Data Mining Model to Predict Fosamax Adverse Events

Neveen Ibrahim
College of Computing and Information Technology
Arab Academy for Science, Technology, and Maritime Transport
Alexandria, Egypt
Email: neveen1010 {at} yahoo.com

Nahla Belal and Osama Badawy
College of Computing and Information Technology
Arab Academy for Science, Technology, and Maritime Transport
Alexandria, Egypt

Abstract—**Fosamax (Alendronate) is an approved drug widely prescribed for osteoporosis treatment and other bone damaging diseases. Fosamax causes a number of serious side effects at the long-term, therefore it is important to discover the hidden patterns between patients' information and Fosamax adverse events to predict the Adverse Drug Events (ADEs) for new patients. In this paper, we investigate many data mining techniques, mainly the multi-label classification methods through a framework for Clinical Decision Support System (CDSS) to extract useful knowledge from the U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (AERS) database for Fosamax, which can help healthcare providers make better decisions and reduce errors. Depending on the multi-label experimental results, BR with SVM obtained the best accuracy (76%), also BR with SVM obtained the best hamming score (77%), and CC and LC with J48 were the highest exact match (23%).**

Keywords-Data mining; Clinical decision support system; Multi-label classification; Adverse drug events; Electronic health record; Osteoporosis disease; Fosamax

I. INTRODUCTION

A. Problem Identification

Osteoporosis is a serious health problem because of the significant morbidity, mortality, and costs of treatment. It can strike at any age but it occurs most often in older people and in women after the age of 50. According to the International Osteoporosis Foundation (IOF) [28], 53.9% of women after age 50 have osteopenia (pre-osteoporosis) while 28.4% have osteoporosis and 21.9% of males aged 20-89 have osteoporosis in Egypt. IOF indicates that by 2020 up to nearly 25% of the population in Middle East countries will be over 50 years old, and will grow to 40% by 2050, as a consequence the osteoporosis infection rates will be increased in the future years [28]. Osteoporosis occurs when bones lose an excessive amount of their protein and mineral content, such as calcium. Over time, bone mass and bone strength, is decreased. As a result, bones become fragile and break easily. The most common sites of osteoporotic fracture are the wrist, spine, shoulder, and hip.

According to the Canadian Organization for Osteoporosis (COO) [29], at least 1 in 3 women and 1 in 5 men will suffer from an osteoporotic fracture during their lifetime, over 80% of all fractures in people 50+ are caused by osteoporosis, 28% of women and 37% of men who suffer a hip fracture will die within the following year. Osteoporosis has many outcomes, including difficulty with balance, weakness, problems with daily activity, poor health, permanent disability, lifetime treatment, hospitalization, or even death. Osteoporosis is often called the "silent" disease, because bone loss occurs without symptoms or signs, people often do not know they have the disease until a bone breaks.

Fosamax (the brand name of Alendronate) is an approved drug widely prescribed in Egypt for osteoporosis treatment and other bone damaging diseases in men and women. The drug causes a number of side effects, according to the drug's manufacturer, Merck, the most common side effects are nausea, diarrhea, cramping, skin rashes, and eye problems [30]. The researchers concluded that long-term (after 5 years) Fosamax usage is a significant risk factor for serious adverse events that may be more dangerous than osteoporosis itself. The U.S. FDA [27] documented the most important of them which are femur fracture, stress fracture, and dead jaw syndrome (osteonecrosis).

B. Motivation

Post-market monitoring for adverse events is an important phase for any drug to guarantee patients safety [3]. Due to the long-term effects for Fosamax and their outcomes which were explored previously, it is important to discover the hidden relationship between Fosamax adverse events and the patients' information. This relationship study will produce useful knowledge that can help predict the Adverse Drug Events (ADEs) for new osteoporosis patients. Moreover, the extracted knowledge can assist physicians through Clinical Decision Support System (CDSS) to make better diagnosis, decision making, and treatment, resulting in improved healthcare service quality.

C. Methodology

The aim of the research is to propose a framework for CDSS to predict the adverse events for Fosamax drug for osteoporosis patients. This system takes advantage of Electronic Health Record (EHR) and data mining techniques

to help healthcare providers in medical decision making to reduce medical errors and guarantee the safety of patients. Data mining techniques are widely used to discover hidden patterns in the biomedical and healthcare fields [1]. It has several algorithms for clustering, classification, and association. According to [15], data mining is very useful to provide decision support in the healthcare settings. Healthcare organizations aim to improve the quality of care while reducing costs. Due to the massive volume of data generated in healthcare settings, healthcare organizations have been interested in data mining to enhance physician practices, disease management, and resource utilization.

