Association Patterns in Open Data to Explore Ciprofloxacin Adverse Events

P. Yildirim
Department of Computer Engineering, Faculty of Engineering and Architecture, Okan University, Istanbul, Turkey

Keywords
Data processing, adverse drug event, clinical decision support, clinical informatics, clinical care

Summary
Background: Ciprofloxacin is one of the main drugs to treat bacterial infections. Bacterial infections can lead to high morbidity, mortality, and costs of treatment in the world. In this study, an analysis was conducted using the U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (AERS) database on the adverse events of ciprofloxacin.
Objectives: The aim of this study was to explore unknown associations among the adverse events of ciprofloxacin, patient demographics and adverse event outcomes.
Methods: A search of FDA AERS reports was performed and some statistics was highlighted. The most frequent adverse events and event outcomes of ciprofloxacin were listed, age and gender specific distribution of adverse events are reported, then the apriori algorithm was applied to the dataset to obtain some association rules and objective measures were used to select interesting ones. Furthermore, the results were compared against classical data mining algorithms and discussed.
Results: The search resulted in 6 531 reports. The reports included within the dataset consist of 3 585 (55.8%) female and 2 884 (44.1%) male patients. The mean age of patients is 54.59 years. Preschool child, middle aged and aged groups have most adverse events reports in all groups. Pyrexia has the highest frequency with ciprofloxacin, followed by pain, diarrhoea, and anxiety in this order and the most frequent adverse event outcome is hospitalization. Age and gender based differences in the events in patients were found. In addition, some of the interesting associations obtained from the Apriori algorithm include not only psychiatric disorders but specifically their manifestation in specific gender groups.
Conclusions: The FDA AERS offers an important data resource to identify new or unknown adverse events of drugs in the biomedical domain. The results that were obtained in this study can provide valuable information for medical researchers and decision makers at the pharmaceutical research field.

Correspondence to:
Pinar Yildirim
Okan Universitesi, Tuzla Kampusu
Muhendislik ve Mimarlık Fakultesi
Bilgisayar Muhendisligi Bolumu
Akfirat, Tuzla
34959 Istanbul, Turkey
Tel.:+90 542 237 04 22
Fax: +90 (216) 677 1647
Email: pinar.yildirim@okan.edu.tr

Appl Clin Inform 2015; 6: 728–747
http://dx.doi.org/10.4338/ACI-2015-06-RA-0076
received: June 19, 2015
accepted: October 18, 2015
published: December 16, 2015
Citation: Yildirim P. Association patterns in open data to explore ciprofloxacin adverse events. Appl Clin Inform 2015; 6: 728–747
http://dx.doi.org/10.4338/ACI-2015-06-RA-0076

1. Introduction

Nowadays, thanks to innovations in life sciences, huge amounts of open data are accessible on the internet and an increasing number of databases are being released [1]. Open data constitutes a novel and important concept for many scientific domains [2], while it additionally possesses a particularly great interest for biomedical research.

Drug use in particular is an important research topic in the biomedical field and clinical pharmacology work on the risk and benefit concerns of drugs [3]. Adverse drug events are defined as injuries "resulting from medical intervention related to a drug" and are the leading cause of iatrogenic harm to patients [4, 5]. National health systems in many countries have their own strategies for the post market surveillance of adverse events to assure drug safety [4].

The Adverse Event Reporting System (AERS) is an open information database created to support the U. S. Food and Drug Administration (FDA)'s post-marketing monitoring program for all approved drugs and therapeutic products. Its main aim is to collect and store safety reports for monitoring community health [6].

As expected, healthcare initiatives showed a great interest in the FDA AERS [7]. Naturally, when dealing with huge amounts of data as in this context, the participation of computer science and especially of data mining technologies becomes inevitable for drug development research and post-marketing surveillance [8]. Pharmacology also plays a key role in assessing the adverse effects and benefits of drugs. Recent years have in fact witnessed the development and application of various data mining algorithms, aiming primarily to track adverse drug events [9].

In this study, an analysis was conducted using the FDA's AERS database on the adverse reactions and outcomes associated with ciprofloxacin, i.e. one of the main medication for bacterial diseases. Bacterial diseases are generally accepted as a serious epidemiological problem in the world due to their high mortality and treatment costs. More precisely, we use FDA's AERS data in order to discover hidden associations by means of association rule mining, among the adverse events of ciprofloxacin and patient demographics (e.g. gender) as well as adverse event outcomes (e.g. death, hospitalization, disability, etc.).

It is known that some drugs are associated with gender or age-specific exposure, and there are some adverse events that are associated with a specific gender or age. Thus, one of our objectives in this study is to highlight whether adverse events are the same or different according to gender or age, because it is suspected that gender or age may be associated with differential risk to drugs. So, patient demographics are an effect modifier [10].

In particular, association rule mining is employed in order to detect significant and useful information in the event reports. Association rule mining is a well known method for discovering undetected relationships between variables in huge databases; and it has been already used in the context of adverse drug event studies. For instance, Harpaz et al. have employed this technique for discovering multi-item adverse drug event associations in FDA's AERS [11]. Moreover, even though large databases imply an increased search space and can lead to eventual performance issues for rule discovery methods, association rule mining can handle these problems simply and efficiently.

In the rest of this paper, after presenting briefly past and related work, the employed method is introduced in detail at methods and results are discussed later at discussion and last, conclusions is devoted to concluding remarks.

1.1 Data mining algorithms

In drug safety research, data mining algorithms have been developed to identify drug-associated adverse events as signals [10]. All of these algorithms extract decision rules for signal detection and/or calculate scores to measure the associations between drugs and adverse events from two by two frequency table of counts that involve the presence or absence of a particular drug and a particular event occurring in case reports (▶Table 1) [10]. For example, the proportional reporting ration (PRR), the reporting odds ratio (ROR), the information component (IC) and the empirical Bayes geometric mean (EBGM) are widely used. These algorithms, however, differ from one another in that the PRR and ROR are frequentist ones, whereas the IC and EBGM are Bayesian ones [12, 13]. The PRR is currently used by the UK Medicines and Healthcare products Regulatory Agency

(MHRA), the ROR by the Netherlands Pharmacovigilance Centre, the IC by the World Health Organization (WHO), and the EBGM by the FDA. When we use each statistical test, we can define the association between drug and adverse event. Using the PRR, an association is detected if the count of co-occurrences is three or more, and the PRR is two or more with an associated Chi-squared value of four or more.

The Chi-squared represents a summed normalized square deviation of the observed values from the corresponding expected values. Corresponding to the values of the Chi-squared and a degrees of freedom count (always 1, for boolean variables) is a p value. This value, between 0 and 1, indicates the probability of witnessing the observed counts were the variables really independent. If this value is low (say, less than 0.05), we reject the hypothesis that the variables are independent. We say a set of items is dependent at significance level α if the p value of the set is at most 1-α. To put the Chi-squared values in perspective, for a p value of 0.05 with one degree of freedom, the Chi-squared cutoff value is 3.84. Thus, any set of items with a Chi-squared value of 3.84 or more is significant at the 1−0.05=95% confidence level [14, 15].

