Predicting Adverse Drug Events from Personal Health Messages

Brant W. Chee, MS, Richard Berlin, MD, Bruce Schatz, PhD
Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL

Abstract

Adverse drug events (ADEs) remain a large problem in the United States, being the fourth leading cause of death, despite post market drug surveillance. Much post consumer drug surveillance relies on self-reported "spontaneous" patient data. Previous work has performed datamining over the FDA's Adverse Event Reporting System (AERS) and other spontaneous reporting systems to identify drug interactions and drugs correlated with high rates of serious adverse events. However, safety problems have resulted from the lack of post marketing surveillance information about drugs, with underreporting rates of up to 98% within such systems[1,2].

We explore the use of online health forums as a source of data to identify drugs for further FDA scrutiny. In this work we aggregate individuals' opinions and review of drugs similar to crowd intelligence[3]*. We use natural language processing to group drugs discussed in similar ways and are able to successfully identify drugs withdrawn from the market based on messages discussing them before their removal.*

Introduction

Post marketing drug surveillance is an important component of drug safety. Clinical trials do not uncover all aspects of drug safety. There are a myriad of co-morbidities, over the counter and prescription drug interactions and food interactions such as grapefruit juice, which may take time to surface once a drug is marketed. The responsibility of post marketing drug safety within the United States lies with the FDA and information relating to adverse drug events is fed to the Adverse Event Reporting System (AERS). Hospitals, pharmaceutical companies and spontaneous reports from patients and doctors populate the AERS with information regarding adverse drug events. Numerous algorithms exist which attempt to mine this database for drugs that cause serious side effects[4]. However, it is widely speculated that AERS grossly underestimates the prevalence of serious adverse events[5].

We believe that online health forum data contains a wealth of information that is not being utilized by the FDA to enrich the AERS database. We explore a system for identifying messages within an online health forum containing information about ADEs. An increasing number of people are using the Internet to search for information about health; 61% of Americans have looked online for health information and of these, 6% have posted information about health or medical matters in an online discussion[6]. This results in approximately 11 million people in the United States alone, providing a sizable patient population discussing topics including drug safety. This number will likely rise as people become increasingly comfortable with social media sites; 33% of young adults aged 18-29 have posted a comment on a website, blog, or newsgroup and 26% of all adults aged 18 and older have done the same[7].

Recent studies have shown that patients' reports have identified previously un-reported ADEs and that their quality is similar to those of health professional reports. There is also evidence that patients report ADEs when they feel their health professionals have not paid attention to their concerns[8]. The FDA discourages the reporting of non serious drug events. However, it is known that these can lead to non-adherence that can have significant medical consequences [9,10].

We hypothesize that drugs that have undergone regulatory action are talked about in similar ways particularly regarding sentiment – one's positive or negative orientation and effect entities (things that a drug causes). Within online health forums people often describe their experiences on a particular drug, both good and bad. We define drugs which have undergone label changes or regulatory action as "watchlist" drugs since they are added to the FDA's watchlist website. We use machine-learning classifiers to compare messages containing watchlist drugs', pre-regulatory action to other messages containing drugs with no regulatory action. We predict that the more often a non-watchlist drug is classified as a watchlist one the more likely it is to need further regulatory scrutiny.

Related Work

Our earlier work has demonstrated that manually generated lexicons can be used to identify drugs and drug effects from online health forums, specifically Yahoo Health discussion forums[11,12]. The resulting data can be used to visualize occurrences of ADEs over time, as well as drugs likely to co-exist in an individual's treatment regimen.

Later work from other researchers has demonstrated that adverse drug reactions can be extracted from comments about drugs made on DailyStrength[9]. That dataset was constrained to short comments about specific drugs; in the current study we attempted to perform similar extraction on less constrained textual data. Specifically, in our dataset it was unknown if a given forum post contained a drug name, and if so, whether the preceding text referenced the aforementioned drug or some other drug. Finally, we sought to determine whether or not the message contains an adverse event or not. The system used an effect lexicon utilizing the Unified Medical Language System (UMLS) Methathesaurus, MedEffect database, SIDER side effect resource, and manually annotated colloquial phrases from DailyStrength.

Other work in this area has focused on extracting side effect information from medical literature[13,14]. While this is a similar task, the underlying data resource is significantly different. For instance, the SIDER side effect resource mines drug-packaging inserts for pairs of drugs and side effects. This type of textual information is highly regularized with few grammatical and spelling mistakes and can be mapped to medical ontology such as UMLS[14]. A different area of work uses natural language processing to analyze electronic health records (EHR) to identify novel adverse drug events[13]. EHRs contain precise medical terminology and can similarly be mapped to medical ontology and resources. We see our work as more similar to processing of narratives within personal health records due to the lack of structured information. However, the information extraction task is similar to analysis of electronic health records.