U.S. FDA is responsible for approving drugs for marketing, in addition, it plays an essential role in monitoring drug safety [27]. It maintains a spontaneous reporting system called Adverse Event Reporting System (AERS). AERS receives ADEs reports from pharmaceutical companies, physicians, nurses, pharmacists, and consumers, and stores these reports in a computerized repository [3]. Here, FDA's AERS for Fosamax was employed to gain useful knowledge.

This paper proposes a framework for a CDSS to predict the ADEs of Fosamax for new patients. The main focus of the paper is the data mining component of the system. The research implements data mining algorithms using patients' data that utilized Fosamax. Moreover, this study analyzes the data collected from the FDA to extract hidden patterns between patient demographics and adverse events for Fosamax using association and classification techniques. The results of both techniques are compared to identify the prediction technique with better results.

The outline of the paper follows. Second section presents a literature review on data mining applications in healthcare field, multi-label classification methods, and different models of CDSS. Third section proposes a framework for a CDSS, data mining processes, data sources, and data preprocessing. Fourth section describes the study implementation. Fifth section displays the results. Finally, the conclusion and future work in the last section.

II. RELATED WORK

Data mining aims to discover pattern and relationship within large datasets [15]. Data mining methods have been used in many industrial areas, mainly in the field of healthcare. [1] defined data mining and its processes, then discussed the use of data mining in biomedical and healthcare fields, and explained data mining algorithms. [2] discussed the use of data mining techniques to handle the medical problems, then demonstrated learning methods in data mining, data mining tasks, scope of data mining, issues, and importance of data mining in healthcare field. [3] searched the hidden relationship between patients' information and adverse events for Fosamax from FDA' adverse event reporting system database. Some association rules were generated, which can be used by medical researchers. [4] presented a study of different types of data mining applications in the healthcare sector, and also compared data mining techniques to discover knowledge from healthcare databases. [5] analyzed FDA' AERS to obtain top 10 drugs associated with outcomes (death, disability, hospitalization, and life-threatening). Naïve Bayes method was applied to rank the drugs based on posterior probability and the result showed that Fosamax is the second drug causing disability. [6] explored drugs interactions from huge document repository. Several machine learning algorithms (SVM, decision tree, and NB) in combination with feature selection techniques were implemented to mine these documents. Sampling techniques were carried out to solve the unbalanced data problem and the decision tree obtained the best results. [7] provided clinical decision support in treatment and defined EHR as patient health-related data such as: age, gender, weight, symptoms, diagnosis, tests results, and treatments. Clinicians must capture and analyze these sources to make a correct and timed decision. Indeed, without using any technology, EHR will be an electronic store for patient data, but applying data mining techniques on patient attributes make EHR more useful and valuable, therefore, it is very important to update it continuously. The study data were extracted from Centerstone's EHR and feature selection methods were applied with WEKA classifier models: NB, NN, KNN, J48, and RF. The highest accuracy in predicting treatment outcomes was between 70-72%. [8] detected Hospital Acquired Infections (HAI) from patients' health records. Filter methods were deployed to reduce the features. Machine learning techniques (NB, SVM, and C4.5) were performed to detect infections. Considering recall, SVM yields the highest rate. [9] presented data mining methods in the healthcare field and the limitations to apply data mining in the health sector were declared. [13] studied the behavior of a classifier using oversampling and undersampling methods in unbalanced databases. [23] demonstrated knowledge discovery and data mining process, application of data mining in healthcare, advantages of data mining application in healthcare, and the limitations for data mining in healthcare.

According to [12], the multi-label classification problem handles a set of instances where each instance is assigned to one or more classes, unlike single-label classification where each instance is assigned to only one class. [10] evaluated the performance of multi-label classification algorithms which were developed based on problem transformation. The experiment provided that Multi-Label K-Nearest Neighbor (MLKNN) was the best, followed by Random k-Label Set (RAkEL), followed by Classifier Chain (CC), followed by Binary Relevance (BR). [11] compared different problem transformation methods (BR, CC, LC, PS, and RAkEL) over different application domains using MEKA software based on many classifiers (SVM, NN, NB, J48, and KNN). With all datasets, BR obtained the better outcome with different evaluation metrics. MEKA (Multi-label Extension to WEKA) software is used to support multi-label and multi-target

classifiers (which WEKA does not). In the multi-label problem, a data instance may be related with multiple labels (attributes), all variables are binary, indicating label relevance (1) or irrelevance (0). In multi-target learning, a data instance is associated with multiple target variables, where each variable takes a number of values (not binary). There are two main steps for handling the multi-label classification problem, the first step is the Problem Transformation (PT), which transforms the multi-labeled data into single-labeled data, and the second is the algorithm adaptation, to apply the traditional single classifiers methods to the transformed data. [12] analyzed PT methods (BR, LC, and PS) in combination with classifiers (NB, ZeroR, and J48) using evaluation measures (accuracy, exact match, and hamming loss). The results indicated that LC and PS were the best.

