ISSN:0975-9646

Detecting Signals of Drug-Drug Interactions Using Association Rule Mining Methodology

Sindhu M.S, Kannan B

Dept. of Computer Applications, Cochin University of Science and Technology, Cochin, Kerala, India

Abstract— A drug-drug interaction (DDI) is a change in the effect of a drug when administered with another drug or group of drugs. Drug interactions are common and cause increased hospital admission rates, treatment failures, avoidable medical complications, and even deaths. Studies have found multiple drug usage, and age related comorbidities to be reasons for the interactions and these demand a general study. In the present study, Adverse Drug Event database is mined via the Association Rule Mining methodology using an appropriate implementation of the Apriori algorithm to detect drug-drug interactions.

Keywords— Association Rule Mining, Adverse Drug Effect, Apriori, Drug-Drug Interactions, Polypharmacy, Pharmacovigilance.

I. INTRODUCTION

Drug-drug interactions (DDIs) may increase or decrease the effects of any of the drugs or may cause an adverse effect that is not usual with either of the drugs administered together. Drug adverse reactions cause up to 5% hospital admissions, 28% emergency room visits, and 5% hospital deaths[1], [2], [3]. The American Food and Drug Administration (FDA) reckons drug-drug interaction as a critical factor in the benefit-risk assessment of a drug during development and regulatory review, and has created a database, the FDA Adverse Event Reporting System (FAERS), which contains worldwide adverse event data [4], [5]. Present adverse event (AE) databases contain AEs of only individual drugs and drug interaction databases contain data from small cohort studies or interaction studies of a small group of drugs administered for the same symptom. Present methods analyze single drug AEs from databases or conduct text search of biomedical literature, but both fail to find novel potential DDIs [6]. The set of textual patterns proposed by pharmacy experts for text mining was found to be inadequate to identify many interactions [7].The cross matching of possibly every drug that has an AE with every other drug must be done before patient exposure. The FAERS database is a repository of mandatory AEs collected from manufacturers and voluntary reports from health care professionals worldwide and can be excavated to find potential DDIs. Association rule mining methodology, which is an established method suitable for finding associations between multiple items in large databases, is used for the purpose of mining new drug-drug interactions with an appropriate implementation of the Apriori algorithm [8]. Data preprocessing and data mining algorithms are developed for the current problem. The rules thus formulated are compared by an expert with standard database for validation.

Systematic and random samplings are used for Association Rule Mining, using the Apriori and FP-growth algorithms and analysis indicate that different sampling methods can be efficiently used and behave similarly in terms of accuracy [9].

The AIS, DHCP, AIS, and Partition algorithms have been compared with the Apriori algorithm and Apriori is found to be a better performer than all the other algorithms [10]. The number of association rules generated by the Apriori algorithm was larger for all confidence and support levels and the margin improved considerably on the increase of transactions.

II. METHODOLOGY

A. DATA RESOURCE

The FAERS database is designed to support the FDA's post-marketing safety surveillance program. As per FDA regulations, the manufacturers are required to submit suspected ADRs before product marketing. The manufacturers are also mandated to report ADRs reported by consumers. FDA also receives voluntary reports from healthcare professionals all over the world. Following a manual review, these reports are entered into the FAERS database. The adverse events are described in the MedDRA (Medical Dictionary for Regulatory Activities) Preferred Term (PT) level. Drug information is in RxNorm context and includes R_xNorm code, method of administration, dosage and brand information for each drug [11], [12]. Patient outcomes, therapy dates, reporting sources, and MedDRA coded indications for reported drugs are also entered into the data base. The data base is reported to be good source for finding potential DDIs and the existing algorithms for mining bivariate relations are incapable to explore higher order ADE associations [13].

The FAERS database is available online and contains over five million reports from 1969 till present. As the size and complexity of the database became unmanageable by the traditional method of manual case reviews, data mining algorithms were designed to explore relations of drugs with adverse events.

The 2012 first three quarter reports show representation from 168 countries. In the present problem, the first three quarter 2012 reports available online were used in the mining process. The latest reports contain information about new drugs and therefore can be sources of potential novel DDIs. Moreover, many reports from the earlier dates would have become obsolete due to the withdrawal of drugs due to various reasons.

B. DATA PRE-PROCESSING

The database required to be pre-processed to suit the problem under consideration, for reducing the unnecessary complexity and duplicity.

FAERS database contains numerous reports which show adverse events caused by single drug usage. These are not sources of drug-drug interactions and are not relevant in the present study. Therefore, these reports are searched and removed based on the count of the drug sequence numbers. Now, the corresponding adverse reaction entries are also deleted being irrelevant. Objective is to find associations of the form *DrugA,DrugB > Adverse Event X*, and not of the form *Drug A>Adverse Event X*.

