Knowledge Discovery and Visualization of Clusters for Erythromycin Related Adverse Events in the FDA Drug Adverse Event Reporting System

Pinar Yildirim[1], Marcus Bloice[2], and Andreas Holzinger[2]

[1] Department of Computer Engineering, Faculty of Engineering & Architecture,
Okan University, Istanbul, Turkey
pinar.yildirim@okan.edu.tr

[2] Medical University Graz, Institute for Medical Informatics, Statistics and Documentation
Research Unit HCI, Auenbruggerplatz 2/V, A-8036 Graz, Austria
marcus.bloice@medunigraz.at, a.holzinger@hci4all.at

Abstract. In this paper, a research study to discover hidden knowledge in the reports of the public release of the Food and Drug Administration (FDA)'s Adverse Event Reporting System (FAERS) for erythromycin is presented. Erythromycin is an antibiotic used to treat certain infections caused by bacteria. Bacterial infections can cause significant morbidity, mortality, and the costs of treatment are known to be detrimental to health institutions around the world. Since erythromycin is of great interest in medical research, the relationships between patient demographics, adverse event outcomes, and the adverse events of this drug were analyzed. The FDA's FAERS database was used to create a dataset for cluster analysis in order to gain some statistical insights. The reports contained within the dataset consist of 3792 (44.1%) female and 4798 (55.8%) male patients. The mean age of each patient is 41.759. The most frequent adverse event reported is oligohtdramnios and the most frequent adverse event outcome is OT(Other). Cluster analysis was used for the analysis of the dataset using the DBSCAN algorithm, and according to the results, a number of clusters and associations were obtained, which are reported here. It is believed medical researchers and pharmaceutical companies can utilize these results and test these relationships within their clinical studies.

Keywords: Open medical data, knowledge discovery, biomedical data mining, bacteria, drug adverse event, erythromycin, cluster analysis, clustering algorithms.

1 Introduction

Modern technology has increased the power of data by facilitating linking and sharing. Politics has embraced transparency and the citizens' rights to data access; the top down culture is being challenged. Many governments around the world now release large quantities of data into the public domain, often free of charge and without administrative overhead.

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 101–116, 2014.

This allows citizen-centered service delivery and design and improves accountability of public services, leading to better public service outcomes [1]. Therefore, open data has been of increasingly great interest to several scientific communities and is a big opportunity for biomedical research [2], [3], [4].

The US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) is such a public database and contains information on adverse events and medication error reports submitted to the FDA [5]. The database is designed to support the FDA's post marketing safety surveillance program for drug and therapeutic biologic products [6], [7], [8]. Adverse events and medication errors are coded using terms from the Medical Dictionary for Regulatory Activities (MedDRA) terminology [9]. Reports can be submitted by health care professionals and the public through the "MedWatch" program. Since the original system was started in 1969, reporting has been markedly increasing. To date, the FAERS is the largest repository of spontaneously reported adverse events in the world with more than 4 million reports [10], [11].

The FDA releases the data to the public, and public access offers the possibility to external researchers and/or pharmacovigilance experts to explore this data source for conducting pharmacoepidemiological studies and/or pharmacovigilance analyses [5].

This study was carried out to describe the safety profile of erythromycin. This is of great importance as erythromycin is one of the main medications for bacterial diseases. Bacterial diseases are of particular interest due to the high morbidity, mortality, and costs of disease management [12]. Previous work has investigated the adverse events of erythromycin. Manchia et al. presented a case of a young man who had symptoms of psychotic mania after the administration of erythromycin and acetaminophen with codeine on 2 separate occasions [13]. Varughese et al. reported antibiotic-associated diarrhea (AAD) associated with the use of an antibiotic such as erythromycin [14].

Bearing the importance of any new insights into erythromycin in mind, the data from the FDA's FAERS was used to discover associations between patient information such as demographics (e.g., age and gender), the route of the drug, indication for use, the adverse event outcomes (death, hospitalization, disability, etc.), and the adverse events of erythromycin were explored. A number of statistically significant relations in the event reports were detected. The automated acquisition, integration, and management of disease-specific knowledge from disparate and heterogeneous sources are of high interest in the data mining community [15].

In the project which this paper describes, data mining experts and clinicians worked closely together to achieve these results.

2 Glossary and Key Terms

Bacteria: are living organisms that have only one cell. Under a microscope, they look like spheres, rods, or spirals. They are so small that a line of 1,000 could fit across a pencil eraser. Most bacteria do no harm - less than 1 percent of bacteria species cause any illnesses in humans. Many are helpful. Some bacteria help to digest food, destroy disease-causing cells, and give the body needed vitamins [42].