Our classification task is most similar to recent work which aims to use a small subset of positively labeled documents to find other documents of the same class within a larger corpus of semi-labeled documents [15]. We are not performing document classification, however, our classification goals are conceptually similar. We aim to classify drugs as watchlist or not with relatively few known positive (watchlist) examples while others likely exist in the data.

None of these methods makes predictions about a drug's safety. Previous work identifies drug and adverse event pairs but does not make a value judgment about the drug. We aim to quantify a drug's safety in some sense by comparing it to its' peers.

Methods

We approach the identification of potential watchlist drugs by using machine learning classifiers over online health forum data. We define a watchlist drug as drugs that have an active FDA safety alert posted for example drugs posted at http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsand Providers/ ucm111085.htm. We pose the problem of identifying drug candidates for further scrutiny as an ensemble classification problem. Ensemble classification is a common technique in machine learning utilizing multiple classifiers to make a decision that is better than any of the constituent parts[16]. We focus on bootstrap aggregating also known as bagging. In this approach each classifier is built over a subset of the total data. Empirically bagging performs quite well and often results in high classification accuracy[17].

Normal machine learning approaches use two sets of data, one for training and the other for testing. A larger set of data is often split into two. To get a better idea of performance the process is run multiple times with the data permuted before each run. Instances that are labeled as positive and classified as positives are considered true positives, instances labeled as negative but classified as positives are considered false positives. Our method focuses on false positives. We believe that if items are consistently labeled as a false positive by many different algorithms across many types of data intuitively they are similar to the positive data in some way. We base our predictions upon multiple false positive classifications over many classifiers and different data sets.

Data

We use messages from online health forums. The online forums we use consist of 27,290 public Health & Wellness Yahoo! Groups. Within these groups there is a total of 12,519,807 messages. These groups range from illness based support groups such as ones focusing on Multiple Sclerosis to groups focusing on herbal home remedies. The messages within these groups span seven years and consist of hundreds of thousands of unique email addresses that we consider as a proxy for people. These messages contain topics including information sharing, support seeking and information about experiences with medications. Two examples of messages taken from Yahoo! groups are in Table 1 below. While no formal content analysis was performed, it appears anecdotally that people who post tend to have more negative responses to medications. Our method relies on drugs being talked about in consistent ways. Despite the apparent overall negative affect, watchlist drugs are identified if there are disproportionately more instances of negative or adverse effect like language describing them compared with other drugs.

Classification using machine learning algorithms typically require large amounts of training data. The performance of these classifiers is often commensurate with the amount of training data available. Due to the nature of the data available relatively few positive examples (watchlist drugs) are available with sufficient amounts of data. There are only 435 drugs with more than 500 unique messages mentioning them; of these there are 63 watchlist drugs. Similarly if we look at drugs with more than 250 unique mentions, there are only 575 drugs and 77 watchlist drugs. Approximately 90% of instances are non-watchlist drugs. This is somewhat comforting in terms of health and drug safety; only 10% of drugs are watchlist and are demonstrated to possibly cause adverse effects. However, in terms of machine learning experiments this leads to problems with bias and data scarcity. With highly unbalanced datasets it is difficult for a machine learning classifier to perform better than the naïve or baseline classifier of always picking the most dominant class.

Table 1: Example messages about Sibutramine highlighting side effects.

```
Date: 4/20/2003
Subject: I will be leaving the group
Hello to All
I have been on Meridia for a year and have lost about 45 pounds but now I am going
off it due to health problems which my doctor feels was caused by it. I had high ,
very high blood pressure, the doctor was going to try me on a blood pressure
medication but felt I should stop Meridia first to see if it could be the possible
trigger. i stopped Meridia two weeks ago and my blood pressure is back to normal. I
was also having hair loss and mental confusions, unless I wrote things down I would
forget all the time. I am so frustrated as I have lots more weight to lose and am
scarred that I will gain all my other weight back but I don't want to die from high
blood pressure. I guess I will just have to deal with being over weight and learn
to love myself. My finance was very upset by the article below and told me to stop
taking the drug also. I also read the following on line. Take care everyone
kimberly

Date: 2/9/2006
Subject: Re:[Meridia Forum] Hi, I'm new!
I was on Meridia and lost over 27 pounds, all of which I have kept off.
However, I had alarming heart rhythm problems so I stopped the Meridia after 1
month.
Please keep in mind that Meridia WAS listed in Consumer Reports as a potentially
unsafe drug. I am living testimony that it was dangerous for  me.
I have lost the rest of the weight by sheer willpower. The Meridia was a wonder
drug but do weigh the risks of obesity with the risks of the drug.
Just my opinion -
```

Classification

The input into machine learning algorithms are feature vectors generated over the words people use to discuss the drugs. Feature selection is an important part of all machine-learning tasks. The goal is to use sufficient numbers of features to enable an algorithm to differentiate between instances both in the training set as well as unforeseen instances while limiting the amount of noise introduced.