For the ROR, a signal is detected if the lower bound of the 95% two-sided confidence interval (CI) of ROR exceeds one. Signal detection using IC is done using the IC025 metric, a criterion indicating the lower bound of the 95% two-sided confidence interval of the IC, and a signal is detected if the IC025 value exceeds zero [13]. The Multi-item Gamma Poisson Shrinker (MGPS) data mining algorithm implements empiric Bayesian models to screen for associations between the drug and previously unidentified adverse drug reactions. The MGPS calculates the EBGM values, which are the ratios of the observed to the expected number of drug-event pairs (reporting ratio). The MGPS adjusts for differences in reporting rates by variables in the dataset, e.g., age, gender and reporting quarter/year. This adjustment shrinks each reporting ratio towards one. This conservative approach considers EBGM values ≥2.0 to be the safety signal threshold. The EBGM values are reported with their 95% CI. Higher values denote a stronger association between the drug and the reported adverse drug reaction [16].

▶Table 2 shows common measures based on two by two table in disproportionately analysis. The RRR, when implemented within an empirical Bayesian framework, is known as EBGM. The IC is a logarithmic RRR metric that is implemented in a Bayesian framework [17].

Disproportional algorithms are widely used but have some drawbacks. For example, the accuracies of the methods for detecting adverse events depend, to a large extent, on the number of suspected drug reports. If we have few reports, the accuracy of adverse events detection is greatly reduced [18]. Therefore, we select association rule mining which explores for frequent itemsets (ciprofloxacin associated adverse events and patient demographics which frequently appear at the same time). In addition, most of traditional methods are not able to offer an approach to reveal age or gender specific adverse events. The MGPS algorithm is similar to our method, it stratifies the FDA AERS by some categories such as age or gender to adjust for differences in relative reporting ratios by these variables, but this method is also based on disproportional analysis like others and it has some limitations to detect frequent adverse event-patient demographics itemsets which are seen at the same time. Hence, this study provides a different approach which has not been previously suggested to analyze FDA AERS based on the categorization of age and gender groups and discover some hidden associations among these groups, the adverse events, and the outcomes of ciprofloxacin.

1.2 Related Work

Several studies have been performed for knowledge discovery on drug adverse event associations so far. Hoog et al. have analysed data on duloxetine-exposed pregnancies as created in both the LSS (Lilly Safety System) and the FDA's AERS databases. They have searched both normal and abnormal pregnancy outcomes as measures. They have used descriptive statistics for LSS data and have conducted a disproportionate analysis with the EBGM for the AERS data. Their study reveals that duloxetine related outcomes in pregnancy cases are frequently reported and the historical view of cases shows consistency with negative effects [19].

Furthermore, Kadoyama et al., have analysed FDA's AERS to explore the adverse event profile of tigecycline, i.e. a glycylcycline-class antibacterial. They have used official pharmacovigilance tools

including the proportional reporting ratio, the reporting odds ratio, the information component given by a Bayesian confidence propagation neural network, and the EBGM and have found several highly frequent adverse events including nausea, vomiting, pancreatitis, hepatic failure, hypoglycaemia, etc. [20].

In 2009, Hochberg et al., have performed a study on drug-versus-drug comparisons. They described the Pharmacoviligance Map approach, ranked adverse event rates in the FDA's AERS database and have compared their consistency against the results of published studies [7, 20].

Both known (e.g. vomiting, diarrhoea, nausea, etc.) and unknown adverse events of ciprofloxacin were investigated extensively. In particular, Haring et al. studied ciprofloxacin and the risk of cardiac arrhythmias. They reported a case of torsades de pointes (Tdp) associated with oral ciprofloxacin in a patient with concomitant risk factors and provided a literature review on pharmacokinetic interactions involving ciprofloxacin related cardiac arrhythmias [21]. In addition, Moffet et al. investigated possible digoxin toxicity associated with concomitant ciprofloxacin therapy with a 27 year old female patient. They reported that she presented to clinic with nausea and anorexia within a few days of the addition of ciprofloxacin to her current regimen of medications, which included digoxin. They concluded that the proposed cause of the nausea and anorexia was digoxin toxicity secondary to a drug-drug interaction with ciprofloxacin [22].

The Observational Medical Outcomes Partnership (OMOP) is a large research initiative that aims to identify the most reliable methods for analyzing large volumes of electronic healthcare data for drug safety surveillance. As part of this ongoing effort, the OMOP investigators have compiled and recently made public an extensive gold standard with which they evaluate their portfolio of methods. The gold standard consists of a total 398 positive and negative test cases, which have been validated to the best of existing knowledge. Each test case represents a drug-event pair. Harpaz et al., carried out a performance comparison of commonly known signal detection algorithms such as the MGPS, PRR and ROR and used the OMOP gold standard in an effort to systematically evaluate and gain a better understanding of the diagnostic potential and operational characteristics of signal detection algorithms that are routinely applied to FAERS [23]. According to their results, MGPS achieved the best performance, with PRR and ROR producing near equivalent performance.

2. Methods

2.1 Data sources

Input data for our study was collected from the website of the FDA's AERS database, spanning the period from the third quarter of 2005 to the last of 2013 [9]. The FDA's AERS is a relational database and complies with the international safety reporting guidance (ICH E2B) issued by the International Conference on Harmonization. The FDA's AERS contains file-tables which can be connected through specific data fields and are packaged in quarterly periods.

All data files are organised in SGML and ASCII data formats and can be processed by well-known relational database management systems such as ORACLE, Microsoft SQL Server and MySQL.

Each quarterly file package contains the following seven data files:

- DEMO: contains demographic data such as "event date", "patient age" and "reporter country";
- DRUG: contains information for drugs such as "primary suspect drug" (PS) and "secondary suspect drug" (SS);
- REACTION: contains all adverse drug reactions described by the MedDRA (Medical Dictionary for Regulatory Activities) terminology system;
- OUTCOME: contains type of outcome, such as disability and death;
- RPSR: contains information on the source of the reports;
- THERAPY: includes information about drug therapy;
- INDICATIONS: contains all MedDRA terms coded for the indications of diagnoses.

Each file can be converted into a specific database table and connected with each other using the "ISR number" (Individual Safety Report) as a primary key field. The ISR number consists of seven

digits and identifies uniquely an AERS report, while providing a common connection between all data files [3, 9].

