Mining Patterns of Adverse Events Using Aggregated Clinical Trial Results

Zhihui Luo, PhD, Guo-Qiang Zhang, PhD, Rong Xu, PhD
School of Medicine, Case Western Reserve University

Abstract
Adverse event reports contain the most important metrics for evaluating the hazards, harms, and risks of a clinical intervention. In this paper, we present an exploratory study of discovering internal association patterns between adverse events. By taking advantage of the published trials reports on ClinicalTrials.gov, we developed an automatic pipeline to create a Clinical Trial Adverse Event Database (cTAED), which currently stores 4,317 clinical trial reports and 11,362 adverse events. The association mining algorithm FP-Tree was applied to the cTAED data to discover patterns between adverse events. We extracted 29,546 patterns and further examined association patterns related to patients’ deaths. The mined results indicate the existence of strong internal association patterns between adverse events. The evaluation results show that the p-value of confidence is smaller than 0.001, which indicates that our method mined association patterns with significantly more confidence than randomly-associated adverse events.

1. Introduction
Adverse events are harmful medical conditions or side effects that occur in a patient resulting from medications or interventions. Adverse events pose significant threats to the health and safety of patients. Over 770,000 patients were injured from adverse drug events each year. The negative results of adverse events also lead to the increase cost of medical care and compensation. The estimated cost to treat patients’ adverse events in national hospitals has quickly risen from $5.6 billion in 1997 to $37.6 billion in 2011.

The underlying mechanisms of the majority of adverse events remain poorly understood due to the complex influential factors of medicines and the complicated pathological and physiological status of patients. Surveillance and reporting of adverse events are crucial steps in evaluating the safety of a drug. Most of the adverse events are collected via the spontaneous reports of the FDA and from clinical trials. Spontaneous report is normally classified as post-marketing surveillance, while clinical trials report is normally a pre-marketing risk assessment. In this paper, we focus on analyzing the adverse event reports of clinical trials. We hypothesize that adverse events have internal correlations, a clinical characteristic similar to disease comorbidities. Our goal is to explore the feasibility of extracting internal association patterns of adverse events from clinical trial reports. The association patterns not only reveal potential correlation between adverse events, but also serve as fundamental elements for other pharmacovigilance research, such as adverse event prediction and prevention.

Our analysis is built upon the result reports available at ClinicalTrial.gov. ClinicalTrials.gov is a public registry of federally and privately sponsored clinical trials. Since 2007, the use of this registry of clinical trial summaries has been mandated by the FDA. Trial results are required to be reported within one year of competition using the Protocol Registration System. The adverse event report module is designed to summarize the incidence of the adverse events occurring in a clinical trial study. Analytic research of adverse events is a very challenging aspect of clinical trials studies. We hypothesize that there are internal association patterns among adverse events in clinical trials. Our goal is twofold: (1) to develop an automatic pipeline to construct an adverse event database using the reported results from ClinicalTrials.gov; and (2) to explore potential association relationships among the reported adverse events in clinical trials. As far as we know, this is the first research focusing on discovering internal patterns among adverse events using clinical trial data. Identification of association patterns among adverse events could also be very useful for improving the safety of clinical trials, such as predicting severe adverse events.

2. Method
First, we downloaded result reports from ClinicalTrials.gov. Result reports are submitted to ClinicalTrials.gov upon completion of trials; they summarize the observed outcome of a clinical trial, including adverse events. In this paper, we keep close focus on mining the adverse event data in the result reports, especially the association patterns among different adverse events. We developed a program to harvest all the documented trials from ClinicalTrials.gov. In total, 128,684 clinical trial reports were collected by the program. We further filtered in 4,317 clinical trials for which final results had been reported. As of this time, many trials do not contain a final result report due to two reasons: (1) the registration of results is only required for completed trials related to FDAAA 801 regulation; (2) the result database was just installed in 2008. Many trials are still recruiting patients or performing ongoing research; hence they are not yet come to the stage of submitting their final results.