Cluster analysis: is the process of grouping data into classes or clusters so that objects within a cluster have high similarity in comparison to other objects in that cluster, but are very dissimilar to objects in other clusters [36].

DBSCAN: a density-based clustering algorithm. A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability. Every object not contained in any cluster is considered to be noise [36].

Drug adverse event: An appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product [43].

FDA FAERS (Food and Drug Administration Adverse Event Reporting System): is a public database that contains information on adverse event and medication error reports submitted to the FDA [5].

Open data: Data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike [44].

Pharmacovigilance: is the science relating to prevention of adverse effects with drugs.

3 Related Work

Several studies have been carried out regarding data mining on drug adverse event relations in the biomedical domain. Kadoyama et al. mined the FDA's FAERS for side-effect profiles of tigecycline. They used standardized, official pharmacovigilance tools using of a number of measures such as proportional ratio, the reporting odds ratio, the information component given by a Bayesian confidence propagation neural network, and the empirical Bayes geometric mean. They found some adverse events with relatively high frequency including nausea, vomiting, and hepatic failure [16].

Malla et al. investigated trabectedin related muscular and other adverse effects in the FDA FAERS database. Adverse event reports submitted to the database from 2007 to September 2011 were retrospectively reviewed and the entire safety profile of trabectedin was explored. They detected that rhabdomyolysis is a life-threatening adverse toxicity of trabectedin [17].

Raschi et al. searched macrolides and torsadogenic risk and analyzed cases of drug induced Torsade de Pointes (TdP) submitted to the publicly available FDA FAERS database. They collected patient demographic, drug, and reaction and outcome information for the 2004-2011 period and performed statistical analyses by using the statistical package SPSS. They concluded that in clinical practice azithromycin carries a level of risk similar to other macrolides; the notable proportion of fatal cases and the occurrence of TdP-related events in middle-aged patients strengthen the view that caution is needed before considering azithromycin as a safer therapeutic option among macrolides. Appropriate prescription of all macrolides is therefore vital and should be based on the underlying disease, patient's risk factors, concomitant drugs, and local pattern of drug resistance [18].

Harpaz et al. have performed a number of studies on data mining for adverse drug events (ADEs). They provide an overview of recent methodological innovations and data sources used to support ADE discovery and analysis [19]. Multi-item ADE associations are associations relating multiple drugs to possibly multiple adverse events. The current standard in pharmacovigilance is bivariate association analysis, where each single ADE combination is studied separately. The importance and difficulty in the detection of multi-item ADE associations was noted in several prominent pharmacovigilance studies. The application of a well-established data mining method known as association rule mining was applied to the FDA's spontaneous adverse event reporting system (FAERS). Several potentially novel ADEs were identified [20]. Harpaz et al. also present a new pharmacovigilance data mining technique based on the biclustering paradigm, which is designed to identify drug groups that share a common set of adverse events in the FDA's spontaneous reporting system. A taxonomy of biclusters was developed, revealing that a significant number of verified adverse drug event (ADE) biclusters were identified. Statistical tests indicate that it is extremely unlikely that the discovered bicluster structures as well as their content arose by chance. Some of the biclusters classified as indeterminate provide support for previously unrecognized and potentially novel ADEs [21].

Vilar et al. developed a new methodology that combines existing data mining algorithms with chemical information through the analysis of molecular fingerprints. This was done to enhance initial ADE signals generated from FAERS to provide a decision support mechanism to facilitate the identification of novel adverse events. Their method achieved a significant improvement in precision for identifying known ADEs, and a more than twofold signal enhancement when applied to the rhabdomyolysis ADE. The simplicity of the method assists in highlighting the etiology of the ADE by identifying structurally similar drugs [22].

The creation and updating of medical knowledge is challenging. Therefore, it is important to automatically create and update executable drug-related knowledge bases so that they can be used for automated applications. Wang et al. suggest that the drug indication knowledge generated by integrating complementary databases was comparable to the manually curated gold standard. Knowledge automatically acquired from these disparate sources could be applied to many clinical applications, such as pharmacovigilance and document summarization [23].

4 Methods

4.1 Data Sources

Input data for our study was taken from the public release of the FDA's FAERS database, which covers the period from the third quarter of 2005 through to the second quarter of 2012. The data structure of FAERS consists of 7 datasets: patient demographic and administrative information (DEMO), drug/biologic information (DRUG), adverse events (REAC), patient outcomes (OUTC), report sources (RPSR), drug therapy start and end dates (THER), and indications for use/diagnosis (INDI). The adverse events in REAC are coded using preferred terms (PTs) from the Medical Dictionary for Regulatory Activities (MedDRA) terminology. All ASCII data files are linked using an ISR, a unique number for identifying an AER. Three of seven files are linked using DRUG_SEQ, a unique number for identifying a drug for an ISR [24], [25].

