior probability was expressed as the “information component (IC)”. The IC given by the BCPNN is applied by the World Uppsala Monitoring Centre (UMC) to monitor safety signals in their SRS (http://www.who-umc.org).

The most widely used analytical methods for signal detection in SRS is Disproportionality Analysis (DPA) (Heba et al., 2014). It requires comparison of observed to expected proportions of Drug-Adverse Event Combinations (DEC). DPA methodologies are generally classified into two categories: frequentist and Bayesian. Both the methods are based on 2 x 2 contingency table to derive the measure of each drug-event pair in SRS. A safety signal is generated when there is a discrepancy between the observed number of drug event combination within the database and the expected number of cases.

**a. Frequentist Approach:**

Frequentist approach includes DMAs like ROR and PRR. These approaches solely depend on the information contained in 2 x 2 contingency table. ROR was first established in Netherlands Pharmacovigilance Foundation Lareb. It deals with the estimation of incidence rate ratio calculating odd of AE occurring in those not exposed to that drug. The ROR with 95% confidence interval (CI) is computed through the following formulae (Heba et al., 2014).

\[
\text{ROR} = \frac{A}{B} \times \frac{C}{D} = \frac{AD}{BC}
\]

\[
95\% \ CI = e^{\ln(\text{ROR}) \pm 1.96 \times \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}}
\]

Proportional Reporting Ratio (PRR) measures the strength of association between the suspected ADR and suspected drugs higher the value of PRR stronger the strength of the signal. The computation of PRR is based on 2 x 2 contingency table, and calculated using the formula (Heba et al., 2014).

\[
\text{PRR} = \frac{A(1-C)}{B(1-D)} = \frac{A(1-C)}{B(1-D)}
\]

\[
95\% \ CI = e^{\ln(\text{PRR}) \pm 1.96 \times \sqrt{\frac{1}{A} + \frac{1}{A+C} + \frac{1}{B} + \frac{1}{B+D}}}
\]

The limitation of these methods are the results tend to become unstable when number of events is small resulting in large estimates with wide confidence and they generate many false positive signals.
Fig. 1: 2X2 contingency table.

b. Bayesian Approach:

The Bayesian approaches include Information Component (IC) and Multi–Item Gamma Poisson Shrinker (MGPS). The role of Uppsala Monitoring Centre is to detect signals in spontaneous reporting system for suspected Adverse Drug Reaction. A method called Bayesian confidence propagation neural network (BCPNN) implements Bayesian statistics within a neural network architecture to search for previously unknown higher order dependencies in the data set (Andrew et al., 2002). The strength between the drug and ADR is determined by logarithmic measure using IC

$$IC = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Where p represents the probability, x denotes drug and y denotes ADR. The currently used data mining algorithm by FAERS is Multi-Item Gamma Poisson Shrinker (MGPS). This algorithm is used for detecting unexpected frequent combination of drug events (DuMouchel, 1999). It measures the Empirical Bayes Geometric Mean (EBGM) based on the Relative Reporting Ratio to enhance the accuracy.

Identify combinations with

$$a > \frac{(a + b)(a + c)}{(a + b + c + d)}$$

Posterior density for RRR can be estimated by Baye’s theorem. It takes only the frequent combinations and does not report about rare events (Almenoff et al., 2005), (DuMouchel & Daryl, 2013).

The above mentioned DPA methods are successful in detecting single Drug-ADR association, but it is essential to identify multi-item ADR association as it may suggest drug interactions. Drug interaction are extremely important (Wilson et al., 2004). A typical SRS database may contain thousands of drug and ADR, so it is unfound to specify all combinations for statistical analysis. Drug interactions may also increase the risk of ADR. Statistical analysis works well with the identification of single drug-and-ADR signals, but not applicable for drug interaction identification. Alternatively, data mining algorithms such as a priori algorithm and clustering algorithms are applicable and useful. It provides an excellent opportunity for computer scientists to develop new algorithms for drug interaction detection. The main objective of this paper is to demonstrate that multi-item ADEs exist in the AERS database and can be discovered by our method. Second, the drug interactions which causes rare ADR are effectively identified. To detect drug interaction with multi-item association rule mining is used (Ma L et al., 2003).

Proposed Approach:

Figure 2 gives the overall processing of the proposed approach.