2.2 Association Rule Mining and the Apriori Algorithm

Frequent sets play an important role in data mining, for detecting interesting patterns within databases, e.g. association rules, correlations and sequences. Association rule mining in particular explores hidden associations among large sets of data items. Given the huge amounts of data being continuously collected and stored, the interest in as well as need of various industries of the aforementioned methods has been increasing dramatically. Indeed, the discovery of interesting association relationships within huge amounts of business data can potentially aid in various business decisions making processes [24].

Association rule mining methods attempt to search frequent items in databases. Given a set of transactions T (where each transaction is a set of items), an association rule can be shown in the form X→Y, where X and Y are mutually exclusive sets of items [25].

The rule's statistical significance is measured by support and the rule's strength by confidence. The support of an itemset expresses how often the itemset appears in a single transaction in the dataset [25]. The support is measured as:

$$\text{Support} = \frac{P(X \cup Y)}{N}$$

The confidence of the association rule is the ratio of the support of the itemset $X \cup Y$ to the support of the itemset X, which roughly corresponds to the conditional probability P (Y/X). The formula of confidence is as follows:

$$\text{Confidence} = \frac{Supp(X \cup Y)}{Supp(X)}$$

Association rule generation is usually split up into two steps:

- Finding all combinations of items whose supports are greater than a user-specified minimum support (i.e. threshold). These combinations are called frequent itemsets.
- Using the items from frequent itemsets to generate the desired rules. More specifically, the confidence of each rule is computed, and if it is above the confidence threshold, the rule is satisfied [25, 26].

The Apriori algorithm is one of the most efficient algorithms for discovering association rules in large databases. The pseudo code for the Apriori algorithm is as follows:

C_k=Candidate itemset of size k
L_k=Frequent itemset of size k
L_1={Frequent items};
for (k=1; L_k!=Ø; k++) **do begin**
C_{k+1}=candidates generated from L_k
for each transaction t in database **do**
#increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1}=candidates in C_{k+1} with minSupport
end
Return Lk;

And now the next section shall focus on the application of the Apriori algorithm to the discovery of association rules within the FDA's AERS database using ciprofloxacin records.

3. Results

The FDA's AERS datasets were downloaded from the FDA's web site. The data in ASCII format was imported into a database using Microsoft SQL Server 2012. Then, ciprofloxacin related records were

selected to create a dataset for association analysis. In total, 6531 adverse event reports for ciprofloxacin were collected from the whole database.

Normalization is an important issue for data mining research because some entities can be expressed by using their synonyms and brand names. These names should be normalized. In this study, ciprofloxacin has also some variations. For example, ciflox and cipro are the variations of ciprofloxacin. These variations were found by searching the Drugbank database and mapped to one specific name [27]. After normalizing, the dataset was created and it contains patient demographics such as age, gender, adverse event outcomes and adverse events and then age was categorized.

▶Table 3 shows data summary of adverse event reports for ciprofloxacin. According to the table, females 3585 (55%) have more adverse event reports than males 2884 (44%). Preschool child 1216 (19%), middle aged 2217(34%) and aged 2097 (32%) patients can be more likely than others (e.g., young and adult groups) to have adverse drug reactions (▶Figure 1). Based on the number of co-occurrences, pyrexia has the highest frequency with ciprofloxacin, followed by pain, diarrhoea, anxiety and nausea in this order (▶Table 3; ▶Figure 2).

Adverse event reactions lead to HO (Hospitalization) 2135(33%), followed by OT (Other) 1949 (30%), DE (Death) 1014 (16%), LT (Life-Threatening) 615 (9%), DS (Disability) 548 (8%), RI (Required Intervention to prevent permanent impairment/damage) 254(4%) and CA (Congenital Anomaly) 16 (0%) sequentially (▶Table 3).

▶Figure 3 shows gender specific distribution of all adverse events. According to this figure, pyrexia, diarrhoea, pneumonia and renal failure acute have the highest frequency with male patients. On the other hands, the majority of the adverse events in female patients are pain, anxiety, pyrexia and nausea.

Top ten adverse events and outcomes for each age group were also listed in ▶Table 4–6. These tables reveal some interesting associations. For example, diarrhoea is the most seen event in child patients, anxiety has the possibility of some strong associations with middle-aged patients and renal failure acute has the highest number of reports with aged patients. In addition, hospitalization is seen as the most frequent adverse event outcome for all of the age groups.

The Apriori algorithm was used to perform association analysis on the dataset. The WEKA 3.6.6 software was used. WEKA contains many machine learning algorithms for data mining tasks and is open source software. It includes tools for data pre-processing, classification, clustering, association rules and visualisation [28, 29]. The application of the Apriori algorithm on the dataset generated many rules (▶Table 7).

After detecting potential associations between ciprofloxacin and adverse events, we investigated the FDA label information for the drug [30]. Tendinopathy induced by fluoroquinolone (FQ) antibiotics is a topic of controversy, with many researchers believing in a direct causal relationship while others believing that the risk is negligible. In a World Health Organization (WHO) survey in Australia of tendon disorders associated with FQ use, ciprofloxacin was found to be the causal agent in 90 percent of cases, with the risk of tendinopathy appearing to be dose independent. As of July 2008, the FDA mandated that all FQ products have a black-box warning indicating an increased risk in adverse events including tendon rupture [31].

Comparing the label information, our study have some drawbacks, for example, tendon disorder is given a black-boxed warning in the FDA website but it was not seen as a suspected signal in our results. There may be some reasons why we could not detect this adverse event. First, we selected the adverse events which have the highest number of reports in the FDA AERS and used first twenty of them for analysis. The number of tendon disorder related reports may be smaller than others and may be ignored. Further, the number of rules may affect the discovery of patterns. For example, if association rule mining algorithms generate a small number of rules, some associations may not be detected. Therefore, Apriori algorithm used in our study may not produce enough number of rules to discover tendon disorder related associations. However, pain and pain in extremity may be associated with tendon disorders and we detected some relationships among these events and patient demographics.

We also compared our results against widely used data mining algorithms. OpenVigil tool was used. OpenVigil contains software packages to analyse pharmacovigilance data which includes disproportionality analyses like Proportional Reporting Ratio (PRR). ▶Table 8 listed the scores of these algorithms which included the Chi-squared, RRR, PRR and ROR [8]. According to the table,

the ROR based signals were pain, anxiety, renal failure acute, pain in extremity, arthralgia, depression, urinary tract infection, abdominal pain, anaemia and renal failure (95% CI>1). On the other hand, considering the PRR, only an association with urinary tract infection was suggested (PRR≥2 and Chi-squared≥4). As seen in the table, the PRR, RRR and ROR methods cannot uncover age or gender specific adverse event associations, but in this study, the Apriori algorithm revealed some potential relationships among the adverse events and particular age and gender groups. ▶Table 9 also summarizes the comparison of the results of Apriori algorithm and traditional methods in terms of threshold criteria and the number of the drug associated event detection.