A Java parser was developed to extract adverse events from the result reports. Due to the inconsistent terminology across different research institutions, we had to further harmonize the extracted adverse event terms by mapping them to UMLS concepts. More than 6 terminology standards were used throughout the reports, including...
MedDRA, WHOART, COSTART, NCI-CTCAE, ICD, and SNOMED-CT. The heterogeneous terminology standards and versions present a challenge for cross-institutional study. To address this problem, a named entity recognition method was developed to align recognizable terms to their corresponding standardized Unified Medical Language System (UMLS) concepts. For example, all of the three terms - *Dysentery, Infectious Diarrhea, and Diarrhoea Infections* - can be mapped to the unique concept C0013360. The concept-alignment procedure eliminated the variations of synonyms and acronyms, which significantly increased the cross-institutional comparability of clinical trial reports. All the extracted information was stored in our Clinical Trial Adverse Event Database (cTAED) which enables us to perform various large-scale analyses and provides convenient data retrieval and update. This is the first goal of this paper. **Table 1** in the results section shows data summary of cTAED.

The second goal of this paper is to explore the possibility of discovering association patterns among adverse events in the clinical trial reports. Pattern mining is a solid approach for discovering interesting relations between variables in a dataset. However, due to the large number of adverse events extracted from ClinicalTrials.gov, it is computationally impractical to exhaust all possible association patterns of adverse events. To overcome this challenge, we use the FP-Tree algorithm to discover strong association patterns.

The FP-Tree algorithm provides an effective and scalable method for mining frequent patterns using a prefix-tree to grow the patterns from smaller fragments. The fundamental concept of this method is the use of a frequent pattern tree to store the associated variables. FP-Tree removes low frequency variables and stores duplicate variables in a compacted data structure. These computational features make FP-Tree very suitable for mining adverse event patterns because there are large numbers of duplicate adverse events in the trial reports. Furthermore, it has been demonstrated that FP-Tree outperforms other popular pattern mining methods, such as the Apriori algorithm and the Tree Projection algorithm. In our implementation, the minimum support parameter of 0.0023 was used, which means any candidate adverse events should appear at least 10 times in the 4,317 trials in order to be included for the pattern mining.

The result of the FP-Tree algorithm is a set of patterns that tells how the adverse events are associated with each other. The pattern between two sets of adverse events is defined by the implication of the form \( X \Rightarrow Y \). \( Y \) is normally called the **pattern head** and \( X \) is called the **pattern body**. For example, the association pattern \( (\text{Anxiety, Fatigue}) \Rightarrow (\text{Insomnia}) \) indicates that when a patient has adverse events Anxiety and Fatigue together, the patient is very likely to have Insomnia. Such information can be used as the basis for decision support of adverse event prediction and prevention.

### 3. Results and Evaluation

#### 3.1. Results of the cTAED database

**Table 1: Clinical Trial Adverse Event Database (cTAED)**

| Data Table          | Data Examples                                           | Total Data Volume (Unit) |
|---------------------|---------------------------------------------------------|--------------------------|
| Adverse Events      | Cough, Sleep disorder, Diarrhoea                       | 11,362 (events)          |
| Clinical Trials     | NCT00403754, NCT00594464, NCT00831701                  | 4,317 (trials)           |
| Disease Conditions  | Asthma, Parkinson's Disease, Ureteral Calculi           | 1,973 (conditions)       |
| Interventions       | Indacaterol, Rotigotine, Tamsulosin                     | 5,447 (interventions)    |

**Table 1** summarizes the current status of the cTAED database. The adverse event table is the core of the database. It contains the event name, UMLS concept of the event, number of affected subjects, number of at risk subjects and event categorization. Clinical trial, intervention and disease condition tables are linked to the adverse event table through their unique reference keys. cTAED not only stores the adverse event data in a structured format, but also provides easy access to the data elements that can be used for data mining.