The highest challenge in the FAERS database is its duplicity [14]. The voluntary reports received are often duplicates reporting the same event. Some duplicate reports are present due to error in health care professionals reporting the same event. Manufacturers also receive reports from consumers or relatives regarding the adverse events. These will also create duplicity. The database does not contain patient name and death date field is no longer populated due to privacy reasons. The date of adverse event occurrence, age, gender, weight and outcome codes are compared to find duplicate reports and the corresponding entries are deleted. Reports containing serious patient outcomes like death, life threatening, hospitalization or disability are to be selected as these can be the indicators of potential DDIs.

The drug information in the database is in the RxNorm context. R_xNorm is a standardized nomenclature for clinical drugs and is produced by NLM, the National Library of Medicine.

In the R_xNorm context, clinical drug is a pharmaceutical product given to patients with a therapeutic or diagnostic intent. Therefore, in R_xNorm, the name of a clinical drug i.e. brand name combines its ingredients, strength and form. In other words, depending on weight and method of administration, the same drug has different codes.

For example,

Consider the 2012 first quarter report of the FAERS database. The drug LIPITOR is given different codes based on dosage as shown in Table 1.

TABLE 1

DIFFERENT DRUG CODES FOR THE SAME MEDICINE

Code	Dosage
1018426025	10 mg tablet
1018422348	20 mg tablet
1017686798	40 mg tablet
1018425841	80 mg tablet

If a medicine is given as injection or syrup, it will have multiple codes depending on dosage. In the mining process, this means that there will be multiple codes for any given drug, increasing the complexity of the algorithm and creating irrelevant combinations i.e. comparing the same drug.

In order to combat this problem, every drug obtained from a report is assigned a UMLS (Unified Medical Language System) drug code using MedLEE, the Medical Language Extraction and Encoding System. The goal of MedLEE is to extract, structure, and encode clinical information in textual patient reports so that the data can be used by subsequent automated processes [15]. The architecture of MedLEE is shown in figure 1.

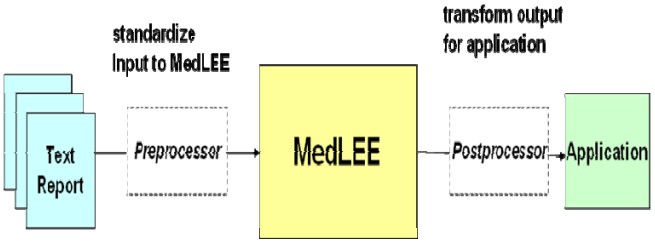

Figure 1 MedLEE Architecture

Finally, the UMLS codes are mapped to the generic using R_xNorm. Therefore, all the codes mentioned in the example corresponding to the brand name LIPITOR can be mapped to the corresponding UMLS-CODE C0286651 for the generic ATROVASTATIN.

C. DATA MINING

The search space of concomitantly administered medications is extremely large and discovering the associations between these drugs is computationally difficult considering the size of the large database. If 100 unique drugs and adverse events are to be considered, the number of DDI associations for 2 drugs and 2 AEs that need to be explored becomes $100^4 = 10^8$

The process of data mining consists of three steps.

1) Step1

The complexity of the algorithm is to be reduced. For the purpose, the drug names are mapped to their corresponding generic names. This will also help in reducing redundant searches and strengthening of signals.

2) Step2

A set of candidate drug-drug ADE associations are generated using the Apriori algorithm.

3) Step3

These associations are filtered to remove unwanted associations to identify potential DDIs.

The FAERS database is huge and the rules to be explored corresponding to two or more drugs is computationally intractable. To improve the efficiency of the algorithm and reduce search space, additional criteria to increase support and confidence of the association rules and the rules with high priority, such as rules containing certain number of items or set of items, are enforced in finding DDIs. The general Apriori algorithm which is modified for the above criteria is implemented.

The Apriori algorithm uses the downward closure property of frequency to prune the search space of association rules. In the process of DDI mining this means that if some combination of drugs and AEs are infrequent, then the larger set of combinations which gets formulated on the infrequent one will also be infrequent and can be eliminated from consideration. This pruning can be called remaining tuple optimization.

The Apriori first explores itemsets for a minimum support, and then formulates association rules based on a certain confidence from these generated itemsets. The item generation process poses more challenge as it is based on these, that all the possible rules are formulated. Hence, in order to tackle the problem with the large quantity of data in the FAERS database, the algorithm demands enhancement.