We focus on two feature sets; the first is comprised of general vocabulary – all words occurring within messages selected using some heuristic such as frequency cutoff. The second feature set consists of meta-features and world knowledge in the form of counts over specialized lexicons. An example of specialized lexicon includes the number of drug mentions, side effect lexicon or positive or negative sentiment words. The second feature set uses the specialized lexicons to select words, for example creating a feature vector of only medical terminology, drugs, diseases, and sentiment containing words or some subset of them.

The specialized lexicons used to generate meta-features and for the second specialized lexicon only approach include drugs, medical terminology, sentiment, adverse drug event lexicon from MedDRA, and lists of diseases. Lexicon are used instead of more advanced named entity recognition techniques due to language processing difficulty of this type of data. Forum posts contain poor grammar, spelling mistakes, emoticons and other extraneous tokens. These problems and the lack of domain specific tools necessitated the use of dictionary-based approaches.

We believe that a person's sentiment value is an important feature to the classifiers. Earlier work demonstrated that sentiment with regard to drugs discussed within an online forum could be reliably extracted [12]. People's perception of a drug is apparent through their sentiment and can have a bearing on a drug's removal from market. We utilize another feature, number of name and effect entities. Much previous work has utilized dictionary-based approaches for identifying drugs and effects. We believe that these features are useful, especially if people attribute negative effects to drugs or if there are significant correlations within messages.

Given a 90/10 split where 10% of drugs are used to evaluate a classifier, 10 runs should ensure each drug is tested at least once and 50 runs statistically speaking, should allow each drug to be classified 5 times against 5 different classifiers. For these experiments each set of features was used to build hundreds classifiers, test and training sets for 10 fold cross validation. Naïve Bayes (NB) and Support Vector Machine (SVM) with a RBF kernel were the two base classification algorithms we used. Some of the transformations we use included both normalized and non-normalized feature vectors and cost weighting. The input to the feature vectors included general word features (5K, 10K, and 15K features selected using Bi-Normal Separation including unigram, bigrams and trigrams), drug entities, effect entities, disease entities, and sentiment lexicon. The number of classifiers included was determined by:

$$c \cdot w \cdot \sum_{i} \binom{n}{i}$$

where c is the number of classification algorithms, w the cost weighting schemes, and n the number of different feature sets. Cross fold validation involves randomizing the labeled data, splitting it into two partitions, a training set and testing set. The machine-learning algorithm is then trained using the training set and performance is calculated against the testing set this is illustrated in figure 1. These steps are performed multiple times. We believe that the output from these classification runs provides insight into future watchlist drug predictions.

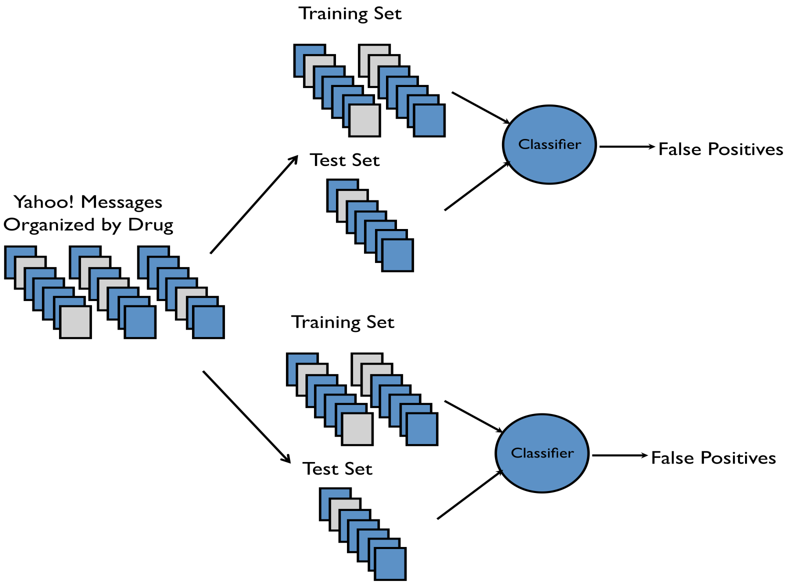

Figure 1: Illustration of how Yahoo! messages are divided and used by multiple classifiers in a bagging approach.

Multiple classifiers are combined in an ensemble like approach taking their combination of features and different classification algorithms to produce a meta classifier where the false positives from each category are combined using a linear combination resulting in a score.

We look at the false positives, drugs that are non-watchlist but are classified as watchlist by a classifier. A false positive occurs when a negative instance is incorrectly identified as a positive one. For SVMs this means that the instance falls on the same side of the hyperplane as the positive instances and is usually close to the boundary[18]. For Naïve Bayes, the maximum likelihood estimate is such that the likelihood of the instance being a watchlist drug is greater than a non-watchlist drug[18].

