Medical Data Mining: Improving Information Accessibility using Online Patient Drug Reviews

by

Yueyang Alice Li

S.B., Massachusetts Institute of Technology (2010)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2011

Author .....
Department of Electrical Engineering and Computer Science
January 4, 2011

Certified by .....
Dr. Stephanie Seneff
Senior Research Scientist
Thesis Supervisor

Accepted by .....
Dr. Christopher J. Terman
Chairman, Masters of Engineering Thesis Committee

Medical Data Mining: Improving Information Accessibility using Online Patient Drug Reviews

by

Yueyang Alice Li

Submitted to the Department of Electrical Engineering and Computer Science
on January 4, 2011, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

We address the problem of information accessibility for patients concerned about pharmaceutical drug side effects and experiences. We create a new corpus of online patient-provided drug reviews and present our initial experiments on that corpus. We detect biases in term distributions that show a statistically significant association between a class of cholesterol-lowering drugs called statins, and a wide range of alarming disorders, including depression, memory loss, and heart failure. We also develop an initial language model for speech recognition in the medical domain, with transcribed data on sample patient comments collected with Amazon Mechanical Turk. Our findings show that patient-reported drug experiences have great potential to empower consumers to make more informed decisions about medical drugs, and our methods will be used to increase information accessibility for consumers.

Thesis Supervisor: Dr. Stephanie Seneff
Title: Senior Research Scientist

Acknowledgments

I would like to express my sincere gratitude to Stephanie Seneff for acting as my advisor. Her invaluable expertise and generous guidance were instrumental to the completion of this thesis, and her eternal enthusiasm kept me motivated throughout the year.

It has been a pleasure being part of the Spoken Language Systems group. Special thanks goes to JingJing Liu for her knowledgeable insight and collaboration in the classification experiments, to Jim Glass for his kind encouragement, and to Victor Zue for his advice on grad school and life beyond. I would especially like to thank Scott Cyphers who was always willing to answer my endless questions about the Galaxy system. Many thanks to everyone in the group for making it such an enjoyable and welcome place to work.

I would also like to acknowledge Tommi Jaakkola for his patient and illuminating instruction on machine learning, and Regina Barzilay for first introducing me to NLP. This work would not have been possible without Victor Costan, who gave me massive help whenever I ran into difficulties with Ruby on Rails. I also deeply appreciate my friends and colleagues at CSAIL, for most enjoyable discussions and treasured memories.

Finally, I am indebted to my wonderful family for their unconditional love and support.

Bibligraphic Note

Portions of this thesis are based on the paper entitled "Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews - Focus on Statin Drugs" with Stephanie Seneff and JingJing Liu, which was submitted to the *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Contents

1 Introduction **17**

1.1 Vision 19

1.2 Contributions 20

1.3 Thesis Overview 21

2 Related Work **23**

2.1 Term Identification 23

2.1.1 Medical Knowledge Resources 24

2.1.2 Statistical Approaches 25

2.2 Medical Applications 26

2.2.1 Dialogue Systems 26

2.2.2 Health Surveillance 28

2.3 Summary 30

3 Data **31**

3.1 Data Collection 31

3.1.1 Data Sources 32

3.1.2 Data Coverage 34

3.2 Example Comments 35

3.3 Spelling Correction 36

4 Automatic Discovery of Side Effects: Focus on Cholesterol-Lowering Drugs **39**

4.1 Side Effects of Cholesterol-lowering Drugs: Brief Literature Review . 40
4.1.1 Statin Drugs . . . 41
4.1.2 Non-Statin Cholesterol-Lowering Drugs . . . 42
4.2 Data . . . 43
4.3 Methods . . . 43
4.3.1 Log Likelihood Statistic . . . 44
4.3.2 Pointwise Mutual Information . . . 45
4.3.3 Set Operations . . . 46
4.4 Results . . . 46
4.4.1 Cholesterol-lowering vs Blood-pressure-lowering Drugs . . . 46
4.4.2 Statins vs Non-statins . . . 47
4.4.3 Gender Differences . . . 50
4.4.4 Lipophilic vs Hydrophilic Statins . . . 51
4.5 Discussion . . . 51
4.5.1 Limitations . . . 52
4.6 Summary . . . 53