Data mining techniques are incorporated in the CDSS to help healthcare providers in decision making. "*Clinical Decision Support System (CDSS) are computer systems designed to impact clinician decision making about individual patients*" [14]. [16] produced information architecture of CDSS which consists of four components. First, patient model which includes measured health parameters, treatment history, and health goals. Second, treatment library which stores treatment procedures, doses, and effects. Third, intelligent agents which are used to provide a recommendation to achieve desired health goals. Finally, an authenticated knowledge base to keep the patient model and treatment library up to date. [17] introduced a knowledge management framework for distributed healthcare systems that consists of data and knowledge bases and using data mining techniques to provide decision making support for the healthcare provider. The patient data is mined off-line and the extracted knowledge is shared through XML documents with other healthcare systems. [18] explained a context model to provide cross boundary decision support in health system. This model of context worked between the domain model and the activity landscapes (individuals, workgroups, and organizations) and between these landscapes and the knowledge resource space model. [19] demonstrated the challenges, process, and outcomes of defining and implementing a CDS architecture which include five components: inference server, authoring environment, interface server and repository, knowledge repository, and service interface. [20] produced a distributed CDSS architecture which involves electronic health record, data mining techniques, clinical databases, domain expert knowledge bases, available technologies, and standards to help healthcare professionals in decision making. [21] presented a Multi Agent System (MAS) to support coupling CDSS with Computerized Prescribe Order Entry (CPOE) and incorporated the MAS in the medical workflow management system which is based on collaborating agents. Hence, each agent plays a role and uses one or several clinical data sources. [22] offered three criteria to provide a better chance for successful deployment of a CDSS which are the data entry and the decision algorithms, the interaction between human and computer, and the output of the CDSS. [24] provided the design of cross boundary decision support in health system, the goal of this system is transfering medical knowledge and practices among clinicians across regional boundaries. [25] proposed a CDSS architecture implementation with knowledge engine.

III. DESIGN

A. *A framework for a Clinical Decision Support System*

The proposed CDSS, shown in "Fig.1", takes as input the patient ID and provides potential ADEs. This framework takes advantage of EHR and data mining techniques to detect ADEs for Fosamax to assist healthcare provider in decision making. The execution of this framework will work as follows:

1) Healthcare provider must enter the patient ID which identifies the patient health record. If the patient has no record, the system will return a message to create a new health record for this patient.
2) The ID will be sent to the EHR database to retrieve the available patient information and provides the patient profile.
3) FDA's AERS for Fosamax is employed with data mining techniques. The patient age, gender, and adverse events attributes were selected. Many techniques were performed on the Fosamax dataset to extract the hidden relationship between patients' information and Fosamax adverse events, as will be shown in the next section. Here, the results were concerned with the top 10 adverse events relevant and irrelevant with an instance.
4) Fosamax adverse events detection is based on patient information from EHR and the extracted knowledge from data mining. The proposed CDSS will apply this knowledge on selected patient attributes (age and gender) to get the potential adverse events for that patient.