We are interested in drugs that are consistently marked as false positives. We hypothesize that drugs that are consistently labeled as watchlist are more likely to be "real" or future watchlist or removed from market drugs in the future. The consistency in being labeled as a false positive provides confidence in the prediction. This prediction is based solely on the word features people use to discuss these drugs in the same way as watchlist drugs it will be identified as a watchlist drug. So drugs that are false positives could be real watchlist drugs in the future given third party confirmation such as from the FDA.

We utilize the binary decision, true or false because different classification algorithm's output cannot be directly compared. It does not make sense to compare a likelihood estimate to a distance to a hyperplane. Multiple rounds of classification with mixed training data increase the confidence in a prediction, as does the use of multiple classifiers.

A weighted ratio is created to score the false positives including the ratio of false positives to number of tests, the number of false positives and the number of classifiers that predicted a false positive: Number of False Positives / Number of occurrences (tests) * Number of False Positives * Number of classifier types.

A weighted average over the number of false positives is important, a ratio of .5 given 1 false positive to 2 occurrences is different from 100 false positives to 200 occurrences. The number of different classifiers is similarly important.

Two experiments were run using drugs withdrawn from the market. Firstly withdrawn drugs were labeled as non-watchlist to determine if the classifiers would accurately identify the withdrawn drugs. This procedure validated this bagging approach of watchlist drug identification. Secondly it demonstrated the robustness of the method for watchlist drug identification. The second experiment removed the watchlist drugs and classified them after the classifier was built for each fold of the cross-validation run. This second method should more accurately identify the withdrawn drugs with greater confidence because their data is not mixed with the other non-watchlist drugs possibly reducing the accuracy of the classifiers.

H1: We will be able to identify data that was intentionally labeled with the incorrect class using ensemble methods described above.

Table 2: Top scoring false positive generating drugs with withdrawn drugs mixed in and labeled as non-watchlist. Next to the drug name, the Pos column denotes the number of times a classifier marked the drug as a false positive. Occ indicates the number of occurrences, or number of times the drug was classified. Class indicates the number of different types of classifiers (1-3) that predicted a false positive for a drug. Score indicates the linear combination of Pos, Occ, and Class resulting in a score of the confidence in prediction. Drugs are arranged in descending order.

Drug	**Pos**	**Occ**	**Class**	**Score**
clozapine OR Clozaril OR FazaClo	31	64	3	45.047
fludarabine OR Fludara OR Oforta	29	61	3	41.361
methylphenidate OR Concerta OR Daytrana OR Metadate CD OR Metadate ER OR Methylin OR Methylin ER OR Ritalin OR Ritalin LA OR Ritalin-SR	25	50	3	37.500
morphine OR Astramorph PF OR Avinza OR Duramorph OR Infumorph OR Kadian OR MS Contin OR MSIR OR Morphine IR OR Oramorph SR OR RMS OR Roxanol	14	38	3	15.474
meloxicam OR Mobic	15	50	3	13.500
Extraneal	10	36	3	8.333
aripiprazole OR Abilify OR Abilify Discmelt	9	30	3	8.100
evening primrose OR Evening Primrose Oil OR Primrose Oil	17	56	1	5.161
quetiapine OR Seroquel OR Seroquel XR	15	52	1	4.327
trazodone OR Desyrel OR Desyrel Dividose OR Oleptro	14	46	1	4.261
(acetaminophen AND diphenhydramine) OR Anacin P.M. Aspirin Free OR Coricidin Night Time Cold Relief OR Excedrin PM OR Headache Relief PM OR Legatrin PM OR Mapap PM OR Midol PM OR Percogesic Extra Strength OR Sominex Pain Relief Formula OR Tylenol PM OR Tylenol Severe Allergy OR Tylenol Sore Throat Nighttime OR Unisom with Pain Relief	12	34	1	4.235
thalidomide OR Thalomid	13	44	1	3.841

Results

For the first experiment, four drugs that were withdrawn from the market were identified, Vioxx, Trovan, Baycol, and Palladone. The following table demonstrates the top scoring results from the first run. All drugs with a score > 0 are available online at http://beespace.cs.uiuc.edu:~chee/amia2011/drugs_run1.csv. Table 2 demonstrates runs of one of the experiments discussed below.

Hydromorphone is a narcotic. It is semi-synthetic opiod derived from morphine. An extended-release version of hydromorphone called Palladone was available United States before it was voluntarily withdrawn after a July 2005 FDA advisory warned of a high overdose potential when taken with alcohol[19]. However, as of March 2007, it is still available in the many other European countries.