3.1 Selection of Meaningful Association Rules

The Apriori algorithm first finds itemsets that meet a minimum support criterion. It then uses these itemsets to generate rules that conform to a minimum confidence criterion. The algorithm can generate a large amount of rules. However, some of the rules can be redundant. This is a drawback for the Apriori algorithm. Usually, data mining tools enable users to determine thresholds for the rules, but afterwards, there is no best approach for selecting effective rules. During the evaluation of the rules, users may deal with huge amount of rules and search some strategies to find good ones [32, 33].

There are two main techniques to evaluate the rules: objective and subjective measures of interest. The objective approach uses the statistical strength or characteristics of the patterns to assess their degree of interestingness. The measures effectively quantify the correlation between the antecedents and the consequent. Subjective techniques incorporate the subjective knowledge of the user into the assessment strategy [33, 34]. In this study, apart from confidence measure, the interestingness of resulting rules was evaluated according to three objective measures: confidence, lift and conviction (▶Figure 4).

Confidence was introduced in methods part in the paper. Lift represents the ratio of probability. Given a rule, X and Y occur together to the multiple of the two individual probabilities for X and Y; that is,

$$\text{Lift} = \frac{Supp(X \cup Y)}{Supp(X) * Supp(Y)}$$

Conviction is another way for measuring the objective interestingness. Conviction compares the probability that X appears without Y if they were dependent from the actual frequency of the appearance of X without Y [27, 28]. Conviction is measured as:

$$\text{Conviction} = \frac{1 - Supp(Y)}{1 - Conf(X \Rightarrow Y)}$$

The interestingness factor (IF) as the lift and conviction can be evaluated as follows:

- IF (X, Y) =1, if X and Y are independent,
- IF (X, Y) >1, if X and Y are positively correlated,
- IF (X, Y) <1, if X and Y are negatively correlated.

Considering both objective and subjective approaches, we assume that a rule which offers as much information as possible can be interesting and useful. As a result, we focused on meaningful and interesting rules with high confidence, lift and conviction values to reveal association patterns in the dataset.

We consider that the confidence is more than 40% and the support is more than 1%, then some association rules were extracted from a large number of rules based on lift and conviction measures. The lift and the conviction signify high interestingness when the measures are greater than one [35].

The statistical significance of the association rule can be estimated by using the Chi-squared analysis. The Chi-squared statistics is defined in terms of the confidence, support and lift of the single rule . We calculated the Chi-squared values to evaluate the association rules [14]:

$$\text{Chi-squared} = n(lift\text{-}1)^2 \frac{Supp * Conf}{(Conf - Supp)(Lift - Conf)}$$

▶Table 7 shows the associations among adverse events, gender and adverse event outcomes for ciprofloxacin. However, considering the Chi-squared analysis results, only rule 1, 2, 3, 7, 9 and 10 indicate the statistical significance of the associations in these associations.

Some events such as, urinary tract infection, pain in extremity and arthralgia are associated with female patients. For example, rule five has high confidence, lift and conviction with 0.62, 1.12 and 1.16 values in the table. Since the lift and the conviction measure have positive correlation when it is greater than one, we can conclude that a positive correlation exists in urinary tract infection and female patients.

Anxiety is an uncommon event for ciprofloxacin. However, it has been discovered that the drug might cause psychiatric disorders in female and middle aged male patients. Rule one and eight highlight positive correlation between psychiatric problems and female and middle aged male patients with lift and conviction values which are greater than one. Furthermore, rule eight reveals that anxiety in female patients can also cause hospitalization outcome.

According to rule two, seven and nine, some adverse events such as renal failure acute and renal failure have some associations with aged and female patients. For example, rule two concerning has a 44% confidence, 1.36 lift and 1.19 conviction values. Related rules such as rule seven and nine have 42% and 41% confidence, 1.3 and 1.27 lift and 1.16 and1.14 conviction values respectively.

Ciprofloxacin is related to pyrexia, hypotension and pneumonia in male patients. For example, rule three concerning the possibility of pyrexia with middle aged male patients has 54% confidence, 1.22 lift and 1.19 conviction values. Rule four has 53% confidence and 1.2 lift and 1.18 conviction values in order. Furthermore, rule eleven shows a relationship between pneumonia and aged male patients with 51% confidence, 1.16 lift and 1.13 conviction values.

3.2 Evaluation of the Performance of Association Rule Mining

Association rule mining is one of the most important tasks in data mining and various effective algorithms have been proposed in the state of the art [36].

The Apriori algorithm generates candidates, and discovers frequent itemsets, by exploiting user-specified support and confidence measures. In the huge quantity of itemsets, the algorithm needs more space and time and consequently the complexity of the algorithm increases [37].

The FP-Growth algorithm was proposed as an alternative to the Apriori-based approach by Han [38, 39]. The FP-Growth algorithm stands for frequent pattern. It generates frequent itemsets without using candidate generation. FP-Growth adopts the divide and conquer strategy. The algorithm encodes the data set using a compact data structure called a FP-tree and then extracts frequent itemsets directly from this structure. It is based on a prefix tree representation of the given database of transactions, which can save considerable amounts of memory for storing the transactions. The basic idea of the FP-Growth algorithm consists in constructing a FP-tree for all the transactions. As a result, every path of the FP-tree represents a frequent itemset and the nodes in the path are stored in decreasing order with respect to the frequency of the corresponding items [39, 40].

The performance evaluation of the following association rule mining algorithms was conducted between Apriori and FP-Growth by execution time. In addition, these algorithms were compared by association rules which have higher confidence, lift and conviction values to detect potential drug signal detection.

4. Discussion

Our study has some limitations, for example, pneumonia is commonly treated with ciprofloxacin and it might not be considered an adverse event. In addition, pyrexia (fever) is a common complication symptom of infectious diseases. Consequently, this causes a problem for data mining studies on adverse drug events. In fact, the evaluation of data mining studies is a rather general problem, since there is no accepted gold standard and there are limited studies for evaluating the accuracy of adverse drug event monitoring. Despite these inconveniences, the rules that have been discovered by the approach under consideration are novel and of significant interest.

▶Table 10 highlighted that the Apriori algorithm took a short time to generate rules and therefore the performance of this algorithm is superior to other. Association rule mining algorithms can generate many rules, but a lot of rules may bring some drawbacks. For example, all of the rules may not be beneficial to researchers. In addition, a huge number of rules need more memory and execution time. On the other hand, sometimes more rules provide more advantages and we can detect unexpected adverse events such as tendon disorders to discover ciprofloxacin related adverse events. Therefore, clinicians and researchers can validate our rules and perform clinical trials to find new ideas for the evaluation of ciprofloxacin's drug safety.

Besides, our study aimed to discover adverse events for a single drug (i.e. ciprofloxacin). However, multi-item adverse drug event associations can be considered as an important research direction targeting the detection of multiple adverse events. We plan to explore multi-item adverse event drug screening in the future.