5 Speech Recognition Experiments 55
5.1 Collection of Spoken Questions Data . . . 55
5.2 Methods . . . 57
5.2.1 Trigram Language Model . . . 57
5.2.2 Data Sparsity . . . 58
5.3 Results and Discussion . . . 59
5.4 Summary . . . 61

6 Additional Preliminary Experiments 63
6.1 Multi-word Term Identification . . . 63
6.1.1 Term Frequency . . . 64
6.1.2 Part of Speech Filter . . . 65
6.1.3 Association Measures . . . 66
6.1.4 Discussion . . . 68

6.2 Side Effect Term Extraction 68
6.3 Review Classification 69
6.3.1 Methods 70
6.3.2 Results 70
6.3.3 Discussion 71
6.4 Topic Modeling 71
6.4.1 Methods 72
6.4.2 Results and Discussion 72

7 Conclusions and Future Work **75**

A Hierarchy for Cholesterol Lowering Drugs **77**

B Anecdotes for AMT Question Collection **79**

C Sample Questions Collected Using AMT **81**
C.1 Cholesterol Lowering Drugs 81
C.2 General Medication 81

D Qualifying Terms Excluded from Side Effects **83**

List of Figures

3-1 Database schema for storing patient comments. 33

3-2 Distribution of comments in cholesterol lowering drug class. Numeric values are total number of reviews in each class. 35

5-1 Prompt presented to Amazon Mechanical Turk workers to collect sample questions about cholesterol-lowering drug experiences. 56

List of Tables

3.1 Sources of data and number of reviews of cholesterol lowering drugs. . 32

4.1 Selected words and phrases that distributed differently over cholesterol-lowering drug reviews and renin-angiotensin drug reviews. The log-likelihood ratio (LLR) and p-value are provided. k_1: cholesterol-lowering drugs. k_2: renin-angiotensin drugs. *Values are essentially 0 ($< 1E-300$). 47

4.2 Twenty terms with highest class preference for statin drug reviews. . 48

4.3 Terms with high class preference for non-statin cholesterol-lowering drug reviews. 49

4.4 Selected words and phrases that distributed differently over statin and non-statin cholesterol lowering drug classes. The log-likelihood ratio (LLR) and p-value are provided. k_1 and k_2: number of statin and non-statin reviews containing the term, respectively. The upper set are far more common in statin drug reviews, whereas the lower set are more frequent in non-statin reviews. 50

4.5 Selected words and phrases in the statin reviews that distributed differently over gender. k_1: male reviews. k_2: female reviews. 51

4.6 Selected words that were more common in lipophilic than in hydrophilic statin reviews. k_1: lipophilic statin reviews. k_2: hydrophilic statin reviews. 52

5.1 Classes used for class n-gram training. 59

5.2 The use of class n-grams slightly improves recognizer performance. . . 60

5.3 Word error rate for various training sets. Additional corpora were used to train the language model, including the comments about statins collected from online forums (and were then used to prompt turkers to ask questions), general medicine-related questions, and the MiCASE corpus. 60

6.1 Bigrams ranked by frequency. 64
6.2 Bigrams ranked by frequency with stop words removed. 64
6.3 Example part of speech patterns for terminology extraction. 65
6.4 Bigrams passed through a part of speech pattern filter. 65
6.5 Bigrams passed through a part of speech pattern filter and containing only letters a-z. 66
6.6 Bigrams ranked by pointwise mutual information. 67
6.7 Bigrams ranked by symmetric conditional probability. 67
6.8 Side effects extracted from the Askapatient corpus. Bolded terms are not found in the COSTART corpus of adverse reaction terms. 69
6.9 Drug review classification performance. BS: baseline; LLR: log likelihood ratio; DN: drug names. Precision, recall, and F-score are for statin reviews. 71
6.10 Examples of latent classes automatically discovered using LDA 73

Chapter 1

Introduction

The last few decades have witnessed a steady increase in drug prescriptions for the treatment of biometric markers rather than overt physiological symptoms. Today, people regularly take multiple drugs in order to normalize serum levels of biomarkers such as cholesterol or glucose. Indeed, almost half of all Americans take prescription drugs each month, which cost over $200 billion in the US in 2008 alone [30]. However, these drugs can often have debilitating and even life-threatening side effects. When a person taking multiple drugs experiences a new symptom, it is not always clear which, if any, of the drugs or drug combinations are responsible.