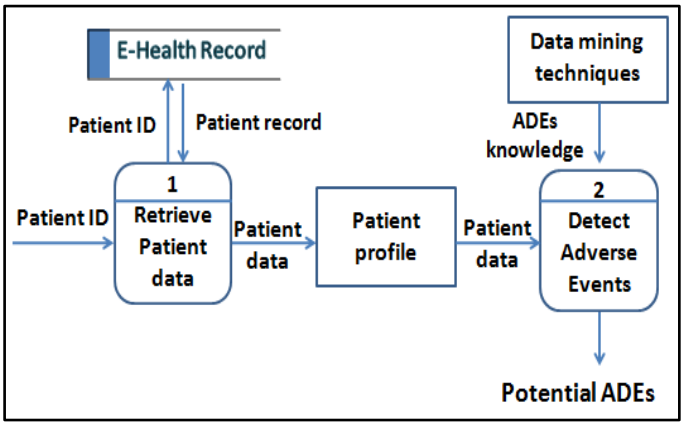

Figure 1. A framework of a clinical decision support system

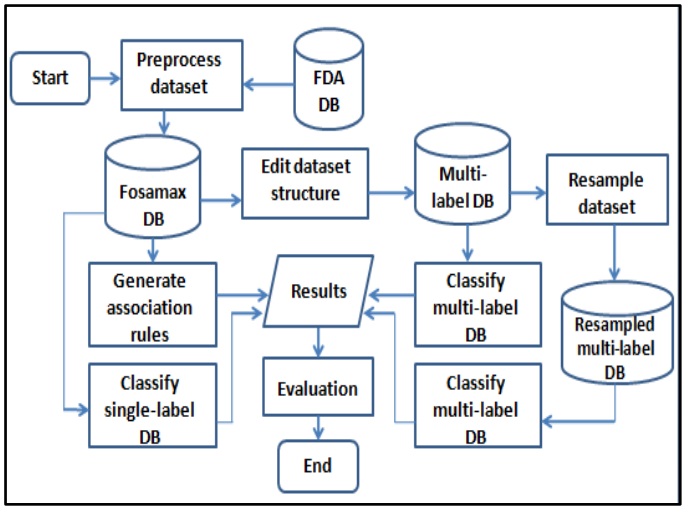

Figure 2. Data mining processes flowchart

B. *Data mining processes*

In the proposed CDSS, data mining process consists of the steps shown in "Fig. 2". In the beginning, Fosamax adverse events dataset was selected from FDA's AERS database, then association rules and single-label classification algorithms were applied using WEKA toolkit [31]. To perform multi-label classification in MEKA software [32], the dataset structure was modified and then resampled. Finally, all algorithms results were compared to evaluate their performance and determine the best one.

C. *Data sources*

The study extracted data from the public release of the FDA's AERS database [27]. In this study the database covers the period from the first quarter of 2004 through the third of 2012. Each extract covers reports received by FDA' AERS during one quarter of the year. The data are provided in two distinct formats in the extract: ASCII files and SGML files. In this research, ASCII files were extracted, in which data elements are separated from each other by a $ sign ("$ delimiter").

The data structure of AERS consists of seven datasets. Those files are: DEMOyyQq.txt, DRUGyyQq.txt, REACyyQq.txt, OUTCyyQq.txt, RPSRyyQq.txt, THERyyQq.txt, and INDIyyQq.txt. The set of the seven ASCII data files in each extract contains raw data for the full quarter covered by the extract. All the ASCII data files are linked using the primary link field ISR, which is a unique number for identifying an AERS report.

D. *Data preprocessing*

The research dataset were extracted from only three ASCII files. The first file is DEMOyyQq.TXT, which contains patient demographic and administrative information, a single record for each event report. The second is DRUGyyQq.TXT, and it contains drug/biological information for as many medications as were reported for the event (1 or more per event). The third file is REACyyQq.TXT, which has all of the "Medical Dictionary for Regulatory Activities" (MedDRA) terms coded for the event (1 or more).

A snapshot of the original dataset is shown in (Table. 1). The original dataset consists of three attributes, which are ISR, Age, and Gender, 51494 instances (with duplicated ISR), and single class (with multi values).

To handle multi-label classification, the dataset structure is changed. The modified dataset is shown in (Table. 2), it also has three attributes, ISR, Age, and Gender, but with 18026 instances (with unique ISR), and ten classes (with binary value), one class for each of the top 10 adverse events.

IV. IMPLEMENTATION

The datasets in the different time periods were imported into Microsoft SQL server 2012 database management system as database tables. Then, Fosamax related records were selected if Fosamax is flagged as the primary suspect (PS) drug causing the adverse event. Age was categorized. This research focused on the most frequent adverse events (the top 10) in the Fosamax dataset. The dataset includes 23531 unique reports. Since each patient has one or more adverse events, therefore the total instances in the dataset are 60775 instances.

Despite FDA's AERS is a rich data source, it has some limitations such as the missing and incorrect data resulting in problems with data quality. In this research, the missing data were removed, as a consequence, the final top 10 adverse events for Fosamax dataset includes 18026 unique reports and 51494 instances.

A. *Association Rules*

The Apriori algorithm was used to perform association analysis on the attributes of patient demographics and adverse events for Fosamax. WEKA 3.7.9 [31] was used with the original dataset which is shown in (Table. 1). Some best rules are shown in the next section.

ISR	Age	Gender	Adverse Events
4617605	>=65	M	Femur Fracture
4617605	>=65	M	Depression
4617605	>=65	M	Hypertension

Table 1. Example of original dataset

ISR	Age	Gender	Femur fracture	Osteon ecrosis	Fall	Depres sion	Anxiety	Osteoa rthritis	Low turnover osteopat hy	Hypert ension	Osteo myelitis	Tooth disorde r
4617605	>=65	M	yes	no	no	yes	no	no	no	yes	no	no

Table 2. Example of modified dataset

B. Single-label Classification

Many single-label classification methods were applied using the original dataset which is shown in (Table. 1) in WEKA such as: SVM, NB, and J48, the results of these classifiers are shown in the next section.