Some of the rules generated by the algorithm may contain only drugs or AEs. These are not agreeable to the ADE association definition. Therefore, the constraint is set so that only itemsets with a set of drugs in the left-hand side and a set of AEs in the right hand side of the association rules are considered. This method decreases the search space of possible multiple drug associations. Thus the rules to be generated will also be considerably reduced. Indexing based on the drug names and adverse events is used to decrease search of the entire database.

Frequent adverse events like headache or nausea usually generate large confidence values in spite of the drugs associated with them. Also, infrequent adverse events will produce small confidence levels even though they are strongly associated to some drugs. This has been verified by the present as well as other studies [15], [17], [18]. Besides this, mining the DDIs encompasses associations between multiple items and these will naturally generate less support and confidence unlike other typical association rule mining applications. FDA and WHO use PRR to monitor safety signals in their databases [19], [20], [21]. The data mining algorithms using Spontaneous Reporting System (SRS) databases use RRR (Relative Reporting Ratios) to quantify drug adverse event unexpectedness [22].

Therefore, in place of the confidence, PRR is used to enhance association strength and rule interestingness [23]. The ratio is a means which can be employed to summarize the extent of occurrence of an adverse event for a person taking a particular drug compared to the occurrence of persons' taking some other drug(s). A PRR value greater than 1 indicates that the adverse event occurs in a person taking a particular drug compared to one taking other drugs. The PRRs are similar to the proportional mortality ratios in epidemiology which are based on the knowledge that the proportional frequency of adverse events reported to the UK Yellow Card systems comparatively constant in spite of the significant increase in total reports under consideration [22].

Consider the following contingency table, Table 2.

2X2 contingency table for a drug (X)-adverse event (Y) combination, in Spontaneous reported data [20], [21].

TABLE 2

COMPUTATION OF PRR

	Adverse event(Y)	**Not adverse event(Y)**	**Total**
Using drug X	a	b	(a+b)
All other adverse events	c	d	(c+d)
Total	(a+c)	(b+d)	(a+b+c+d)

The table is constructed based on the number of reports of the combinations of interest. Drugs are reported as suspected or concomitant medications. The count of reports of a combination of interest can be based on all the reports for a drug, or only those where it is suspected as casual. When seeking interactions, all drugs are to be considered.

Here the cell 'a' is the observed number of reports of drug X and AE Y and the expected count is $\frac{(a+b)(a+c)}{a+b+c+d}$

assuming no association between X and Y.

Using measures of disproportionality, unexpectedness relative to the background of the rest of the database:

$$\mathrm{PRR} = \frac{a}{a+b} \Big/ \frac{c}{c+d}$$

Previous quantitative analysis studies involving the spontaneous databases have used the proportional reporting ratios and the characteristics of their performance post-marketing surveillance and regulatory databases have been evaluated [24], [25]. With a PRR value greater than two for all adverse events occurring with frequency greater than 2, in the UK Yellow Card database, it was found that around 60% of the signals were of known adverse events. The PRR threshold used in the present study is set to 2 based on similar studies [19], [22], [25]. The support threshold to validate the adverse drug event association rules is set to 50. This is done considering the size of the database as well as to mine more frequent patterns. The threshold resulted in less variation in content instead of larger set of associations generated by smaller values.

To suit the problem of finding associations between drugs and their adverse effects, PRR is computed for each drug of a particular report based on other drugs of that report. The Apriori is implemented based on PRR instead of confidence to find the associations. Finally the rule base is populated with the associations containing at least two drug codes and an adverse event.

III. METHOD OF EVALUATING RESULTS

Micromedex is recommended to be a reliable reference for the result evaluations by clinical subject matter experts [26]. This healthcare information system is developed by Truven Health Analytics. It is an online database that includes referenced information about drugs, toxicology, diseases, acute care, and alternative medicine for healthcare professionals to make informed clinical diagnosis and treatment decisions.

Micromedex uses different types of online systems.

1) CareNotes System

CareNotes is a patient education product with information about aspects of patient care, medical conditions and treatment, medications, and health, in up to 15 languages.

2) RED BOOK Online

This can be accessed through micromedex 2.0. Daily access is provided to drug pricing and descriptive information for more than 200,000 active and deactivated FDA-approved prescription and over-the-counter (OTC)

medications, nutraceuticals, bulk chemicals as well as for some medical devices and supplies.

3) Formulary Advisor

Formulary Advisor is an easy-to-use online formulary management tool to effectively manage and update a hospital's formulary and communicate the most current formulary information facility-wide.