Table 1. Characteristics of dataset

Attribute	Type
Age	**Numeric** Minimum: 6 Maximum: 91 Mean: 41.759 StdDev: 23.409
Gender	**Nominal** Male, Female, NS
Route	**Nominal** Oral, Transplacental, Ophthalmic, Intravenous, Topical, Parenteral, Disc, Nos
Indication for use	**Nominal** 48 distinct values (MedDRA terms)
Adverse event outcome	**Nominal** HO-Hospitalization, OT-Other, DE-Death, DS-Disability, LT-Life threatening, RI- Required Intervention to Prevent Permanent Impairment/Damage, CA- Congenital Anomaly
Adverse event	**Nominal** 220 distinct values (MedDRA terms)

The data in this study was created from the public release of the FDA's FAERS database by collecting data from the DEMO, DRUG, REAC, OUTC and INDI datasets [17]. The data, in ASCII format, were combined and stored in a database using Microsoft SQL Server 2012. Erythromycin related records were then selected to create a dataset for cluster analysis. In total, 8592 patients involved in adverse event reports for erythromycin were collected from the FDA database [25].

The dataset contains patient demographics such as age, gender, route, indication for use, adverse event outcome, and adverse event (Table 1). The attributes of the dataset were directly collected from the database. The dataset consists of 8592 instances.

4.2 Cluster Analysis by DBSCAN Algorithm

Cluster analysis is one area of unsupervised machine learning of particular interest for data mining and knowledge discovery. Clustering techniques have been applied to medical problems for some time and there are many different algorithms available, all with very different performances and use cases [26], [27], [28], [29].

Cluster analysis provides the means for the organization of a collection of patterns into clusters based on the similarity between these patterns, where each pattern is represented as a vector in multidimensional space [30], [31].

In clustering schemes, data entities are usually represented as vectors of feature-value pairs. Features represent certain attributes of the entities that are known to be useful for the clustering task. In numeric clustering methods, a distance measure is used to find the dissimilarity between the instances [32]. The Euclidean distance is one of the common similarity measures and is defined as the square root of the squared discrepancies between two entities summed over all variables (i.e., features) measured. For any two entities A and B and k=2 features, say, X_1 and X_2, d_{ab} is the length of the hypotenuse of a right triangle. The square of the distance between the points representing A and B is obtained as follows:

$$d^2_{ab} = (X_{a1}-X_{b1})^2 + (X_{a2} - X_{b2})^2 \quad [1]$$

The square root of this expression is the distance between the two entities [33], [34].

In this study, we used the DBSCAN algorithm to analyze adverse events reports. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability. Any object not contained in a cluster is considered to be noise. The DBSCAN algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases, even those that contain noise. It defines a cluster as a maximal set of density-connected points. The basic principles of density-based clustering involve a number of definitions, as shown in the following:

- The neighborhood within a radius ε-neighborhood of the object.
- If the ε-neighborhood of an object contains at least a minimum number, *MinPts*, of objects, then the object is called a core object.

- Given a set of objects, D, we say that an object p is the directly density-reachable from object q if p is within the ε-neighborhood of q, and q is a core object.
- An object p is density-reachable from object q with respect to ε and *MinPts* in a set of objects, if there is a chain of objects $p_1,\ldots,p_n$, $p_1=q$ and $p_n=p$ such as p_{i+1} is directly density-reachable from p_i with respect to ε and *MinPts*, for $1 \leq i \leq n, p_i \in D$.

Density reachability is the transitive closure of direct density reachability, and this relationship is asymmetric. Only core objects are mutually density reachable. Density connectivity, however, is a symmetric relation.

DBSCAN searches for clusters by checking the ε-neighborhood of each point in the database. If the ε-neighborhood of a point p contains more than *MinPts*, a new cluster with p as a core object is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merger of some density-reachable clusters. The process terminates when no new point can be added to any cluster [37]. If a spatial index is used, the computational complexity of DBSCAN is $O(nlogn)$, where n is the number of database objects. Otherwise, it is $O(n^2)$. The algorithm is therefore sensitive to the user-defined parameters [38].