C. Multi-label Classification

Multi-label classification algorithms were performed using the modified dataset which is shown in (Table. 2) in MEKA 1.5 software in order to discover a set of Fosamax adverse events which are associated with each patient based on his/her demographics information.

D. Transformation Methods

Three problem transformation methods in MEKA were performed: binary relevance, classifier chains and label power set.

1) Binary Relevance (BR)

Binary Relevance transforms the original dataset into q datasets (q= total number of classes in a dataset), one for each label, where each dataset includes all the instances of the original dataset and trains binary classifier on each of these datasets. To classify a new instance, BR outputs the union of the labels that are predicted by the q classifiers. It is used only in applications which have data independency. An example is shown in "Fig. 3".

2) Classifier Chains (CC)

Classifier Chain contains classifiers which are linked along a chain, where each classifier handles the binary relevance problem associated with each label. It creates a chain of classifier C1, C2, ... , CL, where L is the total number of labels. To classify a new instance, CC starts from C1 and runs down along the chain. Each classifier determines the probability of being classified into L1, L2, ... , LL. The chain method passes label information between classifiers to take into account label correlation. It combines the advantages of binary relevance and label dependency.

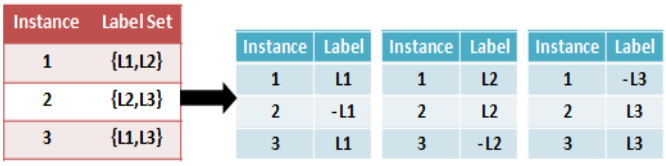

Figure 3. Binary Relevance example

3) Label Power set (LP/LC)

Label Power set considers each unique occurrence of a set of labels as one class. It takes into account label dependency. For example, if an instance is associated with three labels L1, L2, and L4, then the new single-label class will be L1,2,4. To classify a new instance, LP/LC outputs the most probable class, which is actually a set of labels. The limitation is its complex computations (depends on the number of distinct label sets) and the large number of label sets makes the training examples limited.

E. Adaptation Methods

Three classifiers in MEKA were performed: J48, NB, and SVM.

F. Resampling Dataset

The modified dataset (Table. 2) is unbalanced, because the patient reports which have no adverse events have higher rates, as shown in the next section. For solving the unbalanced dataset problem, resample technique is employed in MEKA to get a more balanced class distribution.

G. Evaluation Measures

The most common evaluation measures for multi-label classification are: accuracy, hamming score, and exact match.

1) *Accuracy:* is the ratio between the correct labels to the total number of labels for each instance, averaged across all instances.

2) *Hamming Score:* is the accuracy for each label (class) to correctly predicted, averaged across all labels.

3) *Exact Match:* is the accuracy of each example where all label relevance must match exactly for an example to be correct.

V. Results

There are some web based tools to analyze the FDA's AERS database. Drugcite [26] is a useful web tool which analyzes the FDA's AERS database between 2004 and 2012 to present detailed information about drugs, but it has no data mining functions [3]. According to Drugcite, the three most frequent adverse events for Fosamax are femur fracture, osteonecrosis, and fall, which is the same result of our study, as shown in "Fig. 4". In addition, Drugcite mentioned that different counts may occur depending on how the data are aggregated.

According to [3] using apriori algorithm generated many rules, as shown in (Table. 3). For all rules, the confidence is ranging between 63% and 75%. The adverse events associated with female patients over 65 are femur fracture, anaemia, hypertension, nausea, depression, and osteonecrosis. While the

adverse events associated with female patients between 44 and 64 are arthralgia, diarrhea, and anxiety.