Cerivastatin (Baycol) is a synthetic statin used to lower cholesterol and prevent cardiovascular disease introduced in 1997. Statins work by inhibiting the enzyme HMG-CoA reductase, which is important in the production of cholesterol. It was voluntarily withdrawn in 2001 due to reports of fatal rhabdomyolysis, which is the breakdown of skeletal muscle that can lead to kidney failure. At the time of withdrawal the FDA had reports of 31 deaths due to rhabdomyolysis.

Trovafloxacin (Trovan) is a broad spectrum antibiotic that was withdrawn from market due to the risk of heptatoxicity - causing liver damage and failure. In 1996, Pfizer violated international law during an epidemic testing the unproven drug on 100 children and infants with brain infections[20]. Currently the FDA is aware of 14 cases of liver failure linked to Trovan and over 100 cases of liver toxicity[21].

Rofecoxib (Vioxx) is a nonsteroidal anti-inflammatory drug (NSAID) first marketed in 1999 as a safer alternative to drugs such as Tylenol or Aleve. It was subsequently withdrawn in 2004 due to a significant increased risk of acute myocardial infarction (heart attack)[22]. Rofecoxib was one of the most widely used drugs to be withdrawn from market. Merck, the maker of Vioxx reported a revenue of $2.5 billion the year before it was withdrawn[23].

Table 3: Top scoring false positive generating drugs with withdrawn drugs removed from the cross validation data. Next to the drug name, the Pos column denotes the number of times a classifier marked the drug as a false positive. Occ indicates the number of occurrences, or number of times the drug was classified. Class indicates the number of different types of classifiers (1-3) that predicted a false positive for a drug. Score indicates the linear combination of Pos, Occ, and Class resulting in a score of the confidence in prediction. Drugs are arranged in descending order. Drugs withdrawn from market are highlighted in yellow.

Drug	Pos	Occ	Class	Score
methylphenidate OR Concerta OR Daytrana OR Metadate CD OR Metadate ER OR Methylin OR Methylin ER OR Ritalin OR Ritalin LA OR Ritalin-SR	30	34	3	79.412
morphine OR Astramorph PF OR Avinza OR Duramorph OR Infumorph OR Kadian OR MS Contin OR MSIR OR Morphine IR OR Oramorph SR OR RMS OR Roxanol	13	38	3	13.342
quetiapine OR Seroquel OR Seroquel XR	14	31	2	12.645
trovafloxacin OR Trovan	33	100	1	10.89
hydromorphone OR Dilaudid OR Dilaudid-HP OR Exalgo OR Palladone	33	100	1	10.89
rofecoxib OR Vioxx	32	100	1	10.24
indomethacin OR Indocin OR Indocin IV OR Indocin SR	19	37	1	9.757
sibutramine OR Meridia	17	34	1	8.500
meloxicam OR Mobic	17	35	1	8.257
vigabatrin OR Sabril	14	31	1	6.323
losartan OR Cozaar	13	28	1	6.036
oxycodone OR ETH-Oxydose OR OxyContin OR OxyIR OR Oxyfast OR Percolone OR Roxicodone OR Roxicodone Intensol	13	30	1	5.633
doxepin OR Adapin OR Prudoxin OR Silenor OR Sinequan OR Zonalon	14	37	1	5.297
aripiprazole OR Abilify OR Abilify Discmelt	13	32	1	5.281
guaifenesin OR Altarussin OR Amibid LA OR Drituss G OR Duratuss G OR GG 200 NR OR Ganidin NR OR Guaifenesin LA OR Guaifenex G OR Guaifenex LA OR Hytuss OR Liquibid OR Mucinex OR Mucinex for Kids OR Muco-Fen 1200 OR Organidin NR OR Q-Bid LA OR Robitussin Chest Congestion OR Scot-Tussin Expectorant OR Tussin	13	34	1	4.971

Thalidomide was also flagged as a high-scoring false positive. Thalidomide was used as an anesthetic but was withdrawn from market in the 1960's after it was found to cause birth defects resulting in babies with no limbs or limbs with finger or toes fused together. However, thalidomide has been remarketed with narrow focus and strong labeling.

Temazepam is an intermediate acting benzodiazepine prescribed as a short term sleeping aid and is sometimes used as an anti-anxiety, anticonvulsant, and muscle relaxant. Sweden and Norway withdrew the drug from market in 1999 due to diversion, abuse, and high rate of overdose deaths in comparison to other drugs of its group. It is still available in the US with strong warnings for severe anaphylactic and anaphylactoid reactions and cautions about complex behavior such as "sleep driving" - driving while not fully awake and having amnesia about the event[24].