Before medical drugs and treatments can be approved in the US, clinical trials are conducted to assess their safety and effectiveness. However, these costly trials have been criticized because they are often designed and conducted by the pharmaceutical company that has a large financial stake in the success of the drug. These trials are often too short, and involve too few people to give conclusive results. A large study recently conducted on the heart failure drug, nesiritude, invalidated the findings of the smaller study that had led to the drug's approval [44]. Marcia Angell, who served as editor-in-chief of the New England Journal of Medicine, also criticized the clinical trials process, noting the conflicts of interest, the ease with which trials can be biased to nearly ensure positive results, and prevalence of the suppression of negative trial results [3].

Beyond clinical trials, regulatory agencies also monitor drug adverse reactions

through spontaneous reporting after the drug has come to market. In the United States, the Food and Drug Administration (FDA) maintains a post-marketing surveillance program called MedWatch, which allows healthcare professionals to report adverse reactions of drugs. However, the difficulty of using these reporting systems and their voluntary nature may contribute to an under-estimation of adverse drug reactions [5,83]. It is difficult to accurately quantify the number of adverse reactions that go unreported, but previous studies have found that voluntary reporting detects less than 1% of adverse drug reactions [38]. In addition, patients and even clinicians may not recognize that certain symptoms are caused by the drug.

Increasingly, consumers are turning to online health websites to seek medical advice. Recently, a number of online communities have developed around sharing medical experiences and expertise. These informal forums are rich and invaluable sources of information on the effectiveness and side effects of drugs because they make it possible to reach a wider audience, and supplement information available from drug manufacturers and health professionals. For psychological reasons, patients are often more comfortable sharing personal experiences in support groups, with other participants who are going through similar issues [15].

These health websites have the added benefit of closing the language gap between clinical language and patient vocabulary, which can cause confusion and misunderstanding. Studies have also shown that misspellings, misuse of words, and ambiguous abbreviations can lead to poor information retrieval results [43, 52, 92].

Online health websites are addressing the issue of terminology mismatch, making it possible to reach a wider audience. However they are subject to a different problem of information overload. The trade-off of their accessibility is difficulty finding relevant information for specific queries. The sheer volume of data and presence of noise masks its true value.

Data mining and content summarization are well studied topics in research, especially in the restaurant and movie domains, where the opinion features of online reviews are often overwhelmed by irrelevant commentary. By using a combination of rule-based parsing and statistical analysis of the distribution and concurrence of

certain words and phrases, consumer comments can be consolidated to provide useful summarizations of individual restaurants with promising results [49].

Analogously, we can perform similar retrieval and summarization techniques in the medical domain on patient anecdotes posted online, to address the dual problems of insufficient clinical studies and mismatched terminology. Natural language analysis of drug effectiveness and side effects could prove invaluable to patients who want to learn more about the experience of taking certain drugs. However, the difficulty of performing natural language analysis is increased in the medical domain because of the highly domain-specific vocabulary, which also makes it interesting for natural language research.

1.1 Vision

We propose an interactive online system that will answer questions about medical drugs by consolidating patient-reported drug experiences and will automatically identify important and relevant information pertaining to drug effectiveness and side effects. The use of natural language understanding will allow more specific queries and accessibility to individuals without medical training. Furthermore, in the absence of relevant patient trials or consolidated and structured physician reports, the information gathered by automatically processing patient reported symptoms may provide invaluable insight on drug adverse reactions and effectiveness.