The DBSCAN algorithm was used to perform cluster analysis on the dataset. Table 1 show the attributes used in the dataset. Weka 3.6.8 was used for the analysis. Weka is a collection of machine learning algorithms for data mining tasks and is open source software. The software contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [38]. The application of the DBSCAN algorithm on the dataset generated 336 clusters (Fig. 3). Some of these are shown in Table 6. The results of the application of the DBSCAN algorithm when run in Weka is as follows:

```
Clustered data objects: 8592
Number of attributes: 6
Epsilon(ε): 0.9;   minPoints(MinPts) : 6
Number of generated  clusters: 336
Elapsed time: 34.97
```

5 Experimental Results and Discussion

We investigated the DrugBank database to get detailed information regarding erythromycin, which is shown in Table 2. The DrugBank database is a bioinformatics and cheminformatics resource that combines detailed drug data (i.e. chemical, pharmacological, and pharmaceutical data) with comprehensive drug target information (i.e. sequence, structure, and pathway data) [39]. In the database, each drug has a DrugCard that provides extensive information on the drug's properties.

The majority of the adverse event reports in the dataset are for males (55.8%) (Table 3) with an average age of 41.759 years (Table 1) [14]. The most frequent indication for erythromycin use was for ill-defined disorders, followed by rosacea, rhinitis allergic, and diabetes mellitus (Table 4). Oral use occurs with the highest frequency (Table 5).

The ten most frequent adverse events associated with erythromycin are shown in Fig 1. Oligohtdramnios is at the top of the list, followed by intra-uterine death, gestational diabetes, and C-reactive protein increase. Fig. 2 shows the graphical representation of the top ten co-occurrences of adverse event outcomes with erythromycin. According to Fig. 2, the most observed outcome is OT(Other) (47%), followed by HO(Hospitalization) (25.4%), DE(Death) (16%), LT(Life-Threatening) (5%), DS(Disability) (3%), RI(Required Intervention to Prevent Permanent Impairment/Damage) (1%), CA(Congenital Anomaly) (0.8%), and Unknown(0.01%) in this order.

Table 2. Erythromycin in the DrugBank database

Drugbank ID	DB00199
Drug name	Erythromycin
Some synonyms	Erythromycin oxime, EM, Erythrocin Stearate
Some brand names	Ak-mycin, Akne-Mycin, Benzamycin, Dotycin
Categories	Anti-Bacterial Agents, Macrolides
ATC Codes	D10AF02, J01FA01, S01AA17

Table 3. The number of reports by gender

Gender	The number of reports
Female	3792 (44.1%)
Male	4798(55.8%)
NS(Not Specified)	2(0.2%)

Table 4. Top ten indications for use

No	Indication for use	The number of co-occurrences (N)
1	Ill-defined disorder	2948(39%)
2	Rosacea	1887(25%)
3	Rhinitis allergic	579(7.7%)
4	Diabetes mellitus	528(7%)
5	Drug use for unknown indication	500(6%)
6	Lower respiratory tract infection	420(5%)
7	Infection	205(2%)
8	Prophylaxis	144(1.9%)
9	Enterobacter infection	126(1.6%)
10	Acne	101(1.3%)

Table 5. Route of drug administration

Route	The number of reports
Oral	7875(91%)
Transplacental	279(3%)
Ophthalmic	270(3%)
Intravenous	96(1%)
Topical	42(0.4%)
Parenteral	18(0.2%)
Disc, Nos	12(0.1%)

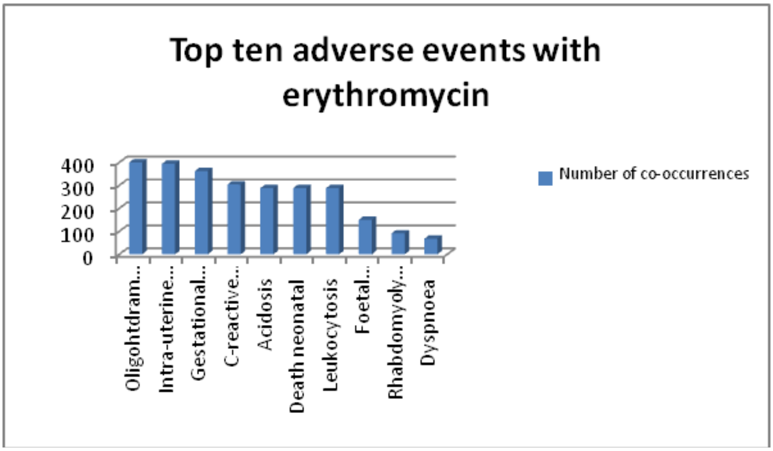


Fig. 1. The number of co-occurrences of adverse events (MedDRA terms) with erythromycin