Primrose oil is derived from Oenothera biennis and is sometimes used to treat eczema, rheumatoid arthritis, menopausal symptoms, premenstrual syndrome, cancer and diabetes. This supplement had an unusually high score given it is a supplement. However, the broad range of uses and associations with other diseases and medications could lead to misclassification especially since one of the classifiers is based solely upon other drug mentions. This misclassification could also apply to acetaminophen, vitamin e, fish oil and Metamucil.

The most striking find was *Sibutramine (Meridia)*, which is an appetite suppressant and is used to treat obesity. The manufacturer has voluntarily removed it from market. During the time of experimentation the drug was under review and was not considered a watchlist drug. A FDA early communication about the drug was posted on 11/20/2009 and a subsequent follow-up on 1/21/2010 indicating an increased risk of heart attack and stroke in patients with a history of cardiovascular disease[25]. This drug was marked as a false positive repeatedly. The last Yahoo! messages mentioning Sibutramine were from 12/11/2008 almost year before FDA advisories and a little over a year before the UK withdrawal[26]. Table 1 contains example messages about adverse events attributed to Sibutramine from the Yahoo! corpus. These are two examples of the numerous messages that exist and contain serious effects including heart arrhythmia, high blood pressure, confusion, etc.

The results of the second run are available at http://beespace.cs.uiuc.edu/~chee/amia2011/ drugs_run1.csv. A re-ordering of the drugs is seen in Table 3, depicting top twelve scoring drugs to compare against the first run. If we take the identification of European Union drugs with higher scores then the second run performs better. Sibutramine is ranked higher, 5th in the list of predictions.

In both cases we see psychiatric drugs such as Ritalin and Clozapine ranked near the top as well as opiods such as morphine and Oxycodone. This might indicate intuitively that these classes of drugs are more dangerous or likely to cause serious effects than other types of drugs.

Table 4, below shows the scores of the drugs withdrawn from market for both runs as well as their relative ranks within the lists of drugs. The scores of all the withdrawn drugs are higher in the second experiment and the relative ranks were lower. A higher score indicates that more classifiers identified these drugs as a "positive". A lower relative rank determines how close the drug is to the top of the list, similar to a page rank in a search results where the lower the score is the better.

Table 4: Table of drugs withdrawn from the market with their associated scores for the two experiments. The first two columns are the scores associated with each experiment. The following columns are the position of each drug within the list of results for each experiment.

Drug	Score Exp 1	Score Exp 2	Rank Exp 1	Rank Exp 2
Palladone	1.929	10.89	33	4
Trovan	1.761	10.89	40	5
Vioxx	1.62	10.24	50	6
Baycol	0.03	0.04	117	107

These scores indicate the classifiers were better at identifying the drugs withdrawn from market in the second run. The raw scores of Palladone, Trovan and Vioxx were almost a magnitude of order higher and similarly ranked almost a magnitude higher in the list. Disappointingly Baycol's score did not improve much and its relative rank while higher was not significantly higher.

Conclusion

Previous work has explored the identification of adverse drug reactions from various types of medical literature and patient data such as hospital exit logs and electronic health records. These systems identify instances of adverse events within data. Further analysis is necessary to identify candidates for closer investigation. The detection of adverse drug reactions makes no predictive assessment of a drug's safety and whether or not it is considered dangerous in context to other drugs. We believe this is the first work at identifying candidates for further investigation with regards to drug safety based upon patients' text in online health forums.

This work aims to provide support for investigative analysis of a drug's safety using multiple machine learning classifiers and identifying those drugs which are most similar to other watchlist and withdrawn drugs. Our hypothesis was proven correct. We are able to identify data that were intentionally labeled with the incorrect class using ensemble methods (H1). Ensemble methods were able to more accurately identify drug instances withdrawn from market when they were withheld from training sets but included in testing sets.

We have predicted candidates for further investigation based upon multiple false positives from many different classifiers. These drugs are false positives in the sense that they are not currently watchlist or recalled drugs. However, these drugs at some point in the future could become watchlist or withdrawn drugs. A list of potential watchlist drugs was produced, the most significant of these is Sibutramine, a weight loss drug. Our dataset only has posts up to a year before it was put on a watchlist then subsequently withdrawn from the European market. This drug has been recently voluntarily withdrawn from the market by its manufacturer, almost three years beyond the data we currently have.

This method is useful for drugs with a wide audience that contains many postings and less serious adverse events. This method like other statistical methods over AERS will find it difficult to detect a signal with few adverse effects.

Limitations

Our method identifies 127 candidate drugs with scores ranging from 79.4 to 0.02. The 127 candidates are a relatively small number of drugs compared to the 11,706 prescription drugs, 390 over the counter drugs, and numerous herbal remedies that exist. However, it is a larger percentage of the 575 total drugs we had data for, illustrating that we have data for relatively few drugs compared to the total numbers of existing drugs. This type of system requires large amounts of data, a weakness of many machine-learning techniques. However, we have demonstrated that this methodology could be useful in identifying drugs for further study.

