# An Early Warning System for Unrecognized Drug Side Effects Discovery

Hao Wu and Hui Fang
Department of Electrical and Computer Engineering
University of Delaware, USA
{haow,hfang}@udel.edu

Steven J. Stanhope
Department of Kinesiology and Applied Physiology
University of Delaware, USA
stanhope@udel.edu

## ABSTRACT

Drugs can treat human diseases through chemical interactions between the ingredients and intended targets in the human body. However, the ingredients could unexpectedly interact with off-targets, which may cause adverse drug side effects. Notifying patients and physicians of potential drug effects is an important step in improving healthcare quality and delivery. With the increasing popularity of Web 2.0 applications, more and more patients start discussing drug side effects in many online sources. In this paper, we describe our efforts on building *UDWarning*, a novel early warning system for unrecognized drug side effects discovery based on the text information gathered from the Internet. The system can automatically build a knowledge base for drug side effects by integrating the information related to drug side effects from different sources. It can also monitor the online information about drugs and discover possible unrecognized drug side effects. Our demonstration will show that the system has the potentials to expedite the discovery process of unrecognized drug side effects and to improve the quality of healthcare.

## Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services - Web-based services

## General Terms

Algorithm, Design

## Keywords

social media, discuss forum, drug side effects

## 1. INTRODUCTION

Each drug has both benefits and risks, because it could interact with “off-targets” in addition to the intended targets. The interaction between a drug and its intended targets could treat the diseases associated with the targets, while the interaction with “off-targets” may make drugs less effective or even cause dangerous side effects such as heart failure. Physicians and patients often need to know possible drug side effects in order to reduce serious accidents resulting from adverse drug-related events.

In the pre-market situation, some side effects of a drug can be recognized in the pre-clinical and clinical trial data. Unfortunately, not all the side effects can be discovered in the lab test and small clinical trials. Instead, after the drug is approved by FDA and enters the market, all the patients on the market end up being part of a large (post-clinical) experiment to identify unrecognized and unexpected drug side effects. Currently, in such post-market situation, drug side effects often come from the reports submitted by physicians based on the information gathered from their patients through MedWatch systems. If the unexpected side effects are dangerous or fatal, FDA or the drug company may decide to withdraw the drug from the market. This process might take up to a few years. Ideally, if we could recognize possible drug side effects and notify FDA and the drug company much earlier, it would be possible to reduce the number of adverse events caused by the drug.

Recently, with the increasing popularity of Web 2.0 technology, more and more patients start sharing their experiences and discussing drug side effects on various online web sites prior to their doctor visits. Such online information forms a valuable resource of drug side effects and makes it possible to discover unrecognized drug side effects much earlier. Unfortunately, such information is currently underutilized.

In this project, we develop *UDWarning*, a novel early warning system to detect unrecognized drug side effects. Figure 1 shows the system architecture. The first component is to create a knowledge base for drug side effects. In particular, we automatically integrate different online sources about the drug side effects, and construct a knowledge base that hopefully includes information of all drugs and their known side effects. The knowledge base enables us to analyze side effects and discover related drug side effects. In order to discover unrecognized drug side effects, we first need to crawl online discussions about drugs and then aggregate the information about a side effect of a drug in order to detect whether there are any potential unrecognized drug side effects. The target users of the developed system are drug companies and FDA, and we aim to notify them of possible unrecognized adverse side effects at a much earlier stage. *UDWarning* is expected to expedite the process of discovering unrecognized drug side effects.

## 2. SYSTEM DESCRIPTION

We now describe the details of *UDWarning* system, an early warning system for unrecognized drug side effects.