#### 3.2. Pattern Mining Results

The result patterns are ranked by the **Lift score** (also called **interest**), which is a measure of the strength of an association pattern. In particular, the lift score measures how often pattern body \( X \) and pattern head \( Y \) occur together more frequently than expected if they were statically independent. The value of lift can be defined as: Lift \( (X \Rightarrow Y) = \frac{P(X \text{ and } Y)}{P(X) \cdot P(Y)} \). Lift scores larger than 1 indicate positive association between \( X \) and \( Y \). The higher the lift score, the stronger the association of the pattern. In this paper, we set the minimum lift to 4, which extracts patterns with strong association. We also restrict the pattern head containing one element to speed up the mining procedure and analysis. The pattern body has no restriction. As a result, we mined 29,546 patterns from the dataset described in the previous section. We report a set of sample patterns generated by the proposed data mining approach. **Table 2**
shows some examples of the top ranked patterns. The patterns indicate that when the adverse events in the pattern body occur together, the event in pattern head is expected to be observed with high likelihood.

The results listed in Table 2 show the existence of different types of associations. For example, pattern 1 in Table 2 indicates that the adverse events Injection Site Pain and Injection Site Swelling are both common early symptoms of Injection Site Erythema. They all belong to Injection Site Reactions (ISRs) that normally happen in immunization trials among patients taking experimental vaccines. Pattern 2 shows co-occurring laboratory test results. All three laboratory tests are related to liver functionality. It is very interesting to see the increased results of the three tests were associated in the clinical trials. Pattern 3 conforms the reasonable assumption that Pain and Decreased Appetite together would be very likely lead to Weight Decrease. Similarly, as pattern 6 suggests, Anxiety and Depression together might lead to Dysgeusia. Pattern 8 indicates that when Atrial Fibrillation complicates Myocardial Infarction, they might lead to the more severe event Congestive Cardiac Failure, an association that has been confirmed and analyzed by Schmitt et al. Pattern 9 shows Headache and Nausea together as common early signs of Vomiting. These results demonstrate that the top-ranked patterns are meaningful association in clinical trial studies.

Table 2: Example Results of Top-Ranked Patterns

| Rank | Pattern Body | Pattern Head | Lift |
|------|--------------|--------------|------|
| 1    | Injection Site pain, Injection Site Swelling | Injection Site Erythema | 13.28 |
| 2    | Alanine Aminotransferase Increased, Blood Alkaline Phosphatase Increased | Aspartate Aminotransferase Increased | 11.74 |
| 3    | Decreased Appetite, Pain | Weight Decreased | 11.10 |
| 4    | Febrile Neutropenia, Neutropenia | Pneumonia | 9.71 |
| 5    | Dizziness, Alopecia | Hypertension | 7.44 |
| 6    | Anxiety, Depression | Dysgeusia | 7.09 |
| 7    | Pyrexia, Sepsis | Deep Vein Thrombosis | 6.37 |
| 8    | Myocardial Infarction, Atrial Fibrillation | Congestive Cardiac failure | 5.74 |
| 9    | Headache, Nausea | Vomiting | 5.57 |
| 10   | Pain in extremity, Mucosal Inflammation, Neutropenia | Oedema Peripheral | 5.34 |

To provide more specific analysis of the mining results, we further analyzed the association patterns of the most severe adverse event – patient’s death. We selected Death as the pattern head and mined other adverse events associated with a patient’s death. The top results are shown in Table 3. As can be seen, vital organ function failure complicated with other serious adverse events are the most common patterns associated with the death of patients. For example, (see Table 3) pattern 1 shows that Cardiac Failure and Convulsion together might result in Death. Similar patterns are shown in examples 2 and 3. Acute inflammation of important organs may also lead to a patient’s death. For example, Bronchopneumonia is the acute inflammation of the walls of the bronchioles, and sepsis triggers whole-body inflammatory reaction which is potentially very deadly. Pattern 6 and pattern 8 show that internal bleedings are another set of dangerous adverse events associated with patient’s death, such as Gastrointestinal Haemorrhage (bleeding in stomach or intestine) and Cerebral Haemorrhage (bleeding in brain tissue). These patterns indicate that there are strong connections between adverse events. Some adverse events may quickly deteriorate to more severe events or even lethal events.