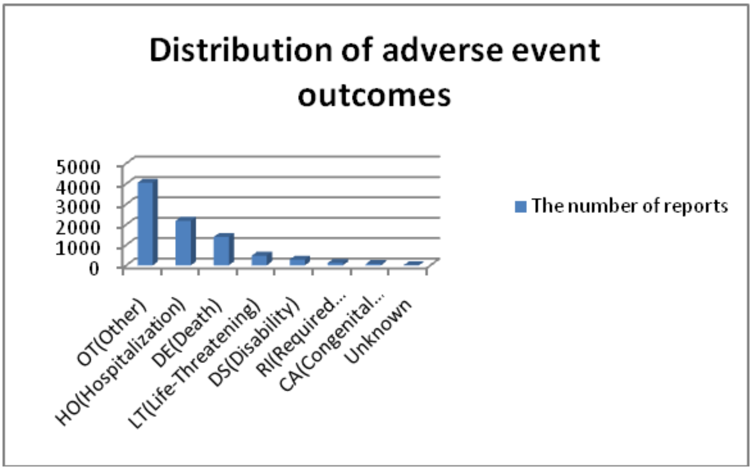
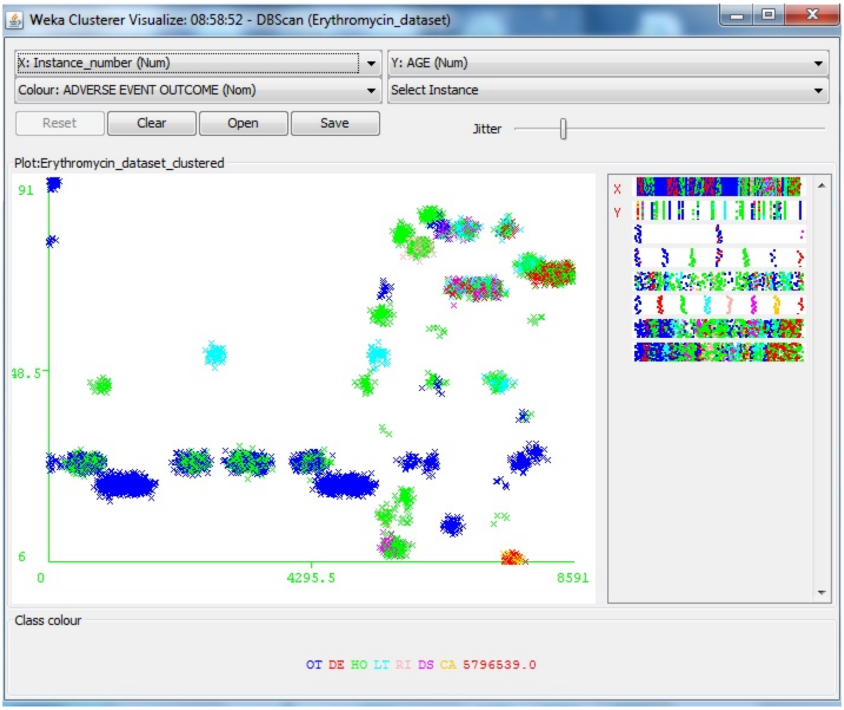


Fig. 2. The number of co-occurrences of adverse event outcomes with erythromycin



Weka Clusterer Visualize: 08:58:52 - DBScan (Erythromycin_dataset)

X: Instance_number (Num)

Y: AGE (Num)

Colour: ADVERSE EVENT OUTCOME (Nom)

Select Instance

Plot:Erythromycin_dataset_clustered

Fig. 3. Visual clusters of the DBSCAN algorithm on the erythromycin dataset

Table 6. Some clusters obtained by the DBSCAN algorithm

Attributes	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
Age	41.758	52	20	83
Gender	Male	Male	Female	Male
Route	Oral	Oral	Topical	Intravenous
Indication for use	Rosacea	Lower Respiratory Tract Infection	Acne	Pneumonia Primary Atypical
Adverse event outcome	Other	Life threatening	Hospitalization	Life threatening
Adverse event	Intra-Uterine Death	Cardiac Failure	Rash Pruritic	Weight Decreased

Erythromycin has some well-known adverse events such as vomiting, diarrhea, and mild skin rash [40]. According to our results, some events such as intra-uterine death, cardiac failure, rash pruritic, and weight decrease, are also seen in the clusters obtained by the DBSCAN algorithm. For example, intra-uterine death has a relationship with middle aged and male patients who are diagnosed with rosacea disease (cluster 1). In addition, young female patients form a cluster and the rash pruritic adverse event is seen with acne disease in the same cluster (cluster 3). Clinicians and researchers can search our results and perform clinical studies to find new hypotheses for the evaluation of drug safety of erythromycin.