*WWW 2012 Companion*, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

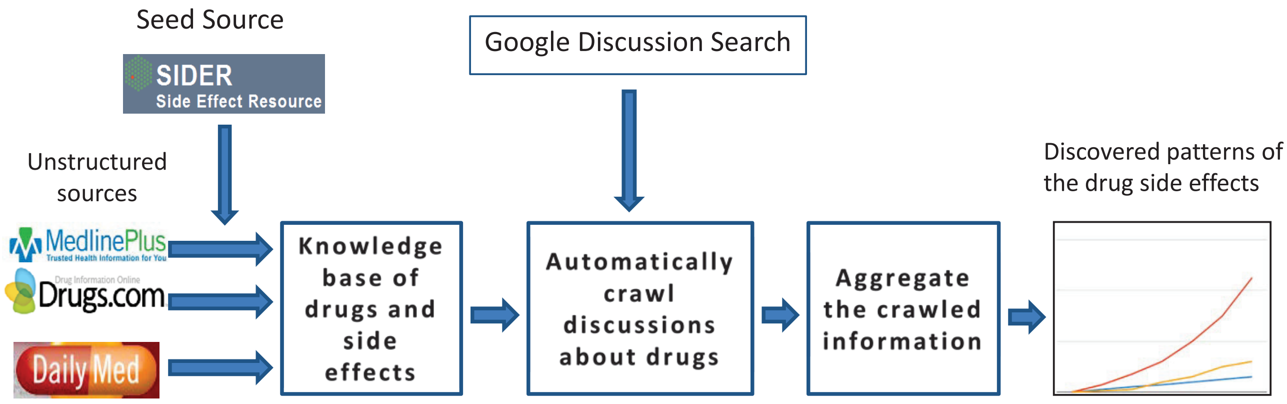
Seed Source
SIDER
Side Effect Resource
Unstructured sources
MedlinePlus
Trusted Health Information for You
Drug Information Online
Drugs.com
Daily Med
Knowledge base of drugs and side effects
Google Discussion Search
Automatically crawl discussions about drugs
Aggregate the crawled information
Discovered patterns of the drug side effects

**Figure 1: System Architecture**

| | *SIDER* | *MedlinePlus* | *Drugs.com* | *DailyMed* | **Our Knowledge Base** |
|---|---|---|---|---|---|
| # of drugs | 4,953 | 1,082 | 11,652 | 12,865 | **15,848** |
| avg. # of side effects per drug | 65 | 17 | 47 | 28 | **66** |

**Table 1: Statistics of Different Drug Side Effect Sources**

## 2.1 Constructing a Knowledge Base for Drug Side Effects

Since our goal is to discover unrecognized side effects of drugs, the first challenge is how to build a knowledge base that includes all of the known drug side effects. Specifically, the knowledge base includes all the marketed drugs as well as their associated side effects.

Many online sources provide information related to drug side effects. For example, *SIDER* [1] contains extracted drug side effects from public documents and provides the information in a well-structured format. *DailyMed* [2] provides high quality information about drugs approved by FDA including FDA labels. *Drugs.com* [3] is one of the most visited drug-related web sites. *MedlinePlus* [4] is a web site created by National Institutes of Health for patients to access medical related information. By comparing the information from these four sources, we find that none of them contain all the drug-related information. Thus, it is necessary to integrate the information from all these sources to construct a more complete knowledge base.

Among all the four sources we consider in the paper, only *SIDER* provides structured information that makes it possible to directly extract drug names and side effects. Unfortunately, the other three sources are unstructured, so it is more challenging to extract drug names and side effects from them. We notice that most pages from these three sources are organized based on drugs. Every page discusses the information of a single drug, and drug names are often mentioned in specific fields such as "title", "drug" or "drug name" in the HTML pages. Thus, a simple yet effective drug name extraction strategy is to utilize the HTML template of each web source, identify the field related to drug names, and use these field values as drug names.

Unlike drug names that are often the values of specific fields, side effect names are often scattered in the plain text with noisy terms such as drug descriptions or drug labels. Thus, the drug name extraction method described above would not work well for side effect name extraction. Fortunately, the structured information from *SIDER* provides a list of side effect names, so we propose to use these names to help us extract side effect names from other three sources. The assumption is that the set of side effects tends to be stable and most side effects are covered in the *SIDER* database. Specifically, instead of using only exact matching for side effect names, we stem the terms and allow non-exact matching. This strategy would allow us to identify variants of a side effect such as "lung cancer" and "cancer of lung".

After extracting drugs and side effects from the four different sources, we may construct a knowledge base of drug side effects with a list of drugs and their associated side effects. Table 1 shows the statistics of the four sources and our integrated knowledge base. It is clear that our knowledge base contains a more complete set of information about drug side effects.