Despite the significance of FDA's AERS database as a resource, it possesses limitations as well. Specifically, as often encountered with large medical databases, the FDA's AERS suffers from the missing data problem [41]. In our study, we removed some data containing missing values. In addition, we faced some data quality and compatibility problems with the datasets created in different time periods and then we merged the datasets which cover the third quarter of 2005 through the last of 2013.

Apart from the FDA's AERS database, electronic medical records created in the hospital information systems or health claims data may be an important resource to track drug adverse events and their outcomes. The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient care [42].These records have higher dimensional and huge amounts of data and they can be analysed by data mining techniques to detect associations between drug related adverse events and patient demographics. Therefore, electronic medical records can be exploited in order to reveal serious drug risks [43].

5. Conclusions

The FDA takes responsibility both for approving drugs for entering the market and for tracking their safety after entering the market. This activity is performed by the FDA AERS. The FDA AERS offers open data sources and tools to identify new or unknown adverse events of drugs and this database is a significant resource for knowledge discovery in the biomedical domain. In this study, the FDA AERS for the ciprofloxacin drug was considered and a research study based on patient demographics, adverse events and adverse event outcome relationships in the AERS reports was performed. Ciprofloxacin is one of the main drugs for bacterial diseases. Therefore, the treatment and management of these diseases has a great interest for biomedical researchers.

We used database and computational methods and highlighted some statistics and we reported the most frequent adverse events and event outcomes of the drug under consideration, then applied the Apriori algorithm to the dataset to obtain some rules and used objective measures to select interesting ones. We searched whether adverse events are the same or different according to gender or age, because it is suspected that gender or age may be associated with differential risk to drugs. It was discovered that patient gender and age can have some relationships with some adverse events and event outcomes of ciprofloxacin. For example, psychiatric disorders are not common events, but our results show that the drug might cause psychiatric disorders in female and middle aged male patients. We also investigated the FDA drug label information and compared our results against both the drug label and the results of traditional methods. We observed some differences among these approaches and discussed the reasons of the differences. Our experiment also revealed some limitations and challenges that there is in fact a lack of methodological standards for post marketing surveillance. Future work will include considering these issues and multi-drug adverse events. Despite some limitations, our study highlights that the FDA AERS is an important resource for post-marketing drug safety studies and computational methods and algorithms provide an important help for researchers to analyze the FDA AERS. In conclusion, we hope that researchers and clinicians can utilize our results and this study can contribute to the research on the evaluation of ciprofloxacin's drug safety.

Clinical Relevance Statement
Patient demographics can affect the action of drugs. This study presents an approach for knowledge discovery on the adverse events of ciprofloxacin and highlights some associations between patient demographics and reactions as well as outcomes. Clinicians and pharmaceutical researchers can validate our results and perform clinical trials to find new ideas for the assessment of ciprofloxacin's drug safety. Clinicians should adjust their benefit/risk rating of ciprofloxacin.

Conflict of Interest
The author has no conflicts of interest to declare.

Protection of Human and Animal Subjects
No human subjects were involved in this research.

