Pharmacovigilance, signal detection using statistical data mining methods

Article *in* International Journal of Engineering & Technology · May 2018

DOI: 10.14419/ijet.v7i2.31.13423

4 authors, including:

G.V. Sriramakrishnan

22 PUBLICATIONS **33** CITATIONS

SEE PROFILE

Marimuthu Muthu Selvam

3 PUBLICATIONS **1** CITATION

SEE PROFILE

Suseendran G.

VELS INSTITUTE OF SCIENCE, TECHNOLOGY & ADVANCED STUDIES (VISTAS), CHE…

149 PUBLICATIONS **1,039** CITATIONS

SEE PROFILE

International Journal of Engineering & Technology, 7 (2.31) (2018) 122-126

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Pharmacovigilance, signal detection using statistical data mining methods

G.V. Sriramakrishnan[1*], M. Muthu Selvam[2], K. Mariappan[3], G. Suseendran[4]

[1]Assistant Professor, Department of IT, VISTAS, Chennai.
[2]Assistant Professor, Department of IT, VISTAS, Chennai.
[3]Assistant Professor, Department of IT, VISTAS, Chennai.
[4]Assistant Professor, Department of IT, VISTAS, Chennai.
**Corresponding author E-mail:greatsri8@gmail.com*

Abstract

Pharmacovigilance programmes monitor and help safeguarding the use of medicines which is grave to the success of public health programmes. Identifying new possible risks and developing risk minimization action plans to prevent or ease these risks is at the heart of all pharmacovigilance activities throughout the product lifecycle. In this paper we examine the use of data mining algorithms to identify signals from adverse events reported. The capabilities include screening, data mining and frequency tabulation for potential signals, including signal estimation using established statistical signal detection methods. We have standard processes, algorithms and follow current requirements for signal detection and risk management activities.

Keywords: *Adverse drug reactions, pharmacovigilance, safety signals, statistical methods.*

1. Introduction

An important challenge for medical industry in developing new drug is the Adverse Drug Reaction (ADR). The study of the adverse drug reactions of the newly released drugs is called Pharmacovigilance which educate the people on the benefit and risk of drugs and warn them. The quintillion of reports about adverse drug events received by the FDA every year by doctors, general public, manufacturers. As of 2016, FAERS contains over 12 million reports, and year by year over a million are added. After being recorded, the reports are inspected by individual reviewers, called Safety Evaluators, who are responsible for a defined class of drugs. Processing and analyzing this amount of data is difficult to accomplish, as the number of reports outweigh the number of report reviewers. Every month, each individual Safety Evaluator has to attempt to read and carry out an average of 3,417 adverse event reports. To support with the analysis of the enormous amounts of data, the FDA has begun deploying data mining techniques.

The Safety Evaluators, who are familiar with the current labeling, known adverse events, and mechanism of actions of their drugs, read the reports, look for particular abnormalities or issues relative to the normal product safety profile, and check the validity of the report. If the collection of reports is regarded important due to abnormalities or issues after this process, the drug and adverse event relationship is investigated more thoroughly and regulatory action may be taken.

In this paper various statistical data mining algorithms and statistical analyses used to find patterns within sets of data at the FDA. With data mining, the FDA can improve its report analysis process by automatically selecting the most significant reports for review as well as allowing reviewers to view the information from all the reports received in an organized manner, instead of having to manually consider each one. The reports that may contain serious and unexpected adverse events.

Although not yet in routine use for most applications, data mining algorithms has been successfully applied by the FDA in past years[10]. For example, in 2010 and 2011, data mining was used to identify warning signs that associated Fluzone with febrile seizures in young children[12]. Researchers calculated an Empirical Bayesian Geometric Mean for each event, which is a value used to determine relevance of reports. This value was adjusted according to the various traits of each report. Next, values fitting within a specific confidence interval were marked for further investigation by reviewers, which may have led to the identification of the safety issue.