Data Source:

FAERS is a relational database prepared in accordance with the international safety reporting guidance (ICH E2B) issued by the International Conference on Harmonization. FAERS includes different files provided in the form of text and ASCII for each quarterly period and they are easily downloaded from FAERS site. Indeed, these files can be imported into all popular applications for relational database such as ORACLE®, Microsoft Office Access, MySQL® and IBM DB2®.

Each classification of the entities present in the text file are provided in Table 1:

<table>
<thead>
<tr>
<th>Number of reports</th>
<th>Target reaction</th>
<th>Other reaction</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target drug</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Other drug</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

Each file is linked with other with "ISR Number" Individual Safety Report which acts as the primary key. Another significant field is "CASE Number" Case number allows identifying all ISRs of the same reports. This field is critical in de-duplication process. Indeed, "duplicate" ISRs (multiple reports of the same event) will generally have the same CASE number (but different ISR numbers).

Mining process:

The overall mining process consists of four steps as specified in Figure 2.
Table 1: Entity description.

<table>
<thead>
<tr>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMO</td>
<td>Includes information about “event date”, patient “age” and “gender”, “reporter country” and “reporter’s type of occupation”.</td>
</tr>
<tr>
<td>DRUG</td>
<td>Includes role codes assigned to each drug: “primary suspect drug” (PS), “secondary suspect drug” (SS), “interacting” (I) or “concomitant” (C).</td>
</tr>
<tr>
<td>REACTION</td>
<td>Includes all adverse drug reactions coded by MedDRA terminology.</td>
</tr>
<tr>
<td>OUTCOME</td>
<td>Contains (type of outcome, such as death, life-threatening, hospitalization, Disability)</td>
</tr>
<tr>
<td>RPSR</td>
<td>Contains information on the source of the reports (i.e. company, literature)</td>
</tr>
<tr>
<td>THERAPY</td>
<td>Contains details about drug therapy start dates and end dates for the reported drugs (0 or more per drug per event).</td>
</tr>
<tr>
<td>INDICATIONS</td>
<td>Contains all MedDRA terms coded for the indications of use (diagnoses) for the reported drugs.</td>
</tr>
</tbody>
</table>

Initially, drug names are assigned to unique code to reduce drug naming redundancy and to avoid complexity during processing. Then the set of multi-item candidates are generated using Apriori algorithm and their support value is calculated for each drug and filtered into two categories namely frequent items and infrequent items. Finally rare events are identified based on the priorities assigned to the class label outcome.

Assigning Drug Code:

One of the main challenges DMA is large variation of terminologies used to describe drugs which in turn leads to misunderstanding. Typographical errors may present in the name of the drug due to the naming conventions followed in various regions. To specify them in a standard format and to overcome such issue each drug obtained from the report is assigned to unique code. Then it can be used throughout the process without any mismatch with other drug.

Association Mining:

Multi-item Adverse Drug Event (ADE) associations are rarely reported but are essential because they could indicate possible DDIs. It is necessary to use an efficient algorithm, and also to employ additional criteria specific to the application, in order to reduce the search space. Additional criteria restrict the search space that not only have high support and giving high priority to the rare events. The general Apriori algorithm, is optimized and tailored in order to detect rare events. The Apriori algorithm is a method designed to economically to identify association rules in large database. Normally Apriori makes use of an iterative approach known as breath-first search, where k-1 item set are used to search k item sets. There are two main steps in Apriori, Join, the candidates are generated by joining among the frequent item sets level-wise. Prune which discard items set if support is less than minimum threshold value. If certain drug combinations and ADR are infrequent they should be given much importance and stored separately.

The novel algorithm known as Prior-Weighted uses filtering and assigning priorities to the class labels.

![Diagram](image)  
**Fig. 2:** Overall Mining process.
Prior-weighted algorithm:

A novel algorithm is developed to find out the rare events in drug-drug interaction. This algorithm has its foundation based on Apriori with quicker generation of candidate set giving more importance to infrequently occurring combinations. The generation of drug-drug interaction item set depend upon the user so that multiple candidate set generation and multiple database scan is avoided.