We envision an integrated system that encompasses a living database of patient-reported anecdotes and supports both text and speech interaction modalities. The system will be a valuable resource for patients who want to learn about and share experiences on the effectiveness and side effects of medical drugs. Users will not only be able to ask questions about drugs or symptoms, but also submit their own comments by typing or speaking about their experiences taking certain drugs. The database will also incorporate information mined from online patient discussions of drugs and publicly available medical data sets, such as the FDA's Adverse Events Report System, which contains reports from MedWatch. As more people use the

system, the database will be augmented with these new entries and thus deliver more relevant results to new queries.

In response to user queries, relevant comments from the database will be returned that may provide the answers the user seeks. To avoid overloading users with too many comments, we will use automatic summarization techniques to highlight the key points relevant to the user query. Statistical analysis may also be performed to answer questions about population statistics, such as the correlation between observed symptoms and certain drugs.

1.2 Contributions

This thesis describes our preliminary experiments in building an interactive medical drug resource for patients. As a preliminary study in this area, we tackle a number of common tasks including spelling correction, tokenization, and term identification. We also explore the degree to which statistical methods such as co-occurrence measures, linear classifiers, and topic models can be used to extract summary information derived from biases in word distributions, and to subsequently detect associations between particular drugs or drug classes and specific symptoms.

The key contributions of this research are:

1. We create a large corpus of over 100,000 patient-provided medical drug reviews and comments.

2. We apply statistical techniques to identify side effects and other terms associated with a specific drug class.

3. We apply topic modeling methods to discover drug side effects and side effect classes.

4. We develop an initial speech recognition system to support spoken queries in the medical domain.

1.3 Thesis Overview

The thesis is organized as follows. First, we provide an overview of related work in natural language processing in the medical domain. We then describe the data collected on medical drug reviews and comments. In chapter 4, we discuss the findings from automatic side effect discovery experiments with a focus on cholesterol-lowering drugs, especially statins. We present results from speech recognition experiments conducted on spoken question data collected from Amazon Mechanical Turk in chapter 5. We discuss additional experiments in review classification and topic modeling, followed by our conclusions in chapter 7.

Chapter 2

Related Work

This thesis builds on a number of areas of previous work, from general tasks such as word sense disambiguation, syntactic parsing, and topic detection, to the domain specific applications of clinical decision making, medical dialogue systems, and diagnosis. With the adoption of electronic health records and increased availability of clinical data in textual form [55], it is becoming increasingly feasible to apply NLP techniques to the medical domain. Natural language processing methods have already been used to supplement health provider education, provide more personalized medical care, and assist in a patient's behavioral compliance, which can greatly reduce the billions of dollars spent each year on health care by encouraging healthier life styles [23]. In this chapter, we will give an overview of term identification methods, which are crucial to many NLP tasks. We also present a survey of applications in the medical domain.

2.1 Term Identification

The development of natural language systems in specialized domains often begins with term identification, an important subtask of information extraction with applications in automatic indexing, language generation, and machine translation. The term identification task can be subdivided into three main steps, (1) term recognition, (2) term classification, and (3) term mapping. As an example, consider the sentence

"*Lipitor caused muscle pain.*" In the recognition step, we would detect two terms of interest (*Lipitor* and *muscle pain*). We would then classify the terms as a drug name and adverse reaction, respectively. Finally, we would map these terms to concepts in a medical lexicon, such as the UMLS Metathesaurus, which is described in detail in section 2.1.1.

Proper treatment of the term identification task may involve parsing techniques that consider contextual information, statistical methods that use measures such as frequency or term frequency inverse document frequency (tf-idf), and lexicon based methods that compare terms against words in a given knowledge base. Term classification is often performed with classifiers using semantic, contextual, and syntactic features, for example, Chowdhury et al.'s work on identifying medical terms, including diseases [10], Settle's study of gene and protein names [69] and Aramaki's experiments on extracting adverse effects from clinical records [4].