Due to data mining's uniqueness in pharmacovigilance, the results of data mining are not depended on stand-alone; instead they are compared to the pharmaceutical knowledge of the FDA Safety Evaluators. As part of the process of testing and moving toward data mining, the FDA has applied its new data mining strategies to existing data, to demonstrate the earlier identification of safety issues [2]. Given the role of data mining algorithms in PV, the Observational Medical Outcomes Partnership is aiming to identify the most reliable algorithms for analyzing large volumes of electronic healthcare data specifically for drug safety surveillance [3].

2. Background

2.1. Data collection

The training of pharmacovigilance is important because unpredicted problems can arise after a drug is released onto the market. Problems that are undetected in the small sample sizes and limited patient demographics may arise once the drug is released into market. In order to ensure the safety of the Nation public, the

FDA collects data on these adverse events, allowing the agency to make informed decisions about what actions are needed to address the long term risks of each drug. The data on adverse events is collected through the FDA Adverse Event Reporting System (FAERS), through a process shown in Figure 1.
Figure 1 shows doctors, consumers, and manufacturers report directly to FDA, providing a detailed description of the adverse events that are assumed to be linked to a specific drug. 96% of reports submitted to the FDA come from manufacturers, who receive reports from consumers and doctors. The other 4% are directly reported to the FDA by doctors and consumers [3].

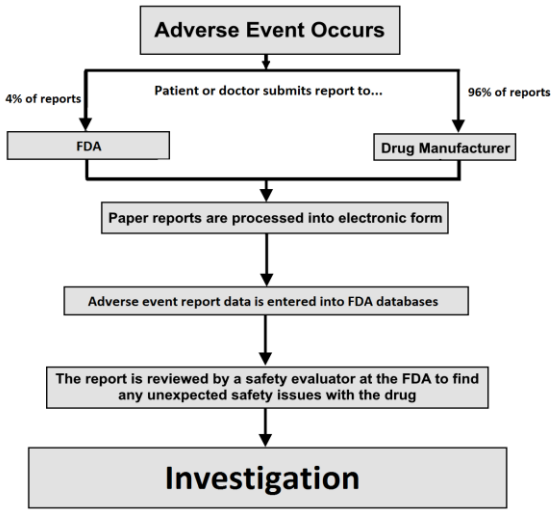

Figure 1: Adverse event monitoring process

2.2 Reporting bodies

The FDA advances data input from different bodies that may be affected by the activities such as public, consumers, nurses, sponsors, pharmacists/pharmacies, physicians, and third party payer. The reporting bodies are encouraged by the agency to use trained health care experts to assist in reporting adverse events.

2.3 Data processing

Adverse event reports are electronically submitted to the FDA or submitted by paper and entered into the FAERS database. All adverse events reports are saved in the FAERS database in the form of Electronic versions and made available for review by Safety Evaluators at the FDA [11][1]. If a report is expedited, it represents a serious adverse event that was not expedited (not in the drug's label). If a report is non-expedited, then it was serious, but expected; non-serious and expected; or non-serious but unexpected. All voluntary reports, usually from patients or health care professionals, are called direct reports.
The other format of submission of report to the FDA is paper. Figure 2 shows the overview of this process. Figure 3 highlights the processing stages. The first step in the processing of paper reports is within the Central Triage Unit (CTU) [14]. The next step for the reports after leaving the DCC is Data Entry. After all the data from the report has been entered, the information is passed to the coding team [19].

2.4 Information of patient safety

A major concern of the FDA is to maintain the privacy and security of a patient's information. This information includes human subject research and reports submitted by individual patients and practitioners. By establishing a standard to maintain this privacy, the FDA is accomplishing the goal of observing and conforming to any international definitions, laws and standards, as appropriate (USDHHS, FDA, CDER, CBER, 2005).