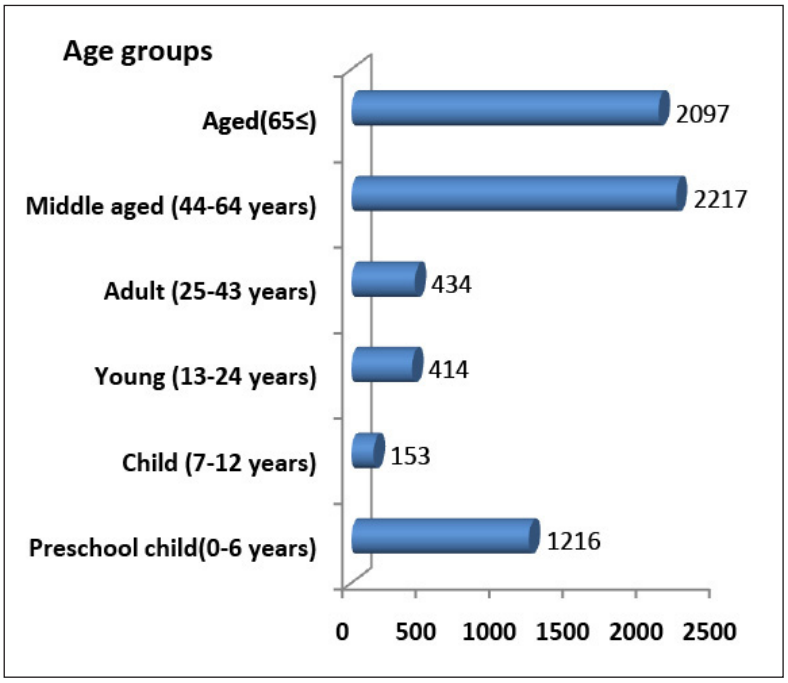

Fig. 1 Age-specific distribution of adverse event reports

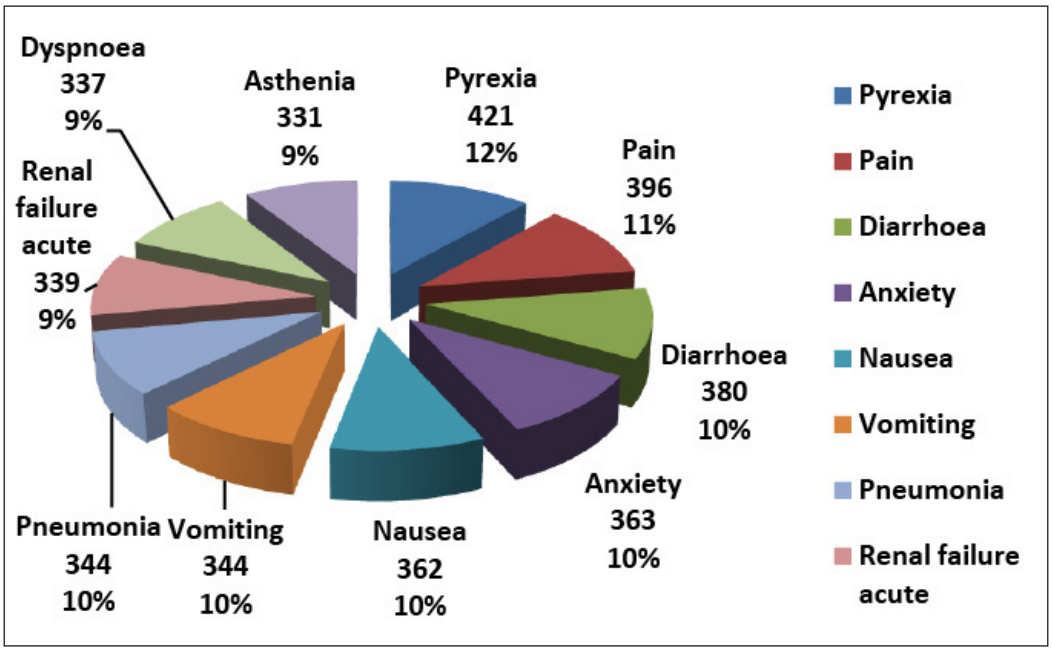

Fig. 2 The distribution of most frequent ten adverse events

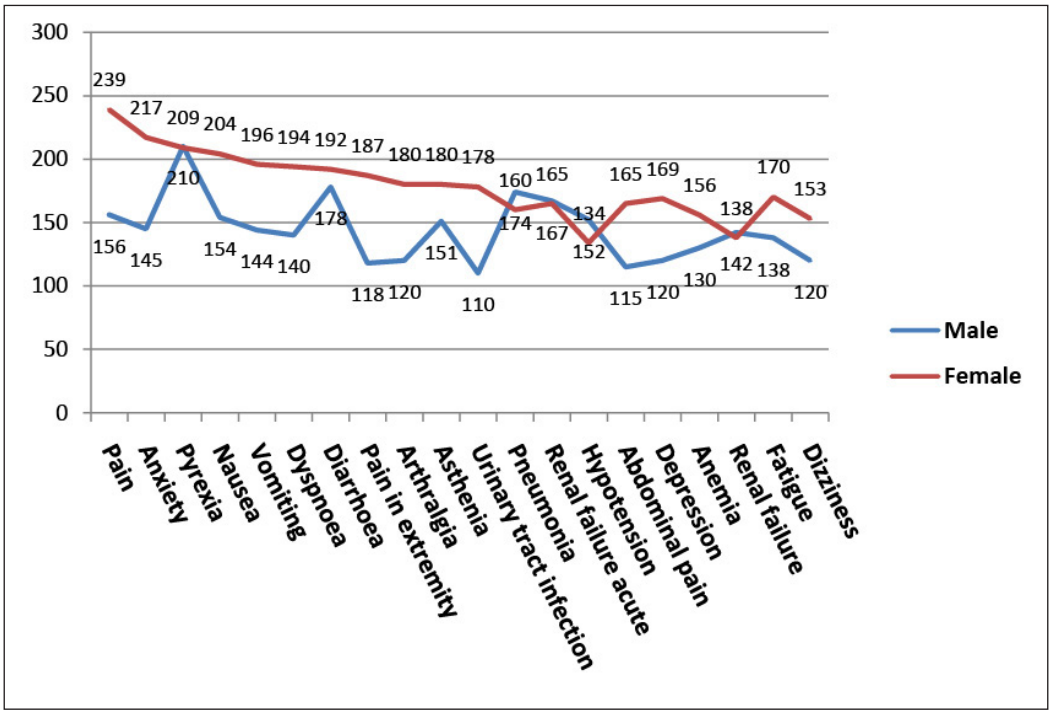

Fig. 3 Gender specific distribution of all adverse events

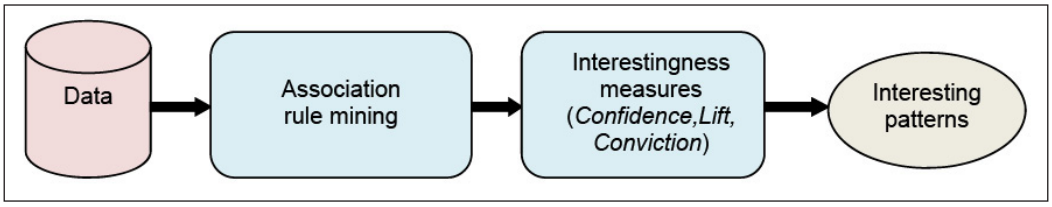

Fig. 4 Interestingness measures in the association rule mining process

Table 1 Two by two frequency table

	Adverse event of interest	All other adverse events	Total
Drug of interest	a	b	a+b
All other drugs	c	d	c+d
Total	a+c	b+d	a+b+c+d

Table 2 Common measures based on two by two table in disproportionately analysis

Measure	Formula	Probabilistic interpretation
Relative reporting ratio (RRR)	$\frac{a(a-b-c-d)}{(a-c)(a+b)}$	$\frac{P(ae \mid drug)}{P(ae)}$
Proportional reporting rate ratio (PRR)	$\frac{a(c+d)}{c(a-b)}$	$\frac{P(ae \mid drug)}{P(ae \mid -drug)}$
Reporting odds ratio (ROR)	$\frac{ad}{cb}$	$\frac{P(ae \mid drug)P(-ae \mid -drug)}{P(-ae \mid drug)P(ae \mid -drug)}$
Information component (IC)	$\log_2 \frac{a(a-b-c-d)}{(a+c)(a-d)}$	$\frac{\log_2 P(ae \mid drug)}{P(ae)}$
Chi-squared (X^2)	$\sum_{i=1}^{n} \frac{(Observed-Expected)^2}{Expected}$	

ae: adverse event

Table 3 Data summary of adverse event reports for ciprofloxacin in FDA AERS

Attribute	Type	(N=6 531 reports)	
Age	**Numeric (Mean=54.59, StdDev=19.71)**		
	Categorized age groups	**N**	**(%)**
	Preschool child (0–6 years)	1 216	19
	Child (7–12 years)	153	2
	Young (13–24 years)	414	6
	Adult (25–43 years)	434	7
	Middle aged (44–64 years)	2 217	34
	Aged (≥65 years)	2 097	32
Gender	**Nominal**	**N**	**%**
	Male	2 884	44
	Female	3 585	55
	NULL	54	1
	UKN (unknown)	2	0
	NS (Not Specified)	6	0

Table 3 Continued

Attribute	Type	(N=6 531 reports)	
Adverse event outcome	**Nominal**	**N**	**%**
	HO-Hospitalization	2 135	33
	OT-Other	1 949	30
	DE-Death	1 014	16
	LT-Life Threatening	615	9
	DS-Disability	548	8
	RI-Required Intervention to prevent permanent impairment/damage	254	4
	CA-Congenital Anomaly	16	0
Adverse event	**Nominal**	**N**	**%**
	Pyrexia	421	6
	Pain	396	6
	Diarrhoea	380	6
	Anxiety	363	6
	Nausea	362	6
	Vomiting	344	5
	Pneumonia	344	5
	Renal failure acute	339	5
	Dyspnoea	337	5
	Asthenia	331	5
	Fatigue	309	5
	Pain in extremity	305	5
	Arthralgia	301	5
	Depression	289	4
	Urinary tract infection	289	4
	Abdominal pain	287	4
	Anaemia	287	4
	Renal failure	286	4
	Hypotension	286	4
	Dizziness	275	4