4) PDR Electronic Library

The PDR Electronic Library provides access to FDA approved drug information like drug interactions, side effects, recommended dosages, contraindications, etc.

The service of a clinical subject matter expert was employed in evaluating the association rule generated using the Micromedex. If an association shows and AE which is not characteristic of any of its component drugs as per the PDR library, it is a clear indication of DDI.

IV. Data And Result Statistics

Table 3 indicates data based on the FAERS database and result generated. Figure 2 indicates the statistics of association rules generated based on comparison with known associations using Micromedex.

TABLE 3

DATABASE AND RESULT STATISTICS

Total Number of reports	593679
Reports involving more than one drug	379184
Total distinct drug names	60189
Total generic drug codes after drug name mapping	17196
Total MedDRA coded AEs	11598
Average number of drugs per association	3.2
Average number of AEs per association	3.5
Total items per association	6.7
Total number of associations generated	1224

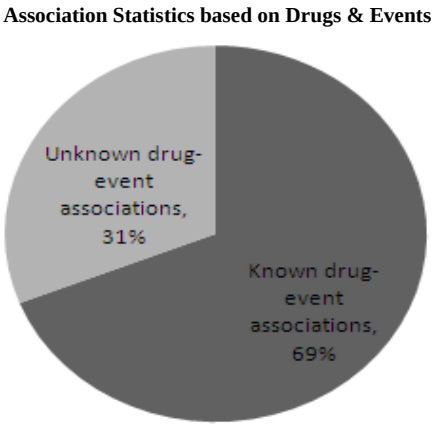

Figure 2 Percentage of Drug-AE Associations- Known & Unknown

1224 associations were generated of which 844 (69 %) were known and 380 (31%) were unknown. 69% represents the DDI identification research till date, in vitro, in vivo and in silico and substantiates for the validity of the result. The success and advantage of the methodology is that unlike the usual pharmacological and small cohort study methods done previously, it can be easily be repeated on an updated database to get new results i.e. results involving new drugs and new AEs.

V. Conclusions

Drug Interaction Management demands tools to improve guideline quality [27]. A number of software tools have been developed to cross match drugs to avoid DDIs during various stages of administration, using the currently known drug interactions. Studies have even been conducted comparing standard tools for known DDI checking [28]. All these tools can be efficiently used for improving pharmacovigilance if and only if new potential DDIs are identified which can be done by the present method.

Acknowledgment

We thank Dr. K.V. Pramod, Professor, Dept. of Computer Applications, Cochin University of Science and Technology for his encouragement in writing this article.

References

[1] M. Pirmohamed, S. James, S. Meakin, C. Green, A. K. Scott, T. J. Walley, K. Farrar, B. K. Park, and A. M. Breckenridge, "Adverse drug reactions as cause of admission to hosptial: prospective analysis of 18,820 patients," BMJ, vol. 329, pp. 15-19, Jul. 2004.

[2] N. Patel, S. Abdelsayed, M. Veve, and C. D. Miller, "Predictors of clinically significant drug-drug interactions among patients treated with nonnucleoside reverse transcriptase inhibitor, protease inhibitor, and raltegravir-based antiretroviral regimens," The Annals of Pharmacotherapy, vol. 45, no. 3, pp. 317-324, 2011.

[3] L. Juntti-Patinen and P. J. Neuvonen, "Drug-related deaths in a university central hospital," European Journal of Clinical Pharmacology, vol. 58, no. 7, pp. 479-482, Oct. 2002.

[4] (2012) The U.S. Food and Drug Administration website. [Online]. Available http://www.fda.gov/

[5] L. Zhang, Y. Zhang, P. Zhao, and S. Huang, "Predicting Drug-Drug Interactions: An FDA Perspective," The AAPS Journal, vol. 11, no. 2, pp. 300-306, Jun. 2009.

[6] I. Segura-Bedmar, P. Martínez, and C. Pablo-Sanchez, "A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents," BMC Bioinformatics, vol. 12, issue Suppl 2:S1, pp. 1-11, Mar. 2011.

[7] I. Segura-Bedmar, P. Martínez, and C. Pablo-Sánchez, "Extracting drug-drug interactions from biomedical texts," BMC Bioinformatics, vol. 11, no. Suppl 5: P9, pp. 1-2, Oct. 2010.

[8] R. Agarwal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD international conference on Management of data 1993, p207-216.

[9] V. Umarani, and M. Punithavalli, "An Empirical Analysis over the Four Different Methods of Progressive Sampling-Based Association Rule Mining," European Journal of Scientific Research, vol. 66, no. 4, pp. 620-30, Dec. 2011.