## 2.2 Crawling Information Related to Drugs

With the development of Web 2.0 technology, patients start discussing side effects of the drugs that they have taken on different online forums. To detect unknown side effects from these discussions, it is necessary to crawl as much drug-related information as possible. One possible strategy is to start with a list of health-related blogs and crawl information from these web sites. However, the limitation is that the selection of the seed blogs are important and the coverage of these blogs is often limited. Thus, we explore another

[1] http://sideeffects.embl.de/
[2] http://dailymed.nlm.nih.gov/dailymed/
[3] http://www.drugs.com
[4] http://www.nlm.nih.gov/medlineplus/

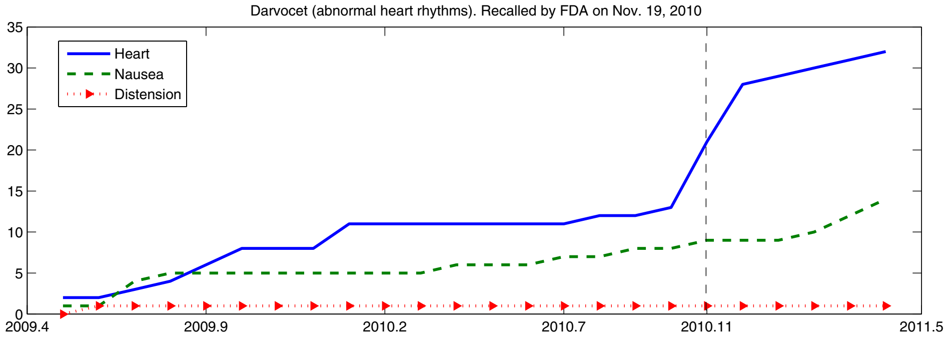

**Figure 2: Darvocet was recalled on Nov. 19, 2010, because it put patients at risk of "fatal heart rhythm abnormalities". "Nausea" is a known side effect of Darvocet while "Distension" is not**

| |
|---|
| Cerebral infarct; Status epileptics;<br>Generalized seizure; Seizure disorder; |
| Allergic reaction; Tongue pain |
| Restlessness; Delirium; |

**Table 2: Examples of side effects groups**

solution in the developed system. In particular, we rely on forum search engines such as "Google Discussion Search" and crawl all drug related information. In particular, we formulate queries based on drug names, and for each drug name, we crawl top 10K search results that are relevant to the drug.

## 2.3 Knowledge Discovery in Drug Side Effects

We now discuss how to discover interesting knowledge from the constructed knowledge base and crawled data, which are described in the previous sections.

### *2.3.1 Identifying Related Drug Side Effects from the Knowledge Base*

The knowledge base provides a valuable source for known drug side effects. It is interesting to study how to infer the relations among different side effects. Intuitively, the relationships between different side effects could be determined by their co-occurrences in different drugs. If two side effects alway co-occur in drug labels, they might be related. As an example, we may infer "anemia" is related to "lightheadedness" or "dizziness" based on their co-occurrences. In order to identify related drug side effects, we propose to compute the similarity between two side effects based on the mutual information value [2], which is calculated through their associations with different drugs. Specifically, the mutual information can be computed as follows:

$$I(X,Y) = \sum_{X,Y \subset \{0,1\}} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)}, \qquad (1)$$

where $X$ and $Y$ are two binary random variables that correspond to the association (related/non-related) of two side effects with each drug. We then apply an agglomerative hierarchical clustering to group side effects based on their similarity. Table 2 shows a few examples of related drug side effects.

### *2.3.2 Detecting Unrecognized Drug Side Effects from the Crawled Data*

The basic idea of detecting unrecognized side effects of a drug is to monitor the number of mentions for every possible side effect for the drug. Intuitively, our targets are the side effects that are associated with the drug and they are unrecognized (i.e., not mentioned in any online sources). In order to do that, we need to determines whether a page is relevant (i.e., a page mentions a side effect), and then count the number of relevant pages for the side effect in the crawled data collection.

Since side effects are related and the mentions of a single side effect could be very small, we propose to compute the relevance score of a page for a given side effect based on not only the mentions of the side effect but also those of its related side effects. Formally, given a side effect $s$ and document $d$, their relevance score can be computed as follows:

$$R(s,d) = c(s,d) + \alpha \sum_{s' \subset G} (c(s',d) \times I(s,s')) \qquad (2)$$

where $c(s,d)$ is the occurrences of the side effect $s$ in document $d$, $G$ is a group of related side effects that are identified using the method described early, and $I(s,s')$ is the mutual information value as shown in Equation (1).

Given a crawled discussion page $d$, if $R(s,d)$ is larger than a threshold (e.g. 0.8), the discussion $d$ will be regarded as relevant to side effect $s$. By monitoring the number of relevant discussions for its possible side effect, we may be able to detect the unrecognized drug side effects for a particular drug.