2.5 Safety signal identification

When an inconsistent amount of adverse events are reported to the FDA, a safety signal is generated. A safety signal is a sign of an unusual number of adverse events compared with a certain product's use [2]. The generated safety signal prompts the FDA to initiate an investigation, during which the analysis of patient demographics, length of exposure to the drug, current dosage and any past dosages, underlying health conditions, and the use of other medications are thoroughly examined (USDHHS, FDA, CDER, CBER, 2005).
Reports and records from consumers and other reporting bodies form the initial data set to identify, interpret, and develop plans to manage safety signals. If a safety signal exists, there are many ways to investigate the signal to determine if there is a potential safety risk. The FDA encourages sponsors to look at all of the various methods for safety signal investigation including pharmacoepidemiologic studies, registries, and surveys.

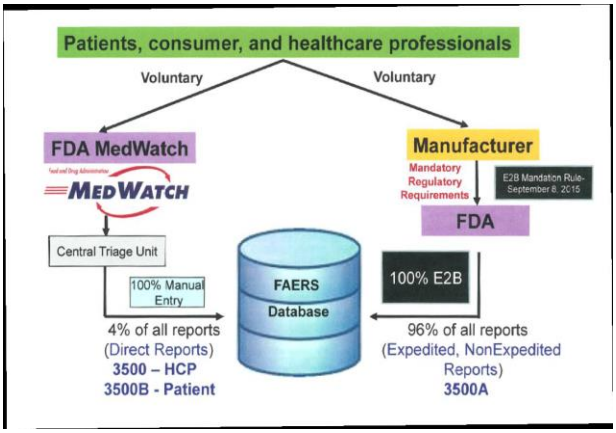

Figure 2: Overview of report processing procedure

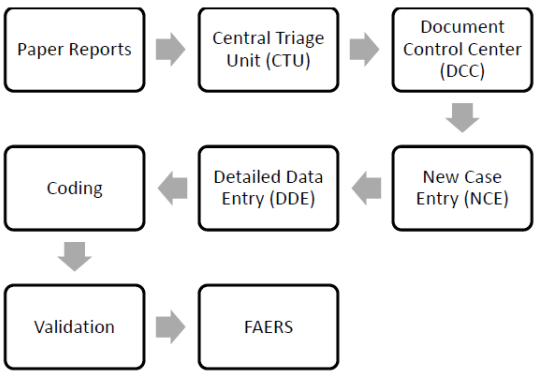

Figure 3: Stages of processing paper reports

2.6 Acting on safety signals

Based on the results of the safety signal investigation, the FDA can conduct further investigation to characterize safety signals and establish whether these signals create potential safety risks. If there is a potential safety risk, the FDA advises the sponsor to submit all safety data and the analysis methods performed. A complete submission contains all case reports: voluntary reports and case studies and literature; background information for the adverse drug event and specific affected populations; associations made from pharmacoepidemiologic studies; biologic and pharmacodynamic effects that were observed through preclinical studies; general marketing history of other similar products; and findings from controlled clinical trials. It makes possible for the FDA to evaluate the level of causality between a particular drug and the associated adverse event.

3. Related work

3.1 Data mining and pharmacovigilance

In pharmacovigilance, data mining is used to support the evaluation process in several ways such as prioritizing reports, analyzing drug-drug interactions, and evaluating both familiar and unfamiliar classes of drugs. Prioritizing safety report is essential to pick out noise and locate reports that could be the source of a safety signal [4]. Analysis of drug-drug interactions can help point out safety signals that might not be found if only performing data analysis on one specific drug. Evaluating a class of drugs, is useful in becoming aware of a class specific trend of Adverse Events (AEs)[14].

