A Review of Different Available Data Sources and Its Limitations for Applying Data Mining Techniques in Pharmacovigilance

Novia Nurain
CSE Department
United International University
Dhaka,
Bangladesh.

Moin Mostakim
CSE Department
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Chowdhury Mofizur Rahman
CSE Department
United International University
Dhaka, Bangladesh.

ABSTRACT
Pharmacovigilance is a science for evaluating and improving the safety of different medicines. It is a collection of activities which are performed to detect, assess, understand, monitor, or to prevent the responses (good or bad) of medicines referred as Adverse Drug Reactions (ADRs). Pharmaceutical industries mainly rely on different data mining techniques for determining the hidden patterns from the huge data collected from healthcare professionals and patients to detect or to reduce Adverse Drug Reactions (ADRs). This recent introduction of different data mining techniques to pharmacovigilance emphasizes the need for the selection of data sources containing quality data. Therefore, in this paper, we investigate the characteristics of different existing conventional data sources such as protein structure based database, chemical structure based database, clinical trials database, spontaneous reporting database, prescription event monitoring database, large linked administrative database, and electronic medical records along with some non-conventional data sources, i.e., biomedical literature and health forums. Besides, we study the advantages and limitations of these data sources, which will eventually help to select effective data source for applying data mining techniques for future research and for developing new drugs. Moreover, after analyzing all the available data sources along with their limitations, in this paper, we emphasizes the need of a hybrid data source in order to apply different data mining techniques.

Keywords
Adverse drug reaction, Data mining, Pharmacovigilance, Post-marketing surveillance, Pre-marketing surveillance

1. INTRODUCTION
Medicines are prescribed to treat, cure or prevent diseases. They can be in different formats i.e., tablets, capsules, liquid or injections. Every medicine in any form may cause diverse adverse reactions varying from person to person. These reactions can be of short term or long term. The response of people to the used medicine is known as Adverse Drug Reaction (ADRs). It is also referred as Adverse Drug Effect (ADEs). Hence, caution should be taken seriously while taking medicines, as severe short-term and long-term adverse drug reactions (ADRs) can lead to serious harm to patients. Recently, adverse drug reactions have been estimated to cause 5% of hospital admissions [1], 28% of all emergency department visits [2], and 5% of hospital morbidity [3] in each year. Besides, over the past 10 years, both reported ADRs and related deaths has increased 2.6 times and a number of drugs have seen to withdrew from the market after exhibiting extreme ADRs.

A drug may create reactions to human body, which may vary from person to person. Being a unique drug user with different life style every person's body react in different way. These responses (reactions) to the medicine are known Adverse Drug Reaction (ADRs). According to Pirmohamed et al. ADRs referred as "any unintended and undesirable effects of a drug beyond its therapeutic effects occurring during clinical use" [1]. ADRs now become a huge concern for the pharmaceutical industry. The pharmaceutical industry undergoes long and expensive processes before introducing a new drug into the market as drugs need to be tested for possible ADRs. Besides, the drugs also need continuous monitoring in post marketing for determining the hidden ADRs, which remain dormant in the drug discovery phase. Many ADRs, such as prescription errors, are easily preventable, however, others ADRs that are unknown in the marketing often lead to the removal of the drugs from the market. For example the interaction between terfenadine and cytochromal p450 enzyme inhibitors causing cardiac arrhythmias was recognized 7 years after the introduction of the drug in the market [4]. Therefore, it is crucial to predict and monitor a drug's ADRs throughout its life cycle, from pre-clinical screening phases to post market surveillance.

Pharmacovigilance (PhV) is the science that concerns with collecting the information from the health providers and patients, detecting, analyzing, understanding, and preventing ADRs to enhance patient's safety. The aims of pharmacovigilance are:

1) Identifying the adverse reactions of existing drugs.
2) Exploring unexpected effects of newer drugs
3) Evaluating the risk factors associated with developments of adverse drug reactions.
4) Quantitative estimation of mortality ratio and admission to the hospitals due to ADRs.

Therefore, PhV can be divided into two stages: 1) pre-marketing surveillance where PhV focus on predicting potential ADRs using information collected from the pre-clinical screening (e.g., bioassay data)and clinical trials; and 2) post-marketing surveillance where PhV predict ADRs using accumulated data throughout a drug's market life.