The FDA's FAERS database is an important resource, but it has some limitations. For example, the database has many missing attribute values such as age and adverse events. We therefore omitted some records containing missing values. In addition, we faced some data quality and compatibility problems with the datasets created during different time periods. We therefore merged the datasets that covered the third quarter of 2005 through to the second quarter of 2012. Apart from the FDA's FAERS database, medical records that are created in hospital information systems are also an important resource for determining drug adverse events and their outcomes. Wang X et al analyzed narrative discharge summaries collected from the Clinical Information System at New York Presbyterian Hospital (NYPH). They applied MedLEE, a natural language processing system, to the collection in order to identify medication events and entities which could be potential adverse drug events. Co-occurrence statistics with adjusted volume tests were used to detect associations between the two types of entities, to calculate the strengths of the associations, and to determine their cutoff thresholds. Seven drugs/drug classes (ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone, and angiotensin-converting-enzyme inhibitors) with known ADEs were selected to evaluate the system [41]. Medical records can therefore be used to reveal any serious risks involving a drug in the future [25].

6 Conclusion

Pharmacovigilance aims to search for previously unknown patterns and automatically detect important signals, such as drug-associated adverse events, from large databases [17]. The FDA's FAERS is a large resource for pharmacovigilance and can be used to detect hidden relationships between drugs and adverse events. In this study, the adverse event profile for erythromycin was analyzed and a research study based on patient demographics, route for drug administration, indication for use, adverse events, and adverse event outcome relationships in the FAERS reports was carried out. Erythromycin is commonly used for the treatment of bacterial diseases and bacterial diseases are one of the most serious causes for health problems in the world. Therefore, the prevention and treatment of these diseases is an important research issue in the medical domain.

We analyzed FAERS reports through the use of computational methods, and subsequently applied the DBSCAN algorithm to the dataset in order to discover clusters. The clusters highlighted that patient demographics can have some relationships with certain adverse events and event outcomes of erythromycin use. Medical researchers must be made aware of these results and the information obtained in this study could lead to new research studies for the evaluation of erythromycin drug safety.

7 Open Problems

The FDA FAERS database offers a rich opportunity to discover novel post-market drug adverse events. However, the exploration of the FDA Adverse Event Reporting System's data by a wider scientific community is limited due to several factors.

Problem 1. FAERS data must be intensively preprocessed to be converted into analyzable and unified format [45]. While preprocessing is common for the effective machine learning analysis of any data, for complex medical datasets this can often require domain-specific medical expertise. This is especially true during, for example, the feature selection phase of data preprocessing. Open datasets, without proper preprocessing, can also be extremely large. Running times for quadratic machine learning algorithms can grow quickly, and when working with medical data that have been made available with no particular research question in mind, proper data preprocessing is especially important to reduce their size.

Problem 2. The data has some data quality issues. For example, the data has many missing attribute values such as age and adverse events. Missing data and noise are two hindrances to using machine learning methods on open data. Open data sets, while free and publicly available, mean no possibility of retroactive refinement by the authors. They must be taken as is, and cannot normally be expanded, refined, or corrected. In the case of medical data, open data is almost always de-identified, which—depending on the research question—can result in too much missing data to make it useful or usable. However, missing values and noise are a reality of any data analysis or collection process. Machine learning techniques and algorithms that are especially designed for data that contain missing values is an active area of research, and specific solutions have been developed in the past.

Problem 3. There are few existing methods and tools to access the data and improve hypothesis generation with respect to potential drug adverse event associations. Those that exist are usually based on limited techniques such as proportional reporting ratios and reporting adds ratios. A generalized method or piece of software for the analysis of adverse event data is not yet available. Whether such a generalized approach would even be feasible, considering for example the level of dataset fragmentation, is fertile ground for future research. With the numbers of datasets that are being made available constantly increasing, novel approaches to properly and more easily analyze this data are sure to increase alongside it.

8 Future Outlook

The FDA FAERS database is used to analyze the safety profiles of several drugs. A number of commercial tools, such as query engines, are now available to analyze the FDA FAERS. These tools provide free-text search query abilities that allow for the primary safety profile of drugs to be viewed. Other tools calculate the probability of an adverse event being associated with a drug. They also allows searching the FDA FAERS database by providing interpretable graphics for the adverse events reported over time, stratified by relevant category, ages, and gender, thus allowing for clinicians to quickly check drug safety information. This would be of benefit for the entire drug safety assessment [46]. However, these tools offer limited statistical techniques and data mining algorithms. Therefore, the automatic preprocessing of data, temporal analysis, and interactive data mining [47], [48] of drug adverse events through the use of state of the art data mining techniques is sorely needed. By increasing access to, and through the analysis of such drug-safety data new insights into ADEs will be discovered, but only when novel approaches in searching, mining, and analysis are discovered and implemented.