Table 4 Top ten adverse events and total outcomes for preschool child and child

Adverse event	N	Adverse event outcome	N
Preschool child (0–6 years)			
Pyrexia	90	HO	407
Anxiety	82	OT	406
Pain	81	DE	130
Diarrhoea	77	DS	112
Nausea	73	RI	57
Pain in extremity	66	CA	4
Depression	62		
Abdominal pain	62		
Dizziness	55		
Anaemia	53		
Child (7–12 years)			
Diarrhoea	18	HO	61
Pyrexia	15	OT	48
Pneumonia	13	DE	34
Renal failure acute	13	LT	9
Anaemia	10	RI	1
Fatigue	10		
Pain in extremity	9		
Vomiting	9		
Asthenia	8		
Nausea	7		

Table 5 Top ten adverse events and total outcomes for young and adult

Adverse event	N	Adverse event outcome	N
Young (13–24 years)			
Pyrexia	46	HO	158
Vomiting	39	OT	155
Nausea	33	LT	30
Diarrhoea	27	RI	29
Pain	27	DE	27
Anxiety	26	DS	15
Dyspnoea	24		
Abdominal pain	21		
Depression	21		
Asthenia	21		

Table 5 Continued

Adverse event	N	Adverse event outcome	N
Adult (25–43 years)			
Pain	37	HO	151
Pyrexia	32	OT	146
Anxiety	30	DS	38
Arthralgia	30	LT	37
Pain in extremity	29	DE	31
Depression	27	RI	30
Diarrhoea	25	CA	1
Vomiting	24		
Asthenia	22		
Urinary tract infection	22		

Table 6 Top ten adverse events and total outcomes for middle aged and aged

Adverse event	N	Adverse event outcome	N
Middle aged (44–64 years)			
Anxiety	134	HO	676
Pain	130	OT	647
Fatigue	129	DE	339
Dyspnoea	124	LT	223
Pneumonia	123	DS	212
Nausea	123	RI	109
Pyrexia	121	CA	11
Asthenia	119		
Arthralgia	116		
Renal failure acute	112		
Aged (≥65 years)			
Renal failure acute	138	HO	682
Pneumonia	135	OT	547
Diarrhoea	122	DE	453
Renal failure	119	LT	216
Dyspnoea	117	DS	171
Pyrexia	117	RI	28
Asthenia	116		
Pain	116		
Vomiting	110		
Anemia	108		

Table 7 Some association rules (MinSupport=1% and MinConfidence=40%)

No	Antecedent	Consequent	Confidence	Lift	Conviction	Chi-squared
1	Gender=Male and Adverse event=Anxiety	Age=Middle aged	0.46	1.34	1.2	**6.8***
2	Gender=Female and Adverse event=Renal failure acute	Age=Aged	0.44	1.36	1.19	**8.5***
3	Age=Middle aged And Adverse event=Pyrexia	Gender=Male	0.54	1.22	1.19	**4.8***
4	Adverse event=Hypotension	Gender=Male	0.53	1.2	1.18	3.9
5	Adverse event=Urinary tract infection	Gender=Female	0.62	1.12	1.16	1.9
6	Adverse event=Pain in extremity	Gender=Female	0.61	1.12	1.16	1.9
7	Adverse event=Renal failure	Age=Aged	0.42	1.3	1.16	**6.8***
8	Adverse event outcome=HO and Adverse event=Anxiety	Gender=Female	0.62	1.12	1.15	0.19
9	Adverse event=Renal failure acute	Age=Aged	0.41	1.27	1.14	**125.3***
10	Adverse event=Pain	Gender=Female	0.6	1.1	1.13	**132.8***
11	Age=Aged and Adverse event=Pneumonia	Gender=Male	0.51	1.16	1.13	2.6
12	Adverse event=Arthralgia	Gender=Female	0.6	1.09	1.11	0.16
13	Outcome=RI	Gender=Female	0.6	1.09	1.11	0.16

*: Statistically significant

Table 8 The results of classical data mining algorithms for ciprofloxacin-associated adverse events

Adverse event	N	Chi-squared	RRR	PRR	ROR (95% CI)
Pyrexia	421	0.004	0.997	0.997	0.997 (0.924; 1.076)
Pain	396	82.564	1.297	1.299	**1.302* (1.23; 1.379)**
Diarrhoea	380	0.205	0.983	0.983	0.983 (0.916; 1.056)
Anxiety	363	254.463	1.633	1.638	**1.654* (1.546; 1.749)**
Nausea	362	97.421	0.719	0.718	0.715 (0.669 ; 0.764)
Vomiting	344	38.696	0.783	0.782	0.78 (0.722; 0.844)
Pneumonia	344	4.488	0.905	0.905	0.904 (0.825; 0.992)
Renal failure acute	339	2.984	1.103	1.104	**1.104* (0.989; 1.233)**
Dyspnoea	337	15.031	0.873	0.872	0.871 (0.813; 0.934)
Asthenia	331	1.969	1.056	1.056	1.057 (0.979; 1.14)
Pain in extremity	305	75.756	1.394	1.397	**1.4* (1.297; 1.51)**
Fatigue	309	31.864	0.816	0.815	0.813 (0.757; 0.874)
Arthralgia	301	249.266	1.687	1.693	**1.699* (1.59; 1.816)**
Depression	289	0.09	1.013	1.013	**1.013* (0.934; 1.099)**
Urinary tract infection	289	**643.267**	**2.557***	**2.557***	**2.588* (2.398; 2.793)**
Abdominal pain	287	54.68	1.357	1.359	**1.361* (1.254; 1.478)**
Anaemia	287	168.161	1.631	1.636	**1.641* (1.521; 1.769)**

Table 8 Continued

Adverse event	N	Chi-squared	RRR	PRR	ROR (95% CI)
Renal failure	286	0	1.001	1.001	**1.001* (0.896; 1.117)**
Hypotension	286	1.021	0.945	0.944	0.944 (0.847; 1.052)
Dizziness	275	66.299	0.715	0.714	0.712 (0.656; 0.773)

*: drug associated adverse event detected
N: number of co-occurrences

Table 9 Comparison of the results of Apriori algorithm and classical data mining algorithms

Algorithm	Treshold criteria	The number of drug associated signal detection
Apriori algorithm	MinConfidence ≥40, Lift ≥1 and Conviction ≥1	13
PRR	PRR >2 and Chi-squared >4	1
ROR	95% CI >1	10

Table 10 Performance comparison of Apriori and FP-Growth algorithms

Algorithm	Execution time (seconds)	The number of interesting association rules
Apriori	0.6	13
FP-Growth	3	9