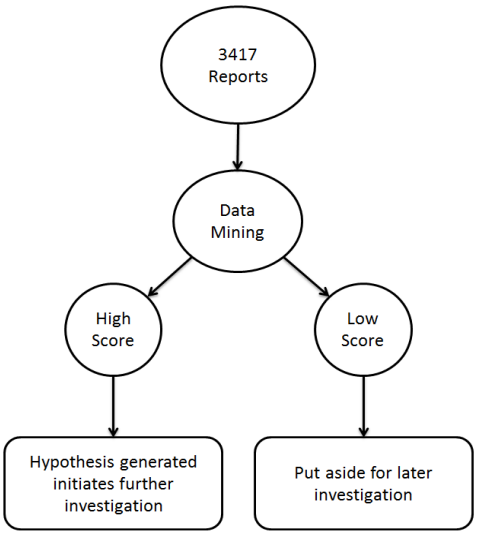

Figure 4: Hypothesis generation

3.2 Hypothesis generation

Data mining is needed in pharmacovigilance to analyze the increasing number of reports received, speed up the identification of potential safety issues, aid in hypothesis generation (shown in Figure 4), and time to in-depth evaluation [20]. Because the number of reports is growing exponentially, it is challenging for Safety Evaluators to view all of the reports within the time constraints [16]. Since not all of the reports are able to be read, reports that point out a potential safety signal might not be found [16]. With data mining, these unviewed reports can be analyzed and used to form a basis that aids evaluators in creating a hypothesis of potential safety signals. Additionally, by easing the amount of manual review that has to be conducted, data mining can give Safety Evaluators more time to focus their efforts on other time sensitive tasks.

3.3 FAERS

The FDA Adverse Event Reporting System (FAERS) is the FDA's post-market safety surveillance database [1][5]. The database contains information found in adverse event reports and medication error reports that are submitted to the FDA[7]. Implemented in 2012, FAERS was designed to support the post-market safety surveillance for drugs and therapeutic products. The database contains the validated and recoded information found in the Adverse Event Reporting System [8]. Figure 5 shows the number of reports submitted to the FDA and entered into FAERS has increased over recent years. For example, in 2004, the total number of reports entered into FAERS was 422,307 and in 2013 the number of reports climbed to 1,178,306 (Food and Drug Administration, 2013).

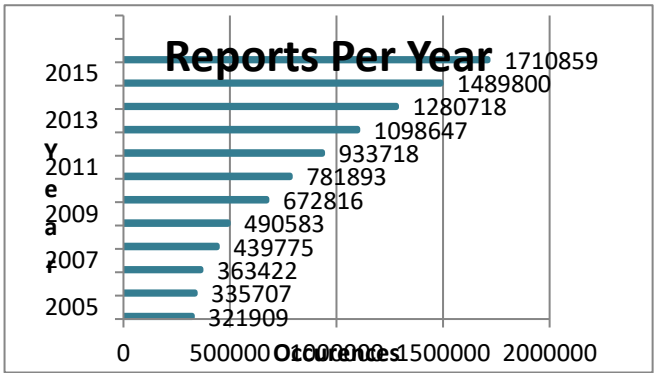

Figure 5: Number of reports submitted to the FDA yearly

4. Methods

The overall goal was to support the FDA to conduct pharmacovigilance efforts more efficiently through the development of drug list on data mining concepts and applications, as well as providing a brief overview of suitable data mining tools to the FDA's work. The objectives are, study current pharmacovigilance strategies, assess the current data mining needs of the FDA. This was accomplished in two stages, as in Figure 6 below.

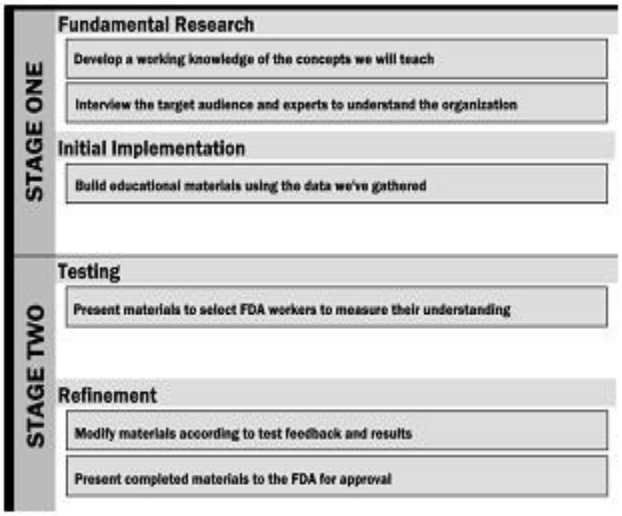

Figure 6: Overview of methodology and goals

4.1 Stage one

4.1.1 Fundamental research

Stage one consists of interviewing with several domain experts and non-experts in the fields of data mining and pharmacovigilance. It began by establishing a foundation from generating initial prototypes of safety signals. The assessment was made up of semi-structured interviews with employees who are experienced in pharmacovigilance, data mining and its applications at the FDA. This paper also involved detailed investigation into the structure of the FDA to understand the needs of the people.