Recently, with the development of large electronic health data storage system, other non conventional storage system: biomedical literatures and patient-reported data in online health forums, powerful computers, and new statistical algorithm, there has been an increased interest in applying

data mining methodologies for predicting potential ADRs. Data mining, also known as knowledge based discovery from data (KDD), is the process for discovering interesting hidden patterns in large datasets. Data mining techniques used sophisticated mathematical algorithms to analyze data to summarize useful information. Besides, these processes, which are used in marketing industry, have gained popularity in various fields such as web mining and information science; however, existing literatures contain very little information on their applications in pharmacovigilance. Therefore, recent studies mainly focus on applications, issues and challenges of data mining techniques in pharmacovigilance. However, the performance of the data mining techniques in most cases depends on the quality of data. Hence, selection of data is considered as an important step in KDD. As a collection of data is very expensive, the data mining techniques for the purpose of pharmacovigilance are applied on the existing data sources. To predict effectively the potential ADRs using mining algorithms we would require high quality large data source with quality data having strong association of ADEs with the drug. Several unique data sources are available for both pre and post marketing PhV. The most common databases used to detect ADRs are: clinical trials database, specialist database, spontaneous reporting database, prescription event monitoring database, linked administrative database, electronic medical records, biomedical records, and health forums.

Therefore, in this paper, we review the characteristics of all the currently available data sources along with their limitations which would helps to select proper database for detecting potential ADRs effectively.

2. A Survey of Different Available Data Sources for Data Mining

Historically, PhV has relied on biological data experiments or manual review of case reports; however, due to the vast quantities and complexity of data to be analyzed, computational methods that can accurately detect ADRs have become a critical component in PhV. Large scale compound databases containing structure, bioassay, and genetic information, such as NIH's molecular Libraries Initiative [5], as well as comprehensive clinical data sets such as electronic medical record database, have become the commonly used resources for computerized ADRs detection methods. As discussed in the previous section we need to monitor a drug's ADRs throughout its life cycle from pre-market to post-market. Therefore, this section focus on describing the characteristics of databases used throughout the world to detect potential ADRs for both pre and post marketing phase.

2.1 Data Sources for Pre-Marketing Surveillance

Drug discovery is an expensive and time consuming process. In order to introduce a new drug to the market, it can take at least 10 long years and billions of dollars due to high failure rate of drug candidates in clinical trials. Therefore, at pre-marketing phase PhV devoted to predict or assess potential ADRs early in the drug development pipelines. To do so, they use data containing preclinical characteristics of the compound (e.g., protein structure, chemical structure) and clinical trial database.

2.1.1 Protein Structure Based Database

Drugs typically work by activating or inhibiting the function of a protein, which in turns results in therapeutic benefits to a patient. Thus, drug design essentially involves the design of small molecules that have complementary shapes and charges to the protein target with which they can bind and interact. Fliri et al. have shown that drugs with similar in protein binding profiles tend to exhibit similar side effects [6]. Therefore, potential ADRs can be detected using protein structure data of drugs. However, construction of this database highly depends on the availability of gene-expression data observed under chemical perturbations by a drug, which make this data source expensive.

Drawbacks:

- Construction of this data set requires to have gene-expression data of a drug.
- Expensive data source

2.1.2 Chemical Structure Based Database

Bender et al. established a link between ADRs to the chemical structure of a drug [7]. However, using only chemical information to predict ADRs fails to investigate other known ADRs. Besides, it has been established by Lie et al. that an integrated database with both biological and chemical information performs better for detecting potential ADRs than using data-source containing only chemical structure information [8].

2.1.3 Clinical Trials Based Database

Before putting a drug into the market, the drug undergoes extensive trials to determine the side effects. These clinical trial databases contain valuable information for detecting ADRs. For example, data mining has been used to explore ADRs of arrhythmia utilizing cardiovascular clinical trial databases [9]. However, these trials are performed on few patients; therefore, rare events most likely remain unexplored.