Acknowledgements. We would like to thank Ozgur Ilyas Ekmekci for help and Dr. Cinar Ceken for medical expert advice.

References

1. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H.: Linked Open Government Data: Lessons from Data.gov.uk. IEEE Intelligent Systems, 1541–1672 (2012)
2. Boulton, G., Rawlins, M., Vallance, P., Walport, M.: Science as a public enterprise: The case for open data. The Lancet 377, 1633–1635 (2011)
3. Rowen, L., Wong, G.K.S., Lane, R.P., Hood, L.: Intellectual property - Publication rights in the era of open data release policies. Science 289, 1881 (2000)
4. Thompson, M., Heneghan, C.: BMJ OPEN DATA CAMPAIGN We need to move the debate on open clinical trial data forward. British Medical Journal, 345 (2012)
5. Sakaeda, T., Tamon, A., Kadoyama, K., Okuno, Y.: Data mining of the Public Version of the FDA Adverse Event Reporting System. International Journal of Medical Sciences 10(7), 796–803 (2013)
6. Rodriguez, E.M., Staffa, J.A., Graham, D.J.: The role of databases in drug postmarketing surveillance. Pharmacoepidemiol Drug Saf. 10, 407–410 (2001)
7. Wysowski, D.K., Swartz, L.: Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: The importance of reporting suspected reactions. Arch Intern Med. 165, 1363–1369 (2005)
8. [Internet] (Internet) U.S. Food and Drug Administration (FDA), http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm
9. [Internet] (Internet) MedDRA MSSO, http://www.meddramsso.com/index.asp

10. Moore, T.J., Cohen, M.R., Furberg, C.D.: Serious adverse drug events reported to the Food and Drug Administration, 1998-2005. Arch. Intern. Med. 167, 1752–1759 (2007)
11. Weiss-Smith, S., Deshpande, G., Chung, S., et al.: The FDA drug safety surveillance program: Adverse event reporting trends. Arch. Intern. Med. 171, 591–593 (2011)
12. (Internet) Bacterial Infections, `http://www.who.int/vaccine_research/diseases/soa_bacterial/en/index4.html`
13. Manchia, M., Alda, M., Calkin, C.V.: Repeated erythromycin/codeine-induced psychotic mania. Clin. Neuropharmacol. 36(5), 177–178 (2013)
14. Varughese, C.A., Vakil, N.H., Phillips, K.M.: Antibiotic-associated diarrhea: A refresher on causes and possible prevention with probiotics–continuing education article. J. Pharm. Pract. 26(5), 476–482 (2013)
15. Chen, E.S., Hripcsak, G., Xu, H., Markatou, M., Friedman, C.: Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. J. Am. Med. Inf. Assoc. 15, 87–98 (2008)
16. 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. Biological & Pharmaceutical Bulletin 35(6), 967–970 (2012)
17. Malla, S., Banda, S., Bansal, D., Gudala, K.: Trabectedin related muscular and other adverse effects; data from public version of the FDA Adverse Event Reporting System. Internatial Journal of Medical and Pharmaceutical Sciences 03(07), 11–17 (2013)
18. Raschi, E., Poluzzi, E., Koci, A., Moretti, U., Sturkenboom, M., De Ponti, F.: Macrolides and Torsadogenic Risk: Emerging Issues from the FDA Pharmacovigilance Database. Journal of Pharmacovigilance 1(104), 1–4 (2013)
19. Harpaz, R., DuMouchel, W., Shah, N.H., Madigan, D., Ryan, P., Friedman, C.: Novel data-mining methodologies for adverse drug event discovery and analysis. Clin. Pharmacol. Ther. 91(6), 1010–1021 (2012)
20. Harpaz, R., Chase, H.S., Friedman, C.: Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics 11(suppl. 9), S7 (2010)
21. Harpaz, R., Perez, H., Chase, H.S., Rabadan, R., Hripcsak, G., Friedman, C.: Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clin. Pharmacol. Ther. 89(2), 243–250 (2011)
22. Vilar, S., Harpaz, R., Chase, H.S., Costanzi, S., Rabadan, R., Friedman, C.: Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: Application to rhabdomyolysis. J. Am. Med. Inform. Assoc. 18(suppl. 1) (December 2011)
23. Wang, X., Chase, H.S., Li, J., Hripcsak, G., Friedman, C.: Integrating heterogeneous knowledge sources to acquire executable drug-related knowledge. In: AMIA Annu. Symp. Proc. 2010, pp. 852–856 (2010)
24. Kadoyama, K., Miki, I., Tamura, T., Brown, J.B., Sakaeda, T., Okuno, Y.: Adverse Event Profiles of 5-Fluorouracil and Capecitabine: Data Mining of the Public Version of the FDA Adverse Event Reporting System, AERS, and Reproducibility of Clinical Observations. Int. J. Med. Sci. 9(1), 33–39 (2012)
25. Yildirim, P., Ekmekci, I.O., Holzinger, A.: On Knowledge Discovery in Open Medical Data on the Example of the FDA Drug Adverse Event Reporting System for Alendronate (Fosamax). In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 195–206. Springer, Heidelberg (2013)