This technique assesses which drugs are talked about in similar ways; like comparing a drug to its peers. This may not be a fair assessment of a drug and may inaccurately predict a drug for safety warning or withdrawal based upon its perception. Perception of a drug is different from its actual effects. Many people may like their drug despite low efficacy or serious side effects [27].

A current drawback of this method is that we aggregate all messages across all disease groups. Drugs have different audiences and certain segments of the population are at greater risk for specific diseases. For example women are more likely to suffer from multiple sclerosis than men, therefore the question remains is it valid to group drugs that are targeted to different segments of the population together? It is unknown whether or not it is correct to group all drugs and messages together.

The number of messages for each drug is not evenly distributed; for example, the numbers of drugs mentioning nonsteroidal anti-inflammatory drugs (NSAIDs) are much greater than those mentioning Tysabri (a multiple sclerosis drug). Many NSAIDs such as Aleve are available over the counter and have multiple applications. This differs significantly from a narrow purpose drug used by a smaller population. As stated previously the prevalence of vitamins or over the counter painkillers that seem innocuous abound on both lists. This might be attributed to their wide spread use among many conditions and use in combination with many different drugs. Both experiments demonstrate a relatively high score for acetaminophen and acetaminophen containing products, however the causality of scores is not established. Like many other machine learning and classification tasks, even if the features used by a classifier are known, there is no established causation, only the correlation between feature and class. It remains to be seen if the outcome is corroborating the recent allegations over the safety of acetaminophen with regard to overdosing and safety of children's products or due to the their widespread use and association with many different drugs and diseases.

We rely on words and groups of words as features. Therefore, phrase ordering and spelling errors introduce problems. Misspelled words are not correctly attributed to their correct word or phrase resulting in lower classification accuracy. Greater numbers of messages help to mitigate this problem but the problem is more pronounced for drugs with fewer mentions or that exist in disease communities with cognitive impairment. Drugs that are talked about in ways that are different from most of the training set will also be misclassified. The same is true for drugs that cause rare but serious effects, or effects that are different from the other watchlist drugs. These are possible reasons why the score for Baycol was low.

This method is a black box. The inputs are groups of messages with drug mentions and the output a single score. This output is lacking in explanation. If one used a single Naïve Bayesian classifier the sets of words or word groups that contributed most to the groups classification could be examined. However when using multiple classifiers and multiple methods this approach is unfeasible. A justification or explanation for a prediction might be more satisfying.

It is unknown if this approach will work for other data sources such as tweets or blogs. Tweets and blogs are different in nature from forum posts, for example tweets are limited to the equivalent of a sentence or two in length. Tweets are much shorter than forum messages. The shortness of a tweet can increase the likelihood of a single topic and clear causality between a drug and effect. However there are also issues with noise such as hashtags or shortening words due to the message length limitations for example using "4" instead of "for". Our method relies on processing the entire message to create a feature vector. Within a message could be multiple drug mentions leading to combined feature vectors and miscalculations.

A limitation of this work is that crowd intelligence can fail due to emotional factors. People want to belong and succumb to peer pressure, discussing one's health is a highly personal and emotional subject. The media or other events could also influence people's perceptions of drugs and cause them to talk about things in similar ways.

Future Work

A first step might include controlling for number of messages, discarding drugs with disproportionately many messages. A small pilot study of a subset of the drugs for a subset of diseases might be performed to see if prediction accuracy increases. However, to perform widespread analysis, more data is necessary to separate the messages by disease or expected demographic of a disease population. Currently the Yahoo! groups do not have enough data available to provide this analysis. Other health sites such as PatientsLikeMe with specific disease communities may provide richer sources of data.

We would like to apply the use of ontology in order to group similar effects together. Grouping effects together would allow us to leverage the sparse data more effectively and apply world knowledge to the groupings in a meaningful way. We can combine small numbers of instances of effects into larger numbers of more general or similar terms leading to more accurate predictions. Language tends to vary within different communities, in one community, "shakes" might be used instead of "tremors" in another. Ontology or automated generation of synonym lists might be useful in reconciling differences between communities. A promising example includes the Consumer Health Vocabularies.

Further exploration of this work with different data sources especially personal narratives that might be found in PHR or spontaneous reporting systems (SRS) like those found in AERS. Current signal detection pharmacovigilance techniques often utilize the structured output of SRS where data from SRS narratives are manually transcribed into structured data. We would like to see collaboration between the FDA, manufacturers and healthcare providers, with integration of personal health message information with EMR and PHR through providers. The FDA and providers should determine the side effects that are to be sought out within the personal health message data. Further, the place for unsolicited and previously unseen side effect discovery needs to be established.