A large-scale dedicated study of fatal adverse event in 11,015 clinical facilities reported a major finding of five most common lethal events, all of which can be closely matched to our mining results (see the * markers in Table 3). We also find our results broadly match to the WHO “Top 10 causes of death” fact sheet of the high income group, except for cancer, diabetes and HIV, which are rarely reported as adverse events. The mined results show good agreement with these previous studies. Comparing to these statistical analysis approaches, our data mining method has two advantages: (1) the mining method can be flexibly applied to discover any strong pairwise or multi-item associations among the data without a hypothetical presumption. We can easily filter in any interested events
for analysis, such as exploring the patterns related to Death, Cardiac Failure, or Pneumonia events; (2) the event patterns provide much stronger predictability than a single event. For example in pattern 2, we calculated that when Renal Failure and Osteoarthritis occurred independently, 34% and 33% of the trials respectively observed death of patients; while when Renal Failure and Osteoarthritis occurred together, the ratio increased to 57%. Trials with patient’s death significantly increased over 70% when the two seemingly unrelated events occurred together. Scientists have suspected the existence of connection between Renal Failure and Osteoarthritis and their destruction to patients’ health when analyzing the cases of two patients. Our method not only captured the phenomenon in the adverse event reports, but also provided data evidences to support the association. First, the high lift score of 5.81 shown in the pattern indicates that the events are very unlikely to be independent. Second, the death of patients observed in 586 of 1030 trials (57%) that have two events occurred together. The observation is supported by a large number of cases.

1.3. Confidence Evaluation of Patterns

Confidence is a common indicator for evaluating how reliable an association relationship is. The value of confidence is between [0, 1] and it can be calculated from formula: \( \text{Confidence}(X \Rightarrow Y) = P(Y|X) \).

We compared the confidence of randomly associated adverse events with the mined association patterns given the same event head (see Table 4). Five different adverse events were evaluated, including Death, Cardiac Failure, Depression, Nausea, and Hypersensitivity. For each adverse event, we randomly selected 10 different adverse events that co-occurred with the selected event and evaluated their association confidence as compared to the confidence of 10 randomly-mined patterns.

The results show that the mined association patterns were significantly (p-value<0.01, t-test, two tails) more confident than randomly associated adverse events. The average confidence of association patterns is also much higher than that of randomly associated events.

4. Discussion and Future Work

This study presents a new analytic pipeline for mining association relationships between adverse events in clinical trials. Our method extracts strong association patterns of adverse events, which could be used to improve the predictability of adverse events and to complement existing event observation. Califf et al. recently reported that the majority (62%) of the trials registered on ClinicalTrials.gov have less than 100 recruited participants. These small trials typically reported fewer adverse events and were less informative than larger trials because of the small participant size. Our research provides a cross-institutional method for analyzing adverse events, which could be used to enrich the observed adverse events of small clinical trials by allowing researchers to see a wider range of associated events.

From the high-level point of view, the mined association pattern can be classified into two types based on the temporal relationship. The first type is co-occurrence association, which is very similar to comorbidity in that a patient presents one or more symptoms or disorders in addition to the observed disorders. For example, the pattern (Dizziness, Headache) \(\Rightarrow\) (Pain) shows the common co-occurrence of symptoms for a patient in pain. The second type of association is sequential, which indicates that some adverse events might happen before others. For example, the pattern (Cardiac Failure, Convulsion) \(\Rightarrow\) (Death) indicates that cardiac failure and convulsion events might lead to the death of a patient. The event pattern helps researchers and clinicians to predict the development of adverse events, which could prevent the deterioration of a patient’s condition in a trial.