4.1.2 Initial implementation

After gathering all of the information from fundamental research, this work began the initial implementation. To conclude stage one, it constructed a first draft of safety signals according to the information gathered during analysis.

4.2 Stage two

Stage two consisted of presenting the first draft to a subset of FDA employees, the target audience, after administering a pre-test.

4.2.1 Testing

The second stage began with preliminary testing of the target audience, to evaluate what knowledge they had regarding data mining. With the pre-test, we established the extent of data mining knowledge already known by the target population. A post-test was administered to evaluate the efficiency. Test results for both pre- and post-test were evaluated using a standard percentage system.

4.2.2 Refinement

In this part of the process, the final draft was created about safety signals. Revisions were made based on the results from the post-test and the learners' feedback.

4.3 Proposed methodology

In pharmacovigilance, data mining is primarily used as a descriptive task to uncover links, patterns, and similarities, allowing for clear analysis. To demonstrate these utility, four main statistical data mining algorithms was useful in pharmacovigilance (PV): Proportional Reporting Ratio, Reporting Odds Ratio, Information Component, and Multi-item Gamma-Poisson Shrinker (Empirical Bayesian Geometric Mean) because they calculate signals of disproportional reporting (SDRs).

Proportional Reporting Ratio (PRR) represents a direct measure of the strength of a safety signal. A PRR is the ratio of the proportion of all reported cases of the event of interest among people exposed to a particular drug compared with the corresponding proportion among people exposed to all or several other drugs [6]. The breakdown of PRR can be seen in the equation:

$$PRR = (A/A+B)/(C/C+D)$$

This algorithm helps with avoiding biases caused by varying details in reports. A limitation of PRR is that signals for a particular drug might reduce the magnitude of the PRR calculation for other signals of the same drug. This is due to the fact that some reports of a particular kind might appear more than others if the symptom is more common.

Reporting odds ratio (ROR), is closely related to PRR and ROR is calculated in the same manner as PRR, but ROR accounts for bias and allows for relative risk assessment.

$$ROR = (A/C) / (B/D)$$

Information component (IC) is a component of IC temporal pattern discovery (ICTPD), based on intra-personal comparison of risk periods and the preceding control period. ICTPD focuses on the exposure to a certain drug [15]:

$$IC = log_2 (A \times (A+B+C+D)/(A+D) \times (A+B))$$

ICTPD uses information from non-cases such as prescription information. The goal of this technique is to identify patterns in the associations between the prescription of a drug and the occurrence of a medical event.

Multi-item Gamma-Poisson Shrinker (MGPS) is calculated in a similar manner to PRR, but incorporates Bayesian "shrinkage" and stratification to produce scores where there is limited data and small number of cases [13]. Bayesian "shrinkage" can be summarized as the improving of an estimate by combining the estimate with other information. Stratification is a procedure for mitigating effects of confounding by adjusting for associations between a drug and a variable and an event and the same variable [9]. The differences in MGPS from PRR diminish the effect of outliers, reducing the number of false-positive safety signals. As a result, MGPS provides a more stable estimate of the relative reporting rate for a particular product.

The four algorithms developed to identify drug-associated adverse events were all developed to calculate signal scores (to assess whether a drug is associated with an adverse event or not). These algorithms are also known as signal detection algorithms (SDAs) [17].