In conclusion, we demonstrate a scalable technique that needs little manually annotated training data, which is a limitation of the application of machine learning for many tasks. We believe that this method can be generalized to different types of data sources such as twitter or blogs with little augmentation due to the lack of custom features or advanced natural language processing techniques such as full syntactic parsing. We demonstrate that our method was able to identify drugs removed from market both when the data was intentionally mislabeled and trained upon as well as when it was used only for testing. We propose this method as a course signal detection technique that can augment existing SRS data and methods using unstructured information directly found in online sources.

References

1. Fletcher AP. Spontaneous adverse drug reaction reporting vs event monitoring: a comparison. *Journal of the Royal Society of Medicine*. 1991;84(6):341.
2. McClellan M. Drug Safety Reform at the FDA—Pendulum Swing or Systematic Improvement? *New England Journal of Medicine*. 2007;356(17):1700–1702.

3. Surowiecki J. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Random House of Canada; 2004.
4. Harpaz R, Haerian K, Chase HS, Friedman C. Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. In: *AMIA Annual Symposium Proceedings*.Vol 2010.; 2010:281.
5. Anon. Making a difference. *Nat Biotechnol*. 2009;27(4):297-297.
6. Fox S, Jones S. The social life of health information. *Washington, DC: Pew Internet & American Life Project*. 2009:2009–12.
7. Lenhart A, Purcell K, Smith A, Zickuhr K. Social Media and Young Adults. *Washington, DC: Pew Internet & American Life Project*.
8. Blenkinsopp A, Wilkie P, Wang M, Routledge PA. Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. *Br J Clin Pharmacol*. 2007;63(2):148-156.
9. Leaman R, Wojtulewicz L, Sullivan R, et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*.; 2010:117–125.
10. Hochberg AM, Reisinger SJ, Pearson RK, O'Hara DJ, Hall K. Using data mining to predict safety actions from FDA adverse event reporting system data. *Drug Inf J*. 2007;41(5):633–44.
11. Chee B, Karahalios KG, Schatz B. Social visualization of health messages. In: *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*.; 2009:1–10.
12. Chee B, Berlin R, Schatz B. Measuring population health using personal health messages. *AMIA Annu Symp Proc*. 2009;2009:92-96.
13. Friedman C. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. *Artificial Intelligence in Medicine*. 2009:1–5.
14. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*. 2010;6(1).
15. Yeganova L, Comeau DC, Kim W, Wilbur WJ. Text Mining Techniques for Leveraging Positively Labeled Data. *ACL HLT 2011*.155.
16. Rokach L. Ensemble-based classifiers. *Artificial Intelligence Review*. 2010;33(1):1–39.
17. Breiman L. Bagging predictors. *Machine learning*. 1996;24(2):123–140.
18. Alpaydin E. *Introduction to machine learning*. The MIT Press; 2004.
19. Anon. Safety Alerts for Human Medical Products - Palladone (hydromorphone hydrochloride). Available at: http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm152047.htm. Accessed March 16, 2011.
20. Anon. FDA statement on Baycol withdrawal. *USA Today*. 2001. Available at: http://www.usatoday.com/money/general/2001-08-08-bayer-fda-statement.htm. Accessed March 17, 2011.
21. Center for Drug Evaluation and Research. Public Health Advisories (Drugs) - Food and Drug Administration 09 June 1999 Trovan (Trovafloxacin / Alatrofloxacin Mesylate). Available at: http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/DrugSafetyInformationforHeathcareProfessionals/PublicHealthAdvisories/UCM053103. Accessed March 16, 2011.
22. Anon. 2004 - FDA Issues Public Health Advisory on Vioxx as its Manufacturer Voluntarily Withdraws the Product. Available at: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2004/ucm108361.htm. Accessed March 16, 2011.
23. Reuters. Merck Sees Slightly Higher 2007 Earnings. *The New York Times*. 2006. Available at: http://www.nytimes.com/2006/12/07/business/07drug.html?ex=1323147600%20&en=19d27b5814f1c1e8&ei=5088&partner=rssnyt&emc=rss. Accessed March 16, 2011.
24. Office of the Commissioner. Drug Safety Labeling Changes - Restoril (temazepam) Capsules. Available at: http://www.fda.gov/Safety/MedWatch/SafetyInformation/Safety-RelatedDrugLabelingChanges/ucm113808.htm. Accessed March 16, 2011.
25. Office of the Commissioner. Safety Alerts for Human Medical Products - Meridia (sibutramine hydrochloride): Follow-Up to an Early Communication about an Ongoing Safety Review. Available at: http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm198221.htm. Accessed March 16, 2011.
26. Anon. Top obesity drug being suspended. *BBC*. 2010. Available at: http://news.bbc.co.uk/2/hi/health/8473555.stm. Accessed March 17, 2011.
27. Silver M. *Success with Heart Failure*. Cambridge, MA: Perseus Publishing; 2002.