2.1.1 Medical Knowledge Resources

The US National Library of Medicine (NLM) has created a set of biomedical lexica and tools known collectively as the Unified Medical Language System (UMLS). First developed in 1986, it is updated quarterly and is used extensively in biomedical NLP research. Resources within the UMLS include the Metathesaurus [1], composed of over 1 million biomedical concepts, the Semantic Network (which provides semantic links among categories such as organisms, anatomical structures, and chemical compounds), and the SPECIALIST Lexicon of both common English and biomedical terms, with syntactic information.

Within the Metathesaurus, we find many specialized vocabularies including RxNorm, a "standardized nomenclature for clinical drugs and drug delivery devices" [50], the World Health Organization (WHO) Adverse Drug Reaction Terminology, and MedlinePlus Health Topics, among 50 others[2]. Concepts found in the Metathesaurus can

[1] `http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`

[2] `http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html`

be mapped to semantic types in the Semantic Network.

Applications

An important application of term extraction is the identification of adverse reaction terms. Penz et al. [60] studied text records of surgical operations in the Veterans Administration database to identify the effect of central venous catheter placements on adverse reactions. Using phrase matching and parsing techniques, they were able to identify adverse reactions with 0.80 specificity and 0.72 sensitivity. Melton and Hripcsak [54] achieved much higher specificity (0.99) at the cost of sensitivity (0.28).

Despite the availability of manually annotated resources such as UMLS, it remains difficult to map terms found in text to concepts in these medical lexica. Historically, dictionary look-up methods in the medical domain have exhibited poor matching [26,35]. The NLM has developed a tool called MetaMap Transfer (MMTx), which automatically maps biomedical documents to terms in the UMLS Metathesaurus using text parsing, linguistic filtering, variant generation, and finally matching to concepts in the Metathesaurus [36]. However, Divita [18] found that MetaMap Transfer had only a 53% success rate at matching terms in free text to concepts in UMLS. Settles's work also suggested that the use of semantic lexica may be of questionable benefit compared to text-based features for entity recognition purposes [69]. The term recognition problem is especially pronounced in the medical domain because of the fast-evolving vocabulary and ambiguity or polysemy of terms.

2.1.2 Statistical Approaches

Statistical and machine learning techniques may prove more successful at term recognition than approaches that rely on accurately mapping free text to controlled vocabularies, especially with the availability of large datasets. For example, Kazama et al. [41] used multi-class support vector machines (SVMs) to learn boundary features of terms in the GENIA corpus[3]. Another study employed Hidden Markov Models

[3] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/`

(HMMs) with orthographic features to discover gene names [13].

With the high density of medical terms in text, we can also use probabilistic collocation extraction methods to identify terms of interest. A number of measures of association have been used in previous research, including simple frequency, pointwise mutual information [11], selectional association [63], log-likelihood [20], symmetric conditional probability [71], and set association measures such as the Dice [17] and Jaccard [37] indices. Many of these measures are defined in more detail in section 4.3, where they are used to detect biases in word distributions.

2.2 Medical Applications

Using tools such as the UMLS, researchers have studied medical text for a wide range of purposes. Weeber et al. found new applications for medical drugs through textual analysis of PubMed articles. They argued that researchers should consider textual databases as an additional source of knowledge. Reeve et al. used various association measures to determine concept saliency in biomedical texts for extractive text summarization. Plaza et al. [61] applied a graph-based approach to map terms in biomedical documents to concepts found in UMLS, also for summarization purposes. These studies, based on documents containing many technical biomedical terms, benefit from the use of the UMLS Metathesaurus for mapping terms to medical concepts. Additional applications include medical dialogue systems and biosurveillance, which are described below.

2.2.1 Dialogue Systems

Personalized medical systems often implement a dialogue system that aims to simulate or supplement the expertise of health care providers [46]. Conversational systems provide a more natural interface for users, and have been applied with limited success to many domains. These systems face the challenges of adapting to unconstrained interaction with patients, and generalization beyond the training data. Speech recognition and language modeling are also challenges faced in this and other constrained

domains, such as weather or flight booking [27, 68]. Furthermore, the usefulness of a question answering system for patients depends not only on its ability to return relevant answers, but on its ability to present these answers in a manner easily accessible to viewers. Improvements in natural language understanding and generation are integral parts of such systems, which would ideally be able to respond to the kind of unconstrained questions patients might direct to their physicians or pharmacists.