In this study, we use the clinical trials report as our data source, which has several distinctive characteristic that might help improve the mining results. First, the patients participating in the clinical trials were carefully screened through very strict eligibility criteria, which helps reduce confronting factors that might obscure the adverse event association. Second, the reported events have been scrutinized by researchers and administrators to reduce potential errors. Third, standardized terminologies have been used to encode most of the reported events in clinical trial reports. Although the terminology sources might not be consistent across different research institutions, they certainly reduce the potential burden of parsing adverse events from other data sources (e.g. clinical notes, discharge summaries), which require sophisticated natural language processing (NLP) procedures and are less precise. Our method can be customized to adapt other data sources such as spontaneous adverse event reports.

| Event Head      | Avg. Confidence of Random Association | Avg. Confidence of Association Pattern | P-Value |
|-----------------|---------------------------------------|---------------------------------------|---------|
| Death           | 0.077                                 | 0.317                                 | 3.55E-8 |
| Cardiac Failure | 0.122                                 | 0.395                                 | 6.69E-6 |
| Depression      | 0.096                                 | 0.352                                 | 2.23E-4 |
| Nausea          | 0.116                                 | 0.387                                 | 5.74E-3 |
| Hypersensitivity| 0.107                                 | 0.402                                 | 6.31E-4 |
There are two major limitations of this research. First, due to the nature of association pattern mining and the characteristics of the data source, mining results only reveal general correlation relationships between adverse events. The implied clinical meanings of the mined patterns still need to be interpreted by domain experts. The example results in Table 2 have clearly shown different types of associations. One possible solution for this problem is to use classification algorithms to categorize the mined association rules. Liu et al. have published a method to integrate classification into association mining. We plan to apply this method into current mining procedure. The second limitation is that we did not make use of the drug information provided in the clinical trials. In this paper, we only focus on internal association of adverse events. Adding the drug information to analysis, we might be able to further stratify the association analysis to different interventions. The association between drugs and event patterns could be very valuable for pharmacovigilance research. This is our next goal in adverse event analysis.

5. Conclusion

We developed an automatic pipeline for creating an adverse event database from clinical trials reports. The FP-Tree algorithm was customized to mine association patterns between adverse events. We discovered that strong association patterns exist between adverse events. The evaluation results show that the adverse event patterns are significantly more confident than randomly associated events. These patterns could be used to help predict adverse events and to improve the safety of patients.

References

1. Cullen DJ, Sweitzer BJ, Bates DW, Burdick E, Edmondson A, Leape LL. Preventable adverse drug events in hospitalized patients: A comparative study of intensive care and general care units. *Critical Care Medicine.* 1997;25(8):1289-1297.

2. Bates D, Spell N, Cullen D, et al. The costs of adverse drug events in hospitalized patients. *JAMA: The Journal of the American Medical Association.* 1997;277(4):307-311.

3. Van Den Bos J, Rustagi K, Gray T, Halford M, Ziemkiewicz E, Shreve J. The $17.1$ billion problem: the annual cost of measurable medical errors. *Health Affairs.* 2011;30(4):596-603.

4. Wysowski D, Swartz L. Adverse drug event surveillance and drug withdrawals in the united states, 1969-2002: The importance of reporting suspected reactions. *Archives of Internal Medicine.* 2005;165(12):1363-1369.

5. FDA. Adverse Event Reporting System (AERS). 2012.

6. NLM. About ClinicalTrials.gov. 2012; http://www.clinicaltrials.gov/ct2/info/about.

7. Kessler RC, Merikangas KR. The National Comorbidity Survey Replication (NCS - R): background and aims. *International journal of methods in psychiatric research.* 2006;13(2):60-68.