Performance in these algorithms can be defined by sensitivity and specificity. In the context of mining adverse event signals, sensitivity is defined as, "the ability of a surveillance or reporting system to detect true health events, i.e. the ratio of the total number of health events detected by the system to the total number of true health events as determined by an independent and more complete means of ascertainment". Specificity is defined as, "a measure of how infrequently a system detects false positive health events, i.e. the number of individuals identified by the system as not being diseased divided by the total number of all person who do not have the disease". The more specific an algorithm is the lower the sensitivity and the slower the production of true signals of disproportionate reporting (SDR). The less specific an algorithm, the greater the sensitivity and faster production of true SDRs.

5. Results and discussion

Though method comparison at the individual level is inconclusive, progress has been made in identifying the advantages, disadvantages and differences between the frequentist methods and the Bayesian methods [10]. The group of frequentist methods consists of Proportional Reporting Ratio (PRR), Reporting Odds Ratio (ROR), and Relative Reporting Ratio (RRR). The frequentist methods use ratios to find and estimate associations and are typically accompanied by hypothesis tests for independence (e.g. chi squared test, Fisher's test) and these tests are used as extra precautionary measures that take into account the sample size used while computing the association [15].

The group of Bayesian methods includes Gamma-Poisson Shrinker (GPS), Multi-item Gamma-Poisson Shrinker (MGPS) and Bayesian Confidence Propagation Neural Network (BCPNN). Methods are categorized as Bayesian if the data mining method incorporates both the disproportionality measure, the measure of how much the drug-event combination occurs "disproportionally" compared to if there was no association between the drug and event, and sample size to "shrink" the disproportionality measure toward the baseline case of no association by an amount proportional to the variability of the measure. Some general advantages and disadvantages of each group are highlighted in Table A.

Table A: Comparison of Frequentists and Bayesian Methods

	Frequentist Methods	Bayesian Methods
Tend to highlight a greater number of DEAs	X	
Tend to highlight a greater variety of DEAs	X	
Tend to highlight DEAs earlier	X	
More computationally intensive		X
More sensitive to low-frequency of reports	X	
More intuitive computations	X	
Ability to sort associations along one single dimension		X
Address reporting biases or confounding		
May result in loss of credible signals		X
Lower impact of random fluctuations of relative reporting ratio ("shrinkage")		X
Produce more false positives	X	
Produce more false negatives		X

This table summarizes the general trends of frequentist methods (e.g. PRR, ROR, IC) and Bayesian methods (e.g. BCPNN, MGPS).

In general, the frequentist group seems to highlight a greater number and variety of drug-event associations (DEAs) than the Bayesian group and tend to highlight these DEAs. By comparison, the Bayesian methods group addresses the low-frequency

reporting issue by adjusting the disproportionality measure to account for these low counts.

The FDA's intake of reports has been exponentially increasing over the years while the number of Safety Evaluators has remained relatively constant. Tables B and C illustrate the overwhelming number of reports that each Safety Evaluator must view.

Table B: Total Number of Reports

Total Reports ***All Versions of Report**	**Total Reports for SE Review** ***Only short version of the reports**
11,198,975	8,425,279

Table C: Average Monthly Safety Evaluator Reports

Average Monthly SE's report				
Months	Expedited	Non-Expedited	Direct	Total Average
Nov-16	962	459	44	1465
Oct-16	2010	1492	107	3608
Sep-16	1918	2467	110	4495
Aug-16	1584	5656	118	7357
Jul-16	1759	1571	105	3436
Jun-16	1681	1103	107	2892
May-16	1588	1746	106	3441
Apr-16	1742	885	113	2740
Mar-16	1786	845	120	2750
Feb-16	1632	1216	83	2931
Jan-16	1454	944	77	2475
Total Average Over 11 Months	1647	1671	99	3417

6. Conclusion

Data mining algorithms are becoming more frequently used as a supplement to traditional expert reviews of reports and to rapidly analyze the large volume of accumulated data. New algorithms are constantly being researched to uncover new trends and associations in data or to improve upon existing algorithms. These algorithms could be routinely applied in order to monitor, prioritize, and identify undiscovered safety signals of adverse drug events that warrant further attention. In choosing an algorithm, the most important question is not which algorithm to use but what is the correct threshold. The benefit of using multiple algorithms is that one may catch a signal that the other does not. To summarize the differences between the algorithms, the algorithms that are frequentist detected a higher number of safety signals than the Bayesian based algorithms. This comparison is only relative to a specific comparison of signals detected from handpicked drugs.