These challenges have been tackled by health dialogue systems; a notable example is Chester, a personal medication advisor prototype developed at the University of Rochester [2]. Chester was designed with the aim of alleviating the increasing burden placed on patients to manage their health and medical treatments, especially in light of the life-threatening complications that may arise from missed pills or drug interactions. Communicating with patients using natural language dialogue makes Chester most accessible to people familiar with the behavior of expert health care providers, and requires minimal training to use. More specialized spoken medical dialogue systems have also been developed, such as Rojas-Barahona et al.'s HomeNL system, which engages in conversation with and offers suggestions to patients who have hypertension [64].

Speech Recognition

An integral part of dialogue systems is speech recognition, which is the process of turning a speech signal into a sequence of recognized words through appropriate representation and the application of acoustic, lexical, and language models. At the acoustic level, a live recognition system must be able to adapt to variations in microphone placements or sound quality. In natural language understanding, difficulties arise from ambiguities in both syntax and word meanings. A given sentence can be produced from multiple parse trees, and the same word has different meanings in different contexts. These problems are compounded with imperfect pronunciation, spelling and punctuation, as is often the case with informal comments posted online. To accurately parse sentences, we must use a combination of semantic rules and probabilistic models. Statistical language models have been found to be very effective

at improving speech recognition without needing complex syntactic rules, by giving more probability to frequently observed word sequences.

However, while acoustic and lexical models are often portable across domains, language models must be more carefully adapted for domain-specific use to achieve higher performance in recognition systems. Adaptation of general language models or cross domain training have been researched, with specific techniques including the use of domain specific corpora [66], model interpolation [88], or training on artificial corpora generated automatically from templates [42].

Of note in such previous research are the steps taken to address the domain-specific data sparsity issues, and the lack of pronunciation data or mispronunciation by users of the system. These health communication systems have also tackled the problem of knowledge representation for the complex relations between drugs, drug effects, and side effects in terms of time and severity.

2.2.2 Health Surveillance

The increased accessibility of public health information through the web has also driven research in text mining for health surveillance. Many Web-based surveillance systems have been developed that focus on event-based monitoring, including the Global Public Health Intelligence Network (GPHIN) [58], HealthMap [25] and BioCaster [12], which gather data from sources such as news reports, official reports, and World Health Organization (WHO) alerts.

BioCaster's system can be decomposed into three major subtasks, namely topic classification, named entity recognition, and event extraction. Document classification was performed using a naive Bayes algorithm, which achieved 94.8% accuracy, and named entity recognition achieved an F-score of 77.0% using a support vector machine. The task faced the challenge of high data volume, the fast response time needed, and out-of-vocabulary terms. It was developed by researchers in Japan, Vietnam, and Thailand, and focuses on Asia-Pacific languages.

These surveillance systems can provide more comprehensive and timely information. For example, GPHIN detected the 2002 outbreak of Severe Acute Respiratory

Syndrome (SARS) through news media analysis three months before official WHO reports [21]. HealthMap, developed in the Harvard-MIT Division of Health Sciences & Technology, mines many online text sources and integrates data from location-aware devices to create a "global disease alert map." It was a useful tool to visualize and track the spread swine flu during the 2009 flu pandemic.

Pharmacovigilance

A special category of health surveillance is pharmacovigilance, or the detection of adverse drug reactions. Postmarketing pharmacovigilance is an area that benefits greatly from NLP methods, as electronic health reports can be analyzed to detect new drug side effects. One of the earliest studies of this kind involved the manual review of patient-reported text comprised of emails sent to the BBC and messages on an online discussion site. Medwara et al. [53] found that the user reports showed a correlation between the antidepressant, paroxetine, and severe withdrawal symptoms and suicide. This study lends support for the use of patient-provided text for detecting drug and drug adverse reaction relationships.