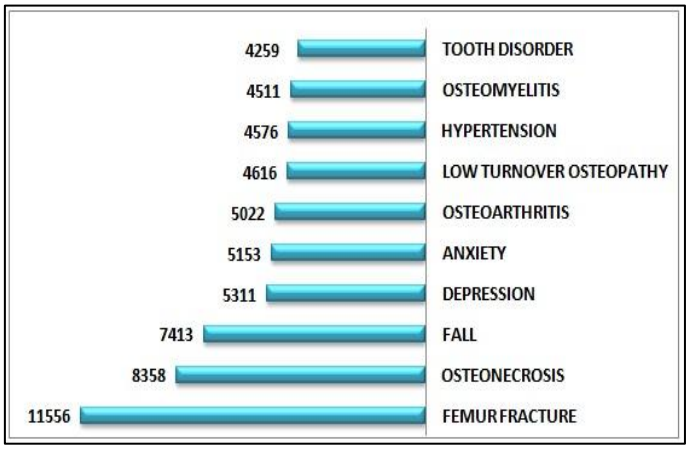

Figure 4. The most frequent Fosamax adverse events

Table 3. The previous study rules

No	Rule	Confidence
1	Age=44-64 Adverse Event=Arthralgia==> Gender=Female	75%
2	Age≥65 Adverse Event = Femur fracture ==> Gender=Female	74%
3	Age≥ 65 Adverse Event =Anaemia ==> Gender=Female	73%
4	Adverse event=Impaired healing ==> Gender=Female	71%
5	Age=44-64 Adverse Event=Diarrhoea ==> Gender=Female	70%
6	Age≥65 Adverse Event=Hypertension ==> Gender=Female	68%
7	Age≥ 65 Adverse Event =Nausea ==> Gender=Female	72%
8	Age ≥65 Adverse Event=Depression ==> Gender=Female	64%
9	Age=44-64 Adverse Event=Anxiety ==> Gender=Female	63%
10	Age ≥65 Adverse Event=Osteonecrosis ==> Gender=Female	63%

Table 4. Our study rules

No	Rule	Confidence
1	Age=>=60 Adverse_Event=FEMUR_FRACTURE ==> Gndr_cod=F	96%
2	Adverse_Event=FALL ==> Gndr_cod=F	96%
3	Age=>=60 Adverse_Event=FALL ==> Gndr_cod=F	96%
4	Adverse_Event=FEMUR_FRACTURE ==> Gndr_cod=F	96%
5	Age=50-59 ==> Gndr_cod=F	95%
6	Age=>=60 Adverse_Event=ANXIETY ==> Gndr_cod=F	95%
7	Adverse_Event=ANXIETY ==> Gndr_cod=F	94%
8	Age=>=60 ==> Gndr_cod=F	94%
9	Age=>=60 Adverse_Event=OSTEONECROSIS ==> Gndr_cod=F	91%
10	Adverse_Event=OSTEONECROSIS ==> Gndr_cod=F	90%

Based on our results using apriori algorithm, as shown in (Table. 4). The confidence is ranging between 90% and 96%. Some adverse events have strong association with female patients over 60 are femur fracture, fall, anxiety, and osteonecrosis. For example, rule 1 means that the possibility of femur fracture with female patients over 60 is 96%.

From the foregoing, there is a similarity between the extracted rules in the two studies, but in our research the confidence is better than the previous study. The dataset in [3] includes 9229 reports and selected Alendronate as a drug name. The dataset in our research has 18026 reports and used Fosamax as a drug name. The reasoning behind the high confidence in our study is due to the high number of instances. On the other hand, each rule in both studies presents the association between patient information and one adverse event, but in fact, each patient is related with one or a set of adverse events. As a consequence, some classification techniques were experimented to solve this problem.

After single-label classification was applied, the best F-score was 20.4% by SVM. The reason for the low F-score obtained is that WEKA is not suitable for this dataset because each patient is related to one or more adverse event, which is a multi-label classification problem.

After the modification on the Fosamax dataset which is shown previously in (Table. 2) the unbalanced problem is shown in "Fig. 5".

According to the evaluation results of multi-label classification (Table. 5), (Table. 6), and (Table. 7), all measures values increased after re-sampling. Based on the accuracy and the hamming score, BR with SVM was the best (76% and 77%), but it was the lowest exact match (4%) and consumed time, so it was eliminated. CC with J48 obtained (75%) accuracy, (76%) hamming score, and it was the highest exact match (23%), so it will be recommended.

Multi-label classification rules (Table. 8) differ from apriori association rules (Table. 3) and (Table. 4) in terms of the number of adverse events in each rule. For example in (Table. 8), rule 1 means that if patient's age is between 0 and 12 years old then the potential adverse events are both tooth disorder and depression for female or male.

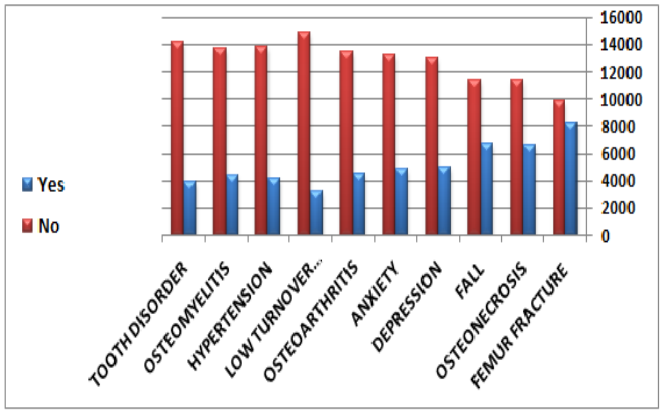

Figure 5. Unbalanced Fosamax dataset

Table 5. Accuracy result

	Binary Relevance (BR)		Classifier Chains (CC)		Label Power-set (LC)	
	Before Resampling	After Resampling	Before Resampling	After Resampling	Before Resampling	After Resampling
J48	0.65	0.71	0.68	0.75	0.68	0.73
NB	0.63	0.70	0.69	0.71	0.68	0.75
SVM	0.71	**0.76**	0.67	0.75	0.68	0.74