References

[1] Meyboom RH, Hekster YA, Egberts AC, Gribnau FW & Edwards IR, "Causal or casual? The role of causality assessment in pharmacovigilance", *Drug Saf*, Vol.17, (1997), pp.374-389.

[2] Harpaz R, Dumouchel W, Shah NH, Madigan D, Ryan P & Friedman C, "Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis", *Clin Pharmacol Ther*, (2012).

[3] Gould AL, "Accounting for multiplicity in the evaluation of "signals" obtained by data mining from spontaneous report adverse event databases", *Biom J*, Vol.49, No.1, (2007), pp.151-165.

[4] Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A & De Freitas RM, "A Bayesian neural network method for adverse drug reaction signal generation", *Eur J Clin Pharmacol*, Vol.54, No.4, (1998), pp.315-321.

[5] Wang X, Chase H, Markatou M, Hripcsak G & Friedman C, "Selecting information in electronic health records for knowledge acquisition", *J Biomed Inform*, Vol.43, No.4, (2010), pp.595-601.

[6] Montastruc JL, Sommet A, Bagheri H & Lapeyre-Mestre M, "Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database", *Br J Clin Pharmacol*, Vol.72,No.6, (2011), pp.905-908.

[7] Lindquist M, Edwards IR, Bate A, Fucik H, Nunes AM & Stahl M, "From association to alert--a revised approach to international signal analysis", *Pharmacoepidemiol Drug Saf*, Vol.8, (1999), pp.15-25.

[8] Xu R & Wang Q, "Large-scale combining signals from both biomedical literature and FDA adverse event reporting system (FAERS) to improve post-marketing drug safety signal detection", *BMC Bioinformatics*, (2014).

[9] Wang X, Hripcsak G, Markatou M & Friedman C, "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study", *J Am Med Inform Assoc*, Vol.16, (2009), 328–337.

[10] Sriramakrishnan & Latha Parthiban, "Data mining techniques for finding serious Adverse Events", *Journal of Chemical and Pharmaceutical Sciences*, Vol.9, No.4, (2016).

[11] Bate A & Evans SJ, "Quantitative signal detection using spontaneous ADR reporting", *Pharmacoepidemiol Drug Saf*, Vol.18, No.6, (2009), pp.427-436.

[12] Hauben M, Madigan D, Gerrits CM, Walsh L & Van Puijenbroek EP, "The role of data mining in pharmacovigilance", *Expert Opin Drug Saf*, (2005), pp.929-948.

[13] Ahmed I & Poncet A, "PhViD: an R package for PharmacoVigilance signal Detection", *R package version 1.0*. (2013).

[14] Evans SJ, Waller PC & Davis S, "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports", *Pharmacoepidemiology and drug safety*,(2001), pp.483-486.

[15] Van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R & Egberts AC, "A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions", *Pharmacoepidemiology and drug safety*, (2002), pp.3-10.

[16] Pearson RK, Hauben M, Goldsmith DI, Gould AL, Madigan D, O'Hara DJ, Reisinger SJ & Hochberg AM, "Influence of the MedDRA® hierarchy on pharmacovigilance data mining results", *International journal of medical informatics*, Vol.78, No.12,(2009), pp.e97-e103.

[17] van der Hooft CS, Sturkenboom MC, van Grootheest K, Kingma, HJ & Stricker BH, "Adverse drug reactionrelated hospitalizations: a nationwide study in the Netherlands", *Drug Saf.* (2006), pp.161-168.

[18] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR & Schacht AL, "How to improve R&D productivity: the pharmaceutical industry's grand challenge", *Nat Rev Drug Discov*, (2010), pp.203-214.

[19] Hopkins AL, "Network pharmacology: the next paradigm in drug discovery", *Nat Chem Biol*, (2008), pp.682-690.