In summary, given a drug and a possible side effect, the developed system would count the number of mentions discussed at a given time period and monitor the trend over the time. If the side effect is unrecognized and we can see consistent high volume of discussions over the time, we would generate a warning message and summarize our finding.

## 3. DEMO DESCRIPTION

The purpose of the demo is to show three main features of the developed system, i.e., *UDWarning*.

- **Search for Known Side Effects of a Drug:** The system would allow users to type a drug and return all the known side effects associated with the drug. The

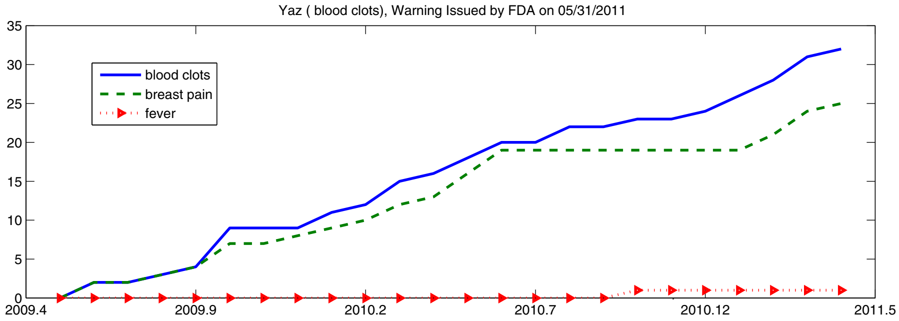

**Figure 3: FDA posted on 05/31/2011 that Yaz can increase the "risk of blood clot". "Breast pain" is a known side effect of Yaz while "fever" is not**

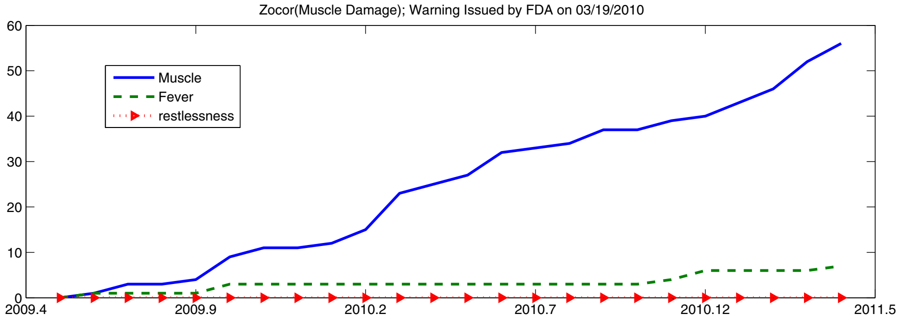

**Figure 4: FDA issued an warning about "increased risk of muscle injury" with Zocor on March 19, 2010. "Fever" is a known side effect of Zocor while "restlessness" is not**

side effects are ranked based on their weights. One possible weighting strategy for a side effect is based on the number of drugs that have the side effect. Intuitively, if a side effect is associated with many drugs, it may not be as serious as others. This idea is similar to the IDF weighting strategy used in IR [1].

- **Search for Drugs Containing a Side Effect:** The system would also allow users to type a symptom (i.e., side effect), and return a list of all the drugs that contain the side effect. The ranking of the drugs could be personalized based on the user profile. The drugs that users have taken in the past would definitely be ranked higher than others.

- **Monitor Online Discussions to Detect Unknown Drug Side Effects:** The system would allow users to monitor the online discussions of all the possible side effects of a drug, and the patterns generated from the system enables us to identify unknown drug side effects. Moreover, we will demonstrate the effectiveness of the developed system through three case studies, where we selected three drugs that have been recently recalled or warned by FDA due to unknown drug side effect and see whether the developed system can make the detection before the official warning from FDA. As shown in Figure 2, 3 and 4, the number of discussions about the unrecoganized drug side effects has started increasing before the official annoucement. Thus, the developed system is able to detect unrecognized drug side effects earlier.

## 4. CONCLUSIONS

*UDWarning* is an early warning system for discovering possible unknown drug side effects. The demo will show its capabilities to integrate online information of drug side effects and to detect unknown drug side effects much earlier than existing strategies.

## 5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number IIS-1017026 and University of Delaware Research Foundation Strategic Initiatives. We thank the anonymous reviewers for their useful comments.

## 6. REFERENCES

[1] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[2] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979.