Steps in implementing prior-weighted algorithm
Input: Large database for transaction.
Output: Candidate set
Get count for combining attributes (i.e no. of combination of drugs)
Method:

Pseudocode for finding frequent item set with limited number of Iterations

M1=find frequent(X)
min-sup=average(support count) //calculate minimum support
for(i=2;M i-1=N;i++)
S i= candidate –gen (M i-1)
for each transaction t ϵ X
{S t=subset(S i,t);
for each candidate s ϵ S t
{ c.count++;
}
S i={s ϵ S t|s.count ≥min_sup }
return M i;

Candidate generation

procedure candidate–gen(M i,frequent(i-itemsets))
for each itemset J 1 ϵ M i-1
for each itemset J 2 ϵ M i-1
( J 1[i-1]>J 2[i-1] )
then
{ S=J 1×J 2
if has-infrequent subset(s, M i-1 ) then
T i=infrequent–itemset
priority(T i); // assign weights to each infrequent item
else add s to S i;
}
return S i;
priority(ArrayList list)
{
for (int i=0;i<list.size;i++)
{
prior.add(T i.get(i));
}
return 0;
}
prior.sort(); //sorting based on weights
for (int i=0;i<prior.size;i++)
{
display (prior.get(i));
}

Filtering:

Support value is generated for each drug in the AERS report this acts as the threshold value. The filtering process is done on the support value calculated. The items which have minimum value than the support are filtered and stored separately. The item set with greater value than the support are automatically added to frequent list.

Support for single Drug = \( \frac{n(D_1)}{n(S)} \)

Support for combination of Drugs = \( \frac{n(D_1 \cup D_2)}{n(S)} \)

where D denotes drug and S denotes total items in Database. The support value obtained here is 2, therefore the item set with value greater than 2 is considered as frequent items and item set whose value is less than 2 is
considered as infrequent items.

**Assigning priorities to class labels:**

Based on the “serious” patients outcome priorities are assigned to detect rare events. Once the drug and drug combinations have been filtered they are mapped to their corresponding outcomes. The outcomes includes Death (DE), Life threatening (LT), Hospitalization (HO), Disability (DS), Congenital Anomaly (CA), Required Intervention to Prevent Permanent Impairment/Damage (RI) and other serious events (OT). After assigning the priorities to the outcome the drug combination are displayed. This can be used by Pharmacovigilance cell (Phv), pharmaceutical industry and physicians so that the combination of drug that results in rare events are reported. The information generated are more useful so that these combination of drug are not prescribed in future.

<table>
<thead>
<tr>
<th>Code</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>1</td>
</tr>
<tr>
<td>LT</td>
<td>2</td>
</tr>
<tr>
<td>HO</td>
<td>3</td>
</tr>
<tr>
<td>DS</td>
<td>4</td>
</tr>
<tr>
<td>CA</td>
<td>5</td>
</tr>
<tr>
<td>RI</td>
<td>6</td>
</tr>
<tr>
<td>OT</td>
<td>7</td>
</tr>
</tbody>
</table>

Fig. 3: List of Outcomes and Priorities.

**Experiments And Results:**

FAERS reports published in the year 2013 contains 32,41104 records with 2540 unique drug names. Sample of 20,000 individual reports are considered for analysis. Prior-weighted algorithm detected frequent DDI’s and infrequent items. Table 2 provides the number of records retrieved with the combination of two drugs for various outcomes. The experimental results shows that combination of drugs like Oxycodone and Zometa are found to be frequent and the association rules generated were similar by the algorithms. However prior weighted algorithm generated more number of strong association rules than apriori algorithm. More number of strong association rules may brought the hidden DDIs present in the data set. Drugs like Duragesic along with Oxycodone, Ondanestron and Alprazolam produces DE with the count of 394. Although the percentage of such outcome is .0197, analyzing such serious outcomes may fine tune the drug safety system. The analysis helps to determine the frequency of multiple drugs with various outcomes. Figure 4 provides the performance analysis of Prior weighted and Apriori algorithm.

Fig. 4: Performance Analysis of Apriori and Prior-weighted Algorithm.

The outcomes like DS, CA and RI are zero for frequent drug combinations, which shows the impact of such outcomes are less in DDIs. The performance of the algorithms are evaluated with the results of association rule generated for the sample considered. Association rules generated by prior weighted and apriori algorithm are provided in Table 3-5.