Drawbacks:

- Information is incomplete
- Performed for short period of time
- Do not detect late-onset or rare adverse-effect
- Perform on few patients excluding patients with co-morbid disease.

2.2 Data Sources for Post-Marketing Surveillance

Although drugs undergo extensive screening in the pre-market surveillance, however many ADRs may still be missed because the information of clinical trials are often short, biased by excluding patients with co-morbid diseases. Pre-marketing surveillance fail to mirror actual clinical use situations for diverse population, thus emphasizes the importance of post-market surveillance. For example, serious risk of cardiac events with rofecoxib was identified one year later the marketing of this drug [9]. Several unique data sources are available for post-marketing PhV. Now, we describe the characteristics of such data sources.

2.2.1 Spontaneous Reporting System Based Database

Spontaneous reporting systems (SRSs) have served as the core data-collections system for post-marketing surveillance. Some prominent SRSs are maintained by US FDA (Food and Drug Administration) and the VigiBase managed by the World Health Organization (WHO). Although the reports may

differ in structure and content, most of them rely on health-care professionals and consumers to identify and report suspected cases of ADRs. SRSs contain information regarding the drugs suspected to cause the ADR, related other drug, indications, suspected events, and limited geographic information. Besides, these databases contain large volume of data, for example the FDA spontaneous reporting database contains over 2 millions reports over a period of 35 years, therefore, efficient to use for data mining to obtain detail ADEs.

Drawbacks:

- Being a passive system SRSs suffers from inconsistent reporting with more frequent reporting for unusual reactions, reactions for new drugs and serious reactions.
- Requires association to be recognized and report to be submitted.
- Much duplication of reports.
- Sensitivity loss due to the use of synonyms for drugs and events.
- Lack of different regulatory reporting requirements.
- Unknown exposure rate.
- For a given report, there is no certainty that a suspected drug caused the reactions.

2.2.2 Prescription Event Monitoring Database

Collecting high-quality data from family doctors, from a selected group of patients exposed to a specific (new) drug for a limited period of time, known as prescription event monitoring (PEM). Drug Safety Research Unit, Southampton, UK maintains PEM database in order to monitor post market PhV on a specific drug. Heeley et al. discuss the role of database exploration in order to detect ADRs signals from a PEM database containing 1millions reports of events from 78 PEM studies [12].

Drawbacks:

- One of the major drawbacks for PEM database mining is the lack of an adequate control group, as the database contains details of cluster of patients exposed to certain drugs. For example, tolterodine did not exhibit presence of hallucinations as an ADR because PEM database contained patients prescribed other drugs known to cause hallucinations. When the other drugs are removed from the prescription the ADR of toltero is discovered [12].

2.2.3 Linked Administrative Database

Large linked health administrative databases, such as Medicaid in the USA and the Ontario provincial databases, the Saskatchewan-linked administrative healthcare utilization database [13], the Tayside Medicines Monitoring (MEMO) containing millions of subjects can be used as a source for data mining. The data are available at a large quantity with relatively small additional costs and are not subject to recall or interviewer bias.

Drawbacks:

- The completeness of details are questionable, therefore, data set may not be accurate for all fields.
- Most of the datasets tend to construct based on collecting information on elderly or low-income population only which fails to represent a whole population.

2.2.4 Electrical Medical Records

Electronic medical records (EMRs) such as COMPETE, Hamilton, and GPRD have emerged recently as a prominent resource for data mining in concomitant. In addition, EMRs contain a large number of data fields, including clinical details such as the use of tobacco products, smoking and nonprescription drugs, symptoms and signs, laboratory data, detailed personal information, social circumstances, on small number of people. Due to extensive data sets EMR can be used to generate new diagnosis or adverse events, hypotheses, which are not restricted to existing diagnosis.

Drawbacks:

- The data sets are constructed by performing investigation on small number of people; therefore, there is a possibility to miss rare events.
- The construction of such massive data sets with both clinical and personal information can be very expensive.

2.2.5 Biomedical Literature

Biomedical literature can be used as a complementary resource for prioritizing drug-ADR associations generated from SRSs. For example, Shetty and Dalal retrieve articles published between 1949 and 2009 that contain mentions of a predefined list of drug-and-ADR pairs with 38 drugs and 55 ADRs from PubMed [13].