26. Belciug, S., Gorunescu, F., Salem, A.B., Gorunescu, M.: Clustering-based approach for detecting breast cancer recurrence. Intelligent Systems Design and Applications (ISDA), 533–538 (2010)
27. Belciug, S., Gorunescu, F., Gorunescu, M., Salem, A.: Assessing performances of unsupervised and supervised neural networks in breast cancer detection. In: 7th International Conference on Informatics and Systems (INFOS), pp. 1–8 (2010)
28. Emmert-Streib, F., de Matos Simoes, R., Glazko, G., McDade, S., Haibe-Kains, B., Holzinger, A., Dehmer, M., Campbell, F.: Functional and genetic analysis of the colon cancer network. BMC Bioinformatics 15(suppl. 6), S6 (2014)
29. Yildirim, P., Majnaric, L., Ekmekci, O., Holzinger, A.: Knowledge discovery of drug data on the example of adverse reaction prediction. BMC Bioinformatics 15(suppl. 6), S7 (2014)
30. Reynolds, A.P., Richards, G., de la Iglesia, B., Rayward-Smith, V.J.: Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5, 475–504 (2006)
31. Yıldırım, P., Ceken, K., Saka, O.: Discovering similarities fort he treatments of liver specific parasites. In: Proceedings of the Federated Conference on Computer Science and Information Systems, FedCSIS 2011, Szczecin, Poland, September 18-21, pp. 165–168. IEEE Xplore (2011) ISBN 978-83-60810-22-4
32. Holland, S.M.: Cluster Analysis, Department of Geology, University of Georgia, Athens, GA 30602-2501 (2006)
33. Hammouda, K., Kamel, M.: Data Mining using Conceptual Clustering, SYDE 622: Machine Intelligence, Course Project (2000)
34. Beckstead, J.W.: Using Hierarchical Cluster Analysis in Nursing Research. Western Journal of Nursing Research 24, 307–319 (2002)
35. Yıldırım, P., Ceken, C., Hassanpour, R., Tolun, M.R.: Prediction of Similarities among Rheumatic Diseases. Journal of Medical Systems 36(3), 1485–1490 (2012)
36. Han, J., Micheline, K.: Data mining: concepts and techniques. Morgan Kaufmann (2001)
37. Yang, C., Wang, F., Huang, B.: Internet Traffic Classification Using DBSCAN. In: 2009 WASE International Conference on Information Engineering, pp. 163–166 (2009)
38. Hall, M., Frank, E., Holmes, G., Pfahringe, B., Reutemann, P., Witten, I.E.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11, 1 (2009)
39. [Internet] (internet) Drugbank, `http://www.drugbank.ca`
40. [Internet] (internet) Erythromycin, `http://www.drugs.com`
41. Wang, X., Hripcsak, G., Markatou, M., Friedman, C.: Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. Journal of the American Medical Informatics Association 16(3), 328–337 (2009)
42. [Internet] (internet) Bacterial infections,
`http://www.nlm.nih.gov/medlineplus/bacterialinfections.html`
43. Edwards, I.R., Aronson, J.K.: Adverse drug reactions: Definitions, diagnosis, and management. Lancet 356(9237), 1255–1259 (2000)
44. [Internet] (internet) Open Knowledge Foundation,Open data introduction,
`http://okfn.org/opendata/`
45. [Internet] (internet) `http://www.chemoprofiling.org/AERS/t1.html`
46. Poluzzi, E., Raschi, E., Piccinni, C., De Ponti, F.: Data mining techniques in Pharmacovigilance: Intech, Open Science (2012)

47. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)
48. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual Data Mining: Effective Exploration ofthe Biological Universe. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. LNCS, vol. 8401, pp. 19–33. Springer, Heidelberg (2014)