The number of association rules generated by the prior weighted algorithm is more than the apriori algorithm. Hence applications which need preventable measures may utilize the methodology of prior weighted algorithm for determining association rules.
Table 2: Number of Frequent vs. Infrequent record count by prior weighted algorithm.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Number of Frequent drugs</th>
<th>Number of Infrequent drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>57</td>
<td>192</td>
</tr>
<tr>
<td>LT</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>HO</td>
<td>26</td>
<td>270</td>
</tr>
<tr>
<td>DS</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>CA</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>RI</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>549</td>
</tr>
</tbody>
</table>

Table 3: List of frequent Association Rules generated by prior weighted algorithm.

<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Percentage of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duragesic Mylan, Fentanyl Transdermal &gt; DE</td>
<td>3.07</td>
</tr>
<tr>
<td>Oxycodone, Prednisone &gt; DE</td>
<td>2.46</td>
</tr>
<tr>
<td>Oxycodone, Zometa &gt; DE</td>
<td>2.46</td>
</tr>
<tr>
<td>Atenolol, Aspirin &gt; HO</td>
<td>1.19</td>
</tr>
<tr>
<td>Prednisone, Prilosec &gt; HO</td>
<td>1.19</td>
</tr>
<tr>
<td>Zometa, Aredia &gt; HO</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Table 4: List of Infrequent Association Rules generated by prior weighted algorithm.

<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Percentage of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duragesic, Oxycodone &gt; DE</td>
<td>1.97</td>
</tr>
<tr>
<td>Duragesic, Alprazolam &gt; DE</td>
<td>1.97</td>
</tr>
<tr>
<td>Duragesic, Ondansetron &gt; DE</td>
<td>1.97</td>
</tr>
<tr>
<td>Duragesic, Oxycodone &gt; HO</td>
<td>.885</td>
</tr>
<tr>
<td>Duragesic, Alprazolam &gt; HO</td>
<td>.885</td>
</tr>
<tr>
<td>Duragesic, Ondansetron &gt; HO</td>
<td>.885</td>
</tr>
<tr>
<td>Atenolol, Acelovir &gt; DS</td>
<td>.08</td>
</tr>
<tr>
<td>Atenolol, Allopurinol &gt; DS</td>
<td>.08</td>
</tr>
<tr>
<td>Atenolol, Lansoprazole &gt; DS</td>
<td>.08</td>
</tr>
</tbody>
</table>

Table 5: List of frequent Association Rules generated by Apriori algorithm.

<table>
<thead>
<tr>
<th>Association Rule</th>
<th>Percentage of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxycodone, Zometa &gt; DE</td>
<td>3.07</td>
</tr>
<tr>
<td>Prednisone, Zometa &gt; DE</td>
<td>2.46</td>
</tr>
<tr>
<td>Zometa, Taxotere &gt; DE</td>
<td>2.46</td>
</tr>
<tr>
<td>Zometa, Aredia &gt; HO</td>
<td>1.62</td>
</tr>
<tr>
<td>Zometa, Tamoxifen &gt; HO</td>
<td>.31</td>
</tr>
<tr>
<td>Zometa, Dexamethasone &gt; HO</td>
<td>.25</td>
</tr>
</tbody>
</table>

Conclusion And Future Work:

The current approach in pharmacovigilance uses bivariate analysis, where each drug-AE combinations are studied separately. In this paper we propose basically a different approach taking into account multi-item combinations. We examined the feasibility of a well known data mining approach that we adopted and modified to discover multi-item combination of drug that detect rare events effectively. The proposed data mining method was applied to FAERS database and the finding demonstrate that multi-item drug combination that causes rare events could be extracted using our methodology. There are several challenges concerning our approach that should be addressed. In future we plan to address some of the issues such as: the FAERS adverse event can be coded according to MedDRA terminology, which allows to classify adverse event information associated with the use of pharmaceutical and other medical products, algorithmic improvements, a more automated and streamlined process.

REFERENCES


Rave Harpaz, Herbert, S. Chase, Carol Friedman, 10-12, 2010. Mining Multi-item Drug Adverse Effect Associations In Spontaneous Reporting Systems, Bioinformatics San Francisco, CA, USA.,


The Upsala monitoring centre. [http://www.who-umc.org].


William DuMouchel and Daryl Pregibon AT&T Labs—Research, 2013. Empirical Bayes Screening for Multi-Item Associations, 180 Park Avenue, Florham Park, NJ 07932, USA.