8. Groves T. Mandatory disclosure of trial results for drugs and devices. *BMJ: British Medical Journal.* 2008;336(7637):170.

9. ClinicalTrials.gov. Adverse Event Reporting in the Protocol Registration System of ClinicalTrials.gov. 2009.

10. Friedman L, Furberg C, DeMets D. *Fundamentals of Clinical Trials:* Springer New York; 2010.

11. Harpaz R, DuMouchel W, Shah N, Madigan D, Ryan P, Friedman C. Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics.* 2012;91(6):1010-1021.

12. Luo Z, Johnson SB, Weng C. Semi-Automatic Induction of Semantic Classes from Free-Text Clinical Research Eligibility Criteria Using UMLS. *American Medical Informatics Association Annual Symposium; Washington, DC.* 2010. 487-491.

13. Luo Z, Zhang GQ, Teagno J, Xu R. Large-Scale Analysis of Serious Adverse Events Reporting in Clinical Trials Funded by For-Profit and Not-For-Profit Agencies. AMIA Joint Summits on Translational Science; San Francisco. 2013.

14. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. *Al Magazine.* 1996;17(3):37-54.

15. Han J, Pei J, Yin Y, Mao R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Min. Knowl. Discov.* 2004;8(1):53-87.

16. Luo Z, Zhang GQ, Teagno J, Xu R. Large-Scale Analysis of Serious Adverse Events Reporting in Clinical Trials Funded by For-Profit and Not-For-Profit Agencies. AMIA Joint Summits on Translational Science; San Francisco. 2013.

17. Luo Z, Zhang GQ, Teagno J, Xu R. Large-Scale Analysis of Serious Adverse Events Reporting in Clinical Trials Funded by For-Profit and Not-For-Profit Agencies. AMIA Joint Summits on Translational Science; San Francisco. 2013.

18. Han J, Pei J, Yin Y, Mao R. Mining Frequent Patterns without Candidate Generation. Proceedings of the 2000 ACM SIGMOD international conference on Management of data; Dallas, Texas, United States. 2000. 1-12.

19. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases. 1994. 487-499.

20. Agrawal RC, Aggarwal CC, Prasad VVV. A tree projection algorithm for generation of frequent item sets. *J. Parallel Distrib. Comput.* 2001;61(3):350-371.

21. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec.* 1996;25(2):255-264.

22. McMahon BJ, Helminiak C, Wainwright RB, Bulkow L, Trimble BA, Wainwright K. Frequency of adverse reactions to hepatitis B vaccine in 43,618 persons. *The American Journal of Medicine.* 1992;92(3):254-256.

23. Schmitt J, Duray G, Gersh BJ, Hohnloser SH. Atrial fibrillation in acute myocardial infarction: a systematic review of the incidence, clinical features and prognostic implications. *European Heart Journal.* May 1, 2009 2009;30(9):1038-1045.

24. Astiz ME, Rackow EC. Septic shock. *The Lancet.* 1998;351(9114):1501-1505.

25. Wester K, Jönsson AK, Spigset O, Druid H, Hägg S. Incidence of fatal adverse drug reactions: a population based study. *British Journal of Clinical Pharmacology.* 2008;65(4):573-579.

26. WHO. Top 10 Causes of Death. http://www.who.int/gho/mortality_burden_disease/causes_death_2008/en/index.html. 2008.

27. Duncan JI, Hurst NP, Disney A, Sebbren B, Milazzo SC. Is chronic renal failure a risk factor for the development of erosive osteoarthritis? *Annals of the Rheumatic Diseases.* March 1, 1989 1989;48(3):183-187.

28. Calif R, Zarin D, Kramer J, Sherman R, Aberle L, Tasseem A. Characteristics of clinical trials registered in clinicaltrials.gov, 2007-2010. *JAMA: The Journal of the American Medical Association.* 2012;307(17):1838-1847.

29. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. KDD-98 Proceedings. 1998.