Drawbacks:

- This resource can contain irrelevant information, therefore, need to remove the irrelevant information applying some classification algorithm.
- The construction of such resource is time consuming and laborious.

2.2.6 Health Forums

Data posted by user on-health-related websites like DailyStrength [14] may also contain valuable drug safety information. These data source could also be used to detect ADRs from the user post and also can be used to identify the withdrawn drugs successfully based on the messages even before they were removed from the market.

Drawbacks:

- Due to absence of any monitoring body, the authenticity of the messages on the health-related website is questionable. Therefore, using this information may often lead to generation of false signal for ADRs.

3. DISCUSSION

This paper, investigates the characteristics of different existing data source for applying data mining in pharmacovigilance along with their limitations. We can divide the data source into two categories: 1) pre-marketing surveillance, 2) post-marketing surveillance. Pre-marketing surveillance concentrate on detecting ADRs using pre-clinical characteristics. Therefore, protein and chemical structure based database are used for this purpose. However, an integrated dataset along with biological and chemical information performs better while investigating ADR signals. In addition, clinical trials data could be used for screening in pre marketing phase. However, the information of clinical

trials is not complete as they are conducted on few people leaving people with co-morbid disease.

Next, we discussed another popular data source, prescription event monitoring database, which contains good quality large volume of data. However, suffers from inadequate control group and deals only with a few selected drugs.

Table I: Summery of all data sources along with their advantages and disadvantages used for data mining in

Phase	Data Source	Advantages	Disadvantages
Pre – Marketing Surveillance	Protein Structure Based Database	a) Consider direct interaction of drugs with protein targets	a)Need gene-expression data b) Expensive
	Chemical Structure Based Database	a) Consider links between ADRs and chemical structures	a) Fails to investigate other known ADRs
	Clinical Trials Database	a) Ensure extensive screening	a) Information is incomplete b) Small, short and biased c) Performed on few patients d) Exclude patients with co-morbid diseases
Post – Marketing Surveillance	Spontaneous Reporting System Based Database	a) Relatively cheap b) Extensively large c) Asses the whole population	a) Not complete b) Requires association to be recognized c) Reports need to be submitted d) Unknown exposure rate e) Duplicity in reports f) Lack of regulatory reporting requirements g) Loss of sensitivity
	Prescription event monitoring database	a) Large b) Good quality data c) No need to identify association	a) Consider only a few selective drugs b) Inadequate control group
	Large linked administrative database	a) Extensively large b) Relatively cheap	a) Don't asses whole population b) Data set may not accurate for all fields
	Electronic Medical Reports	a) More extensive dataset b) Include clinical information along with personal information	a) Data set considers small number of people b) Expensive
	Biomedical literatures	a) Used as complimentary resource for prioritizing drug-ADR associations	a) Contain irrelevant information b) Time consuming and laborious
	Health forums	a) Contain valuable information b) Easy to construct	a) Absence of monitoring body b) Authenticity of data is questionable

For post-marketing surveillance, there exist several unique data sources. Examples include, spontaneous reporting database, prescription event monitoring database, large linked administrative database, electronic medical records, biomedical literature, and health forums.

Here, spontaneous reporting database served as the core data-collection system for post-marketing drug surveillance. This data source is relatively large as it assessed the whole population, at the same time it is cheap. However, it requires recognizing associations and to reports needing to submit. This data source suffers from lack of regularity reporting, duplication of reports, and loss of sensitivity.

In addition, we also investigate large linked administrative database which is extensively large and relatively cheap. However, don't take into consideration the whole population. Therefore, data set may not accurate for all fields.

Next, we discuss another prominent data source useful for observational research which is electronic medical records. This data base contains both extensive clinical and personal information. However, this data source considers only a small number of people.

Finally, we investigate some non conventional data source such as biomedical literature and health problem. Here, biomedical literature requires extensive collection of

literatures, removal of irrelevant articles which makes the construction of this databases time consuming and laborious. On the other hand, different health forums contain valuable information, however, suffers from lack of monitoring body and authentication of data sets.