[10] F. H. Al-Zawaidah, Y. H. Jbara, and M. A. Abu-Zanona, "An Improved Algorithm for Mining Association Rules in Large Databases," World of Computer Science and Information Technology, vol. 1, no. 7, pp. 311-316, 2011.

[11] (2012) MedDRA - The Medical Dictionary for Regulatory Activities website. [Online]. Available http://www.meddramsso.com/

[12] (2012) The National Library of Medicine, USA website.[Online]. Available http://www.nlm.nih.gov/research/umls/rxnorm/

[13] W. P. Stephenson, and M. Hauben, "Data mining for signals in spontaneous reporting databases: proceed with caution," Pharmacoepidemiology & Drug Safety, vol. 16, no. 4, pp. 359-65, Apr. 2007.

[14] M. Hauben, L. Reich, J. DeMicco, and K. Kim, "Extreme duplication in the US FDA Adverse Events Reporting System database," Drug Safety, vol. 30, no. 6, pp. 551-554, Jun. 2007.

[15] (2012) MedLEE- Medical Language Extraction and Encoding System website. [Online]. Available http://www.medlingmap.org/

[16] S. E. Brosette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser, "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance," Journal of the American Medical Informatics Association, vol. 5, no. 4, pp. 373-81, Jul-Aug. 1998.

[17] L. Ma, F. Tsui, W. R. Hogan, M. M. Wagner, and H. Ma, "A Framework for Infection Control Surveillance Using Association Rules," AMIA Annual Symposium Proceedings Archive, vol. 2003, pp. 410-414, 2003.

[18] M. Rouane-Hacene, Y. Toussaint, and P. Valtchev, "Mining Safety Signals in Spontaneous Reports Database Using Concept Analysis," Artificial Intelligence in Medicine Lecture Notes in Computer Science, vol. 5651, pp. 285-294, 2009.

[19] A. Szarfman, S. G. Machado, and R. T. O'Neill, "Use of Screening Algorithms and Computer Systems to efficiently signal higher-than-expected combinations of drugs and events in the US Spontaneous Reports Database," Drug Safety, vol. 25, no. 6, pp. 381-392, 2002.

[20] A. Bate, M. Lindquist, I. R. Edwards, S. Olisson, R. Orre, A. Lansner, and R.M. De Freitas, "A Bayesian neural network method for adverse drug reaction signal generation," European journal of clinical pharmacology, vol. 54, issue. 4, pp. 315-321, Jun. 1998.

[21] A. Bate and S. J. Evans, "Quantitative signal detection using spontaneous ADR reporting," Pharmacoepidemiology & Drug Safety, vol. 18, no. 6, pp. 427-36, Jun. 2009.

[22] M. Hauben, D. Madigan, C. M. Gerrits, L. Walsh, and E. P. Van Puijenbroek, "The role of data mining in pharmacovigilance," Expert Opinion on Drug Safety, vol. 4, no. 5, pp. 929-948, Sept. 2005.

[23] Stephane Lallich,Olivier Teytaud and Elie Prudhomme, "Association Rule Interestingness: measure and statistical validation," Quality measures in data mining Springer (Ed.) (2006) 25.

[24] R. H. Meyboom, A. C. Egberts, I. R. Edwards, Y. A. Hekster, F. H. de Koning, and F. W. Gribnau, "Principles of signal detection in pharmacovigilance," Drug Safety, vol. 16, no. 6, pp. 355-65, Jun. 1997.

[25] S. J. W. Evans, P. Waller, and S. Davis, "Proportional reporting ratios: the use of epidemiological methods for signal detection," Pharmacoepidemiology Drug Safety, vol. 7, Suppl. 2:S102, 1998.

[26] (2012) Truven Health Analytics Micromedex website. [Online]. Available http://www.micromedex.com/

[27] A. Floor-Schreudering, P. A. G. M. De Smet, H. Buurma, S. Amini, and M. L. Bouvy, "Clarity and Applicability of Drug-Drug Interaction Management Guidelines: A Systematic Appraisal by General Practitioners and Community Pharmacists in the Netherlands," Drug Safety, vol. 48, no. 8, pp. 683-90, Aug. 2011.

[28] N. A. Perkins, J. E. Murphy, D. C. Malone, and E. P. Armstrong, "Drug Interactions: Performance of Drug-Drug Interaction Software for Personal Digital Assistants," The Annals of Pharmacotherapy, vol. 40, no. 5, pp. 850-55, May. 2006.