Table 6. Hamming score result

	Binary Relevance (BR)		Classifier Chains (CC)		Label Power-set (LC)	
	Before Resampling	After Resampling	Before Resampling	After Resampling	Before Resampling	After Resampling
J48	0.68	0.74	0.69	0.76	0.69	0.75
NB	0.67	0.73	0.71	0.74	0.69	0.76
SVM	0.71	**0.77**	0.70	0.76	0.69	0.76

Table 7. Exact match result

	Binary Relevance (BR)		Classifier Chains (CC)		Label Power-set (LC)	
	Before Resampling	After Resampling	Before Resampling	After Resampling	Before Resampling	After Resampling
J48	0.03	0.14	0.10	**0.23**	0.11	**0.23**
NB	0.02	0.15	0.10	0.16	0.10	0.22
SVM	0	0.04	0.09	0.22	0.10	0.22

Table 8. Multi-label classification rules

Age	Gender	Adverse events
0-12	Female / Male	Tooth Disorder and Depression
13-24	Female	Femur Fracture and Fall
	Male	Femur Fracture, Tooth Disorder, and Low Turnover Osteopathy
25-43	Female	Femur Fracture
	Male	Osteonecrosis
44-64	Female / Male	Osteonecrosis
>=65	Female / Male	Femur Fracture

VI. CONCLUSION AND FUTURE WORK

In this research, we propose a framework for a CDSS to predict the most common ADEs for Fosamax, mainly the top 10 events. The proposed CDSS takes advantage of EHR and data mining knowledge. Many association rules and single-label classification methods using WEKA, in addition, many multi-label classification methods using MEKA, were performed on FDA's AERS database and their performance was evaluated in order to extract useful and valuable knowledge.

According to the experimental results, the association rule obtained the relationship between patients' information and only one side effect in each rule, and the single-label classification is not appropriate for this kind of dataset. On the other hand, the multi-label classification obtained the best outcomes, because it presents a set of Fosamax adverse events associated with a patient based on his/her age and gender. Based on the accuracy and hamming score measures, BR with SVM was the best. Based on the exact match, CC and LC with J48 were the best.

The main contribution of this research is the investigation of multi-label classification methods to forecast the adverse events for a drug to a certain patient. Moreover, resampling the research dataset in MEKA improved all measures' values with all algorithms. Finally, the multi-label classification outcomes can help the physicians in decision making through the CDSS, especially in our country where the bones doctors

pointed out that there is no system for monitoring the effect of drugs on patients on long-term. For future work, medical experts can explain and interpret the extracted knowledge. Moreover, this study should be continued to detect the less frequent adverse events for the same drug.

REFERENCES

[1] I. Yoo, P. Alafaireet, et al., "Data mining in healthcare and biomedicine: a survey of the literature," Journal of Medical Systems, Springer, Volume 36, Issue 4, pp 2431-2448, August 2012.

[2] R. Anand and S. K. Srivatsa, "A Data Mining Framework For Building Healthcare Management System," International Journal of Engineering Research and Technology, Volume 2, Issue 5, May 2013.

[3] P. Yildirim, I. O. Ekmekci, and A. Holzinger, "On Knowledge Discovery in Open Medical Data on the Example of the FDA Drug Adverse Event Reporting System for Alendronate (Fosamax)," Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Springer, Volume 7947, pp 195-206, 2013.

[4] M. Durairaj and V. Ranjani, "Data Mining Applications in Healthcare Sector: A Study," International Journal of Scientific and Technology Research, Volume 2, Issue 10, October 2013.

[5] A. B. Banu, S. A. A. Balamurugan, and P. Thirumalaikolundu, "Ranking Drugs in Spontaneous Reporting System by Naïve Bayes," Journal of Biological Sciences, Volume 13, Issue 4, pp 293-297, 2013.

[6] J. Mata, R. Santano, D. Blanco, M. Lucero, and M. Mana, "A Machine Learning Approach to Extract Drug - Drug Interactions in an Unbalanced Dataset," Drug-Drug Interaction (DDI) Extraction Workshop, pp 59-65, September 2011.