The summary of all the existing data sources along with their advantages and disadvantages is shown in **Table 1** for a quick glimpse. Analyzing the advantages and disadvantages of all the data source, it is evident that each of the single data sources fail to detect all the potential ADRs, however, it is desirable to incorporate various data source into one framework to understand ADRs well by applying different data mining techniques.

It becomes quite evident from our discussion that each of the existing data source has some limitations. None of them solely can be used for applying data mining techniques. Hence we feel the need to introduce a hybrid data source. Our proposed data source will contain information on both clinical and personal information of patients along with the protein and chemical structural information of the drugs corresponds to each patient. This hybrid data source will be able to detect ADRs more efficiently comparing the other conventional sources as the association of ADRs with drug will be prominent.

4. CONCLUSION AND FUTURE WORK

ADRs are common phenomena and result in significant mortality, and despite of existing systems drugs with adverse effects have been withdrawn many years after licensing. Therefore, due to the development of massive data storage system and powerful computer systems data mining techniques are used to detect potential ADRs more efficiently. Besides, performance of data mining approaches depends mostly on data sets. Therefore, in this paper, we perform elaborate analysis on the characteristics of all the existing data sources at the same time we highlight the limitations of each data sets. We hope the study would help to select an effective data source for applying the data mining techniques to detect potential ADRS. In our future work, we plan to construct a hybrid database and evaluate the performances of different data mining techniques to detect or reduce ADRs using the hybrid data source.

5. REFERENCES

[1] Pirmohamed, M., Breckenridge, A. M., Kitteringham, N. R., and Park, B. K. 1998 Adverse drug Reactions. Medical Journal, pp. 1295-8.

[2] Patel, P. and Zed, P.J. 2002 Drug-related visits to the emergency department: how big is the problem?. In Pharmacotherapy, pp. 915–23.

[3] Juntti-Patinen, L. and Neuvonen, P. J. 2002 Drug-related deaths in a University central hospital. In European Journal of Clinical Pharmacology, pp. 479–82.

[4] Honig, P. K., Woosley, R. L., Zamani, K., Conner, D. P. and Cantilena, L. R. 1992 Changes in the pharmacokinetics and electrocardiographic pharmacodynamics of tertenadine with concomitant administration of erythromycin. Available at Clinical Pharmacological Therapy, pp–231-8.

[5] Austin, C. P., Brady, L. S., Insel, T. R. and Collins, F. S. 2004 NIH Molecular Libraries Initiatives. Available at Science, vol. 306, pp. 1138–1139.

[6] Fliri, A. F., Loging, W. T., Thadeio, P. F. and Volkmann, R. A. 2005 Analysis of drug-induced effect pattern to link structure and side effects of medicine. Available at National Chemical Biology, pp. 389–397.

[7] Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., Urban, L., Whitebread, S. and Jenkins, J. L. 2007 Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. Appeared in ChemMedChem, vol. 2, pp. 861–873.

[8] Liu, M., Wu, Y., Chen, Y., Sun, J., Zhao, Z., Chen, X. W., Matheny, M. E. and Xu, H. 2012 Large-scale Prediction of Adverse Drug Reactions Using Chemical, Biological, and Phenotypic Properties of Drugs. Available at Journal of Medical Information Association.

[9] Cerrito, P. 2001 Application of data mining for examining poly-pharmacy and adverse effects in cardiology patients. Available at Cardiovascular Toxicology, pp. 177–9.

[10] Bombardier, C., Laine, L. and Reicini, A. 2000 Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group, pp. 1520-8.

[11] Heeley, E., Wilton, L. V. and Shakir, S. A. 2001 Automated signal generation in prescription-event Monitoring. Appeared in Drug Saf, pp. 423–32.

[12] Rawson, N. S. and Rawson, M. J. 1999 Acute adverse event signaling scheme using the Saskatchewan Administrative health care utilization darafiles: results for two benzodiazepines. Appeared in Canadian Journal of Clinical Pharmacology, pp. 159–66.

[13] Shetty, K. D. and Dalal, S. R. 2011 Using information mining of the medical literature to improve drug safety. Appeared in Journal of Journal of Medical Information Association, pp. 668–674.

[14] DailyStrength http://www.dailystrength.org .