References

1. Hrynaszkiewicz I. The need and drive for open data in biomedical publishing. Serials 2011; 24(1): 31–37.
2. Boulton G, Rawlins M, Vallance P, Walport M. Science as a public enterprise: the case for open data. Lancet 2011; 377: 1633–1635
3. Poluzzi E, Raschi E, Piccinni C, De Ponti F. Analysis of the Publicly Accessible FDA Adverse Event Reporting System(AERS). Data Mining Applications in Engineering and Medicine 2012. http://www.intechopen.com/books/data-mining-applications-in-engineering-and-medicine/data-mining-techniques-in-pharmacovigilance-analysis-of-the-publicly-accessible-fda-adverse-event-re
4. Johnson KB, Lehmann CU, Council on Clinical Information Technology of the American Academy of Pediatrics. Electronic prescribing in pediatrics: toward safer and more effective medication management. Pediatrics 2013; 131(4): 824–826.
5. Institute of Medicine, Committee on Quality in Healthcare in America. To Err is Human: building a Safer Health System. Washington DC: National Academies Press 1999.
6. Walport M, Brest P. Sharing research data to improve public health. Lancet 2011; 377: 537–539.
7. U.S. Food and Drug Administration. Postmarketing surveillance programs, 2009. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/ucm090385.htm.
8. Sakaeda T, Tamon A, Kadoyama K, Okuno, Y. Data Mining of the Public Version of the FDA Adverse Event Reporting System. Int J Med Sci 2013; 10(7): 796–803.
9. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System. Clin Pharmacol Ther 2013; 93(6): 539–546.
10. O'Neill RT, Szarfman A. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. American Statistician 1999; 53(3): 190–196.
11. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics 2010; 11(Suppl 9): S7.
12. Hauben M, Horn S, Reich L, Younus M. Association between gastric acid suppressants and clostridium difficile colitis and community-acquired pneumonia: analysis using pharmacovigilance tools. Int J Infect Dis 2007; 11(5): 417–422.
13. Tamura T, Sakaeda T, Kadoyama K, Okuno Y. Omeprazole and Esomeprazole-associated Hypomagnesaemia: Data Mining of the Public Version of the FDA Adverse Event Reporting System. Int J Med Sci 2012; 9(5): 322–326.
14. Alvarez SA. Chi-squared computation for association rules: preliminary results. Technical Report, BC-CS-2003–01 July 2003.
15. Silverstein C, Brin S, Motwani R. Beyond Market Baskets: Generalizing Association Rules to Dependence rules. Data Mining and Knowledge Discovery 1998; 2: 39–68.
16. Ali AK. Pharmacovigilance analysis of adverse event reports for aliskiren hemifumarate, a first-in-class direct renin inhibitor. Ther Clin Risk Manag 2011; 7: 337–344.
17. Hauben M, Bate A. Decision support methods for the detection of adverse events in post marketing data. Drug Discov Today 2009; 14 (7–8): 343–357.
18. Wang C, Guo XJ, Xu JF, Wu C, Sun YL, Ye XF, Qian W, Ma XQ, Du WM, He J. Exploration of the association rules mining technique for the signal detection of adverse drug events in spontaneous reporting systems. PLoS One 2012; 7(7): e40561.
19. Hoog SL, Cheng Y, Elpers J. Duloxetine and Pregnancy Outcomes: Safety Surveillance Findings. Int J Med Sci 2013; 10(4): 413–419.
20. Kadoyama K, Sakaeda T, Tamon A, Okuno Y. Adverse Event Profile of Tigecycline: Data Mining of the Public Version of the U.S. Food and Drug Administration Adverse Event Reporting System. Biol Pharm Bull 2012; 35(6): 967–970.
21. Haring B, Bauer W. Ciprofloxacin and the risk for cardiac arrhythmias: culprit delicti or watching bystander? Acta Cardiol 2012; 67(3): 351–354.
22. Moffett BS, Valdes SO, Kim JJ. Possible digoxin toxicity associated with concomitant ciprofloxacin therapy. Int J Clin Pharm 2013; 35(5): 673–676.
23. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of Pharmacovigilance signal detection algorithms for the FDA Adverse Event Reporting System. Clin Pharmacol Ther 2013; 93(6): 1–20.
24. Han J, Kamber M, Pei J. Data mining: concepts and techniques. Oxford: Elsevier Ltd 2011.
25. Zhu AL, Li J, Leong TY. Automated Knowledge Extraction for Decision Model Construction. A Data Mining Approach. AMIA Annu Symp Proc 2003; 2003: 758–776.

26. Deora CS, Arora S, Makani Z. Comparison of interestingness measures: support-confidence framework versus lift-irule framework. IJERA 2013; 3(2): 208–215.
27. Drugbank. http://www.drugbank.ca
28. Hall M, Frank E, Holmes G, Pfahringe B, Reutemann P, Witten IE. The WEKA data mining software: an update. SIGKDD Explorations 2009; 11(1): 10–18.
29. WEKA: Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka
30. Medication guide Cipro® 2008. http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/019537s70198474419857512078028214732 5L.pdf
31. Kim GK. The risk of Fluoroquinolone-induced Tendinopathy and Tendon Rupture: What Does The Clinician Need to Know? J Clin Aesthet Dermatol 2010; 3(4): 49–54.
32. McGarry K. A survey of interestingness measures for knowledge discovery. The Knowledge Engineering Review 2005; 20(1): 39–61.
33. Lee DG, Ryu KS, Bashir, M, Bae JW, Ryu KH. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J Med Syst 2013; 37(2): 9896.
34. Tang J, Chuang LY, Hsi E, Lin YD, Yang CH, Chang HW. Identifying the association rules between clinicopathologic factors and higher survival performance in operation centric oral cancer patients using the apriori algorithm. BioMed Res Int 2013; 2013: 359634.
35. Sheikh LM, Tanveer B, Hamdani MA. Interesting Measures for Mining Association Rules. Multitopic Conference, Proceedings of INMIC 2004. 8th International 2004: 641–644.
36. Kaimal LB, Metkar AR, Rakesh G. Self Learning Real Time Expert System. IJSCAI 2014; 3(2): 13–25.
37. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM-SIGMOD Record 2000; 29(2): 1–12.
38. Said AM, Dominic D, Abdullah AB. A Comparative Study of FP-growth Variations. IJCSNS 2009; 9(5): 266–272.
39. Borgelt A. An Implementation of the FP-Growth Algorithm. OSDM'05, August 21, 2005, Chicago, IL, United States.
40. Wu B, Zhang D, Lan Q, Zheng J. An Efficient Frequent Patterns Mining Algorithm Based on Apriori Algorithm and the FP-tree Structure. Third International Conference on convergence and Hybrid Information Technology 2008, Washington DC, United States.
41. Pearson RK. The problem of disguised missing data. SIGKDD Explorations 2006; 8(1): 83–92.
42. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems 2014; 2(3): 1–10.
43. Yildirim P, Ekmekci OI, Holzinger A. On Knowledge Discovery in Open Medical Data on the Example of the FDA Drug Adverse Event Reporting System for Alendronate (Fosamax). Lecture Notes in Computer Science 2013; 7947: 195–206.