[7] C. Bennett and T. W. Doub, "Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice," International Conference on Data Mining (ICDM), pp 313-318, 2010.

[8] C. Ehrentraut, H. Tanushi, H. Dalianis, and J. Tiedemann, "Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records," Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data (AND), 2012.

[9] F. E. Bekri and A. Govardhan, "Association of Data Mining and healthcare domain: Issues and current state of the art," Journal of Computer Science and Technology, Volume 11, No 21, 2011.

[10] A. Tawiah and V. S. Sheng, "A Study on Multi-label Classification," Industrial Conference on Data Mining (ICDM), Springer, Volume 7987, pp 137-150, 2013.

[11] P. K. A. Chitra and S. A. A. Balamurugan, "Performance Analysis of Transformation Methods in Multi-Label Classification," International Conference on Advanced Computing, Networking, and Informatics, Springer, Volume 243, pp 1233-1239, 2014.

[12] H. Modi and M. Panchal, "Experimental comparison of different problem transformation methods for multi-label classification using MEKA," International Journal of Computer Applications, Volume 59, No 15, December 2012.

[13] O. Loyola-Gonz´alez, M. Garc´ıa-Borroto, et al., "An Empirical Study of Oversampling and Undersampling Methods for LCMine an Emerging Pattern Based Classifier," Pattern Recognition, Springer, Volume 7914, pp 264-273, 2013.

[14] E. S. Berner and T. J., et al., "Overview of Clinical Decision Support Systems". Clinical Decision Support System, Health Informatics, Springer: 3-22, 2007.

[15] J. M. Hardin and D. C. Chhieng, "Data Mining and Clinical Decision Support System," Clinical Decision Support System, Health Informatics, Springer: 44-63, 2007.

[16] D. E. Robbins, V. P. Gurupur, and J. Tanik, "Information architecture of a clinical decision support system," Southeastcon, Proceedings of IEEE, pp 374-378, March 2011.

[17] R. S. Kazemzadeh and K. Sartipi, "Interoperability of data and knowledge in distributed healthcare systems," 13th IEEE International Workshop on Software Technology and Engineering Practice, pp 230-240, 2005.

[18] O. Anya, H. Tawfik, and A. Nagar, "Cross-Boundary Knowledge-based Decision Support in e-Health," International Conference on Innovations in Information Technology, IEEE, pp 150-155, April 2011.

[19] I. Cho, J. Kim, J. H. Kim, H. Y. Kim, and Y. Kim, "Design and implementation of a standards-based interoperable clinical decision support architecture in the context of the Korean HER,".International Journal of Medical Informatics, Volume 79, Issue 9, pp 611-622, September 2010.

[20] S. H. El-Sappagh and S. El-Masri, "A distributed clinical decision support system architecture," Journal of King Saud University - Compter and Information Sciences, Volume 26, Issue 1, pp 69-78, January 2014.

[21] L. Bouzguenda and M. Turki, "Coupling clinical decision support system with computerized prescriber order entry and their dynamic plugging in the medical workflow system," International Conference on Information Technology and e-Services (ICITeS), IEEE, pp 1-6, March 2012.

[22] M. Frize, S. Weyand, and E. Bariciak, "Suggested criteria for successful deployment of a Clinical Decision Support System (CDSS)," IEEE International Workshop on Medical Measurements and Applications Proceedings (MeMeA), IEEE, pp 69-72, May 2010.

[23] B. Milovic and M. Milovic, "Prediction and decision making in Healthcare using Data Mining," International Journal of Public Health Science, Volume 1, No 2, pp 69-78, December 2012.

[24] O. Anya, H. Tawfik, and A. Nagar, "CaDHealth: Designing and Prototyping for Cross-Boundary Decision Support in E-Health," Developments in e-Systems Engineering (DeSE), IEEE, pp 111-116, December 2011.

[25] J. A. Kim, S. T. Kim, Shim, and J. Lee, "Implementation of guideline-based CDSS," International Conference on Ubiquitous Computing and Multimedia Applications (UCMA), IEEE, pp 96-99, April 2011.

[26] Drugcite tool, www.drugcite.com

[27] U.S. Food and Drug Administration for Adverse Event Reporting System database, www.fda.gov

[28] International Osteoporosis Foundation, www.iofbonehealth.org

[29] Canadian Organization for Osteoporosis, www.osteoporosis.ca

[30] DrugWatch: dangerous side effects resource, www.drugwatch.com

[31] Weka: Machine Learning Toolkit, www.cs.waikato.ac.nz/ml/weka

[32] Meka: A Multi-label Extension to Weka, meka.sourceforge.net