A more recent study conducted on a wider range of drugs show even more promise that user comments contain information that can be used in pharmacovigilance. Leaman et al. [48] studied user comments posted on the DailyStrength[4] health site and found that the incidence of patient-reported side effects were in line with documented incidence from the FDA online drug library. They compared patient comments against a lexicon of medical terms found in the FDA's COSTART vocabulary set.

In another study, Cable [8] manually examined 351 patient-reported comments on statin adverse reactions and found that not only all patients experienced side effects, but more than 60% reported that they discontinued the drug because of the severity of the side effects. While one may question the validity of using self-reported anecdotes rather than controlled studies, in aggregate, anecdotes can provide useful information, as Cable demonstrates. Furthermore, his findings are backed by research literature, described in more detail in section 4.1.1.

[4] `http://www.dailystrength.org`

2.3 Summary

Prior work has focused in part on improving term recognition, one of the largest bottlenecks to medical text mining. The increased availability of electronic health information and the development of medical lexica have enabled a number of projects in personalized medical care and health surveillance. However, to improve the accessibility of health information, we still face the challenge of a large language gap between consumers and clinical documents, and the overwhelming volume of text now available online. In our research, we take a contrasting approach to previous methods, placing emphasis on statistical and parsing techniques, instead of relying on manually created knowledge sources such as the UMLS.

Chapter 3

Data

A large part of the drug reports system is the large database of patient-provided drug reviews and drug experience comments collected from various health-related sites. This corpus of comments will be referred to as the DrugReports corpus hereafter. In this chapter, we describe our data collection process and give an overview of the data collected.

Because of the constant addition of new comments posted to online health sites, we designed a comment collection system that would regularly update the database of comments while being (1) extensible to new sites, (2) easy to configure for new drug classes, and (3) minimal in bandwidth consumption.

3.1 Data Collection

For each web site, data collection is performed with the following steps:

1. Given a search term, URLs of relevant pages are collected.

2. URLs for all search terms are collected and a unique set of URLs are recorded.

3. Web pages corresponding to the URLs are downloaded and cached. Cached web pages which are less than a week old are skipped, to reduce unnecessary network bandwidth usage.

4. Comments are extracted from the HTML pages, along with supplementary information such as author and time posted.

5. The comments are loaded into the database following the schema in Figure 3-1.

3.1.1 Data Sources

Each web site follows a different format, so we implemented site-specific scrapers that collect all comments given the name of a drug. Drug reviews were harvested from five sites dedicated to (or containing sections dedicated to) reviews of pharmaceutical drugs: (1) WebMD[1], (2)Askapatient[2], (3) Medications[3], (4) iGuard[4], and (5) DrugLib[5]. Many of these sites were established almost ten years ago (WebMD and Askapatient), while some were established as recently as 2007 (iGuard). WebMD is one of the largest online health portals, with over 17 million unique monthly visitors in 2007.

These sites each allow users to post reviews of specific drugs, providing comments labeled with the drug name. Some sites encourage users to specify supplementary information such as gender, age, side effects and ratings, similar to product and restaurant review sites. Table 3.1 presents a numerical overview of the collected data with contributions from each site.

Site	Review count	Contribution
WebMD	4124	34%
Askapatient	3960	33%
Medications	3055	25%
iGuard	897	7%
DrugLib	82	1%

Table 3.1: Sources of data and number of reviews of cholesterol lowering drugs.

In addition, many health websites allow users to post general comments in forums,

[1] http://www.webmd.com/
[2] http://www.askapatient.com/
[3] http://www.medications.com/
[4] http://www.iguard.com/
[5] http://www.druglib.com/

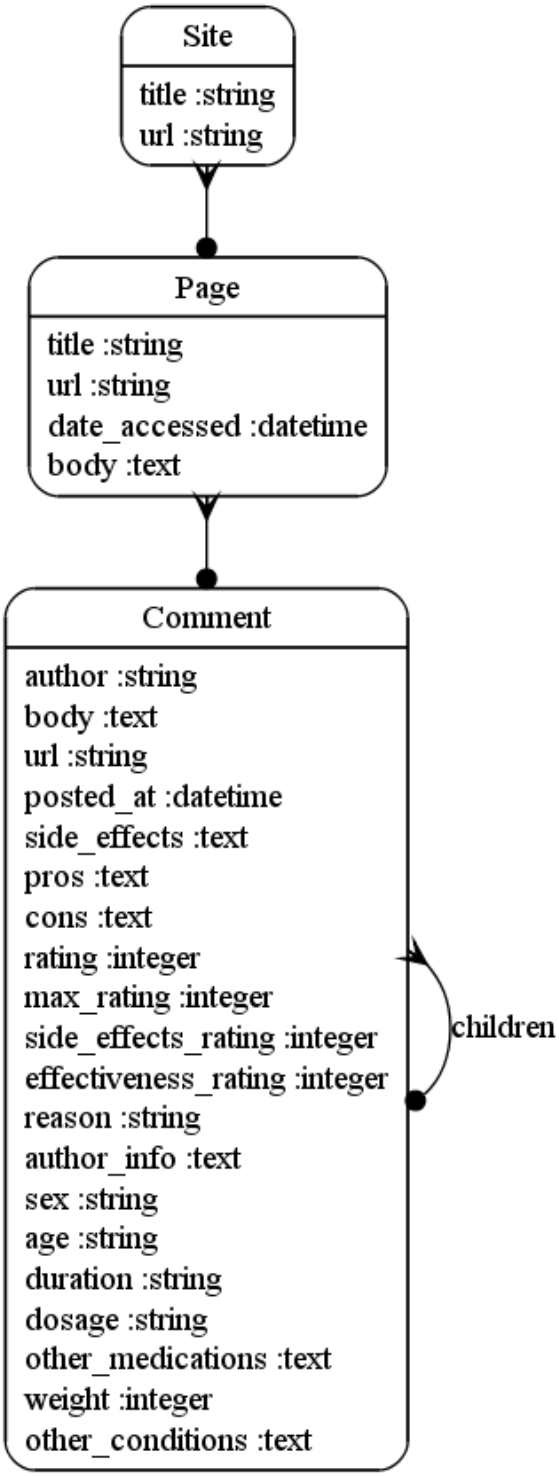

Figure 3-1: Database schema for storing patient comments.

or as responses to articles posted by the site's editors. These sites include: (1) WebMD Blog[6], (2) People's Pharmacy[7], (3) Healing Well[8], and (4) Spacedoc[9]. Most of these are general health web sites with the exception of Spacedoc.net, which has forums focused on cholesterol related drugs. Unlike the sites dedicated to drug reviews, these sites tend to contain comments that are less relevant to specific drugs.

3.1.2 Data Coverage

Because many substances are marketed under country-specific brand names, we collected reviews for all brand names popular in English speaking countries, as well as the generic names. For example, simvastatin is marketed as Zocor in the US and Lipex in Australia. The drug classes covered are separately configured in a file that contains the names of all drugs and the hierarchy. The drug hierarchy is adapted from the Anatomical Therapeutic Chemical (ATC) Classification System, which is managed by the WHO Collaborating Centre for Drug Statistics Methodology, and organizes drugs based on their therapeutic use and chemical characteristics. A portion of the drug hierarchy we use can be found in Appendix A.

For the scope of this thesis, we focused on cholesterol-lowering drugs, which rank among the most prescribed pharmaceuticals ever. Their prevalence allows for a large quantity of patient-reported data. Furthermore, preliminary examination of online medicine and patient forums shows a large number of responses which include reported drug side effects such as muscle weakness and memory loss [1]. We collected a total of over 12,000 reviews about drugs falling under ATC class C10, which includes all lipid modifying drugs. These drugs may be referred to interchangeably as cholesterol lowering drugs. Figure 3-2 presents an overview of the size and distribution of comments over different classes of cholesterol lowering drugs.

[6] http://blogs.webmd.com/
[7] http://www.peoplespharmacy.com/
[8] http://www.healingwell.com/
[9] http://www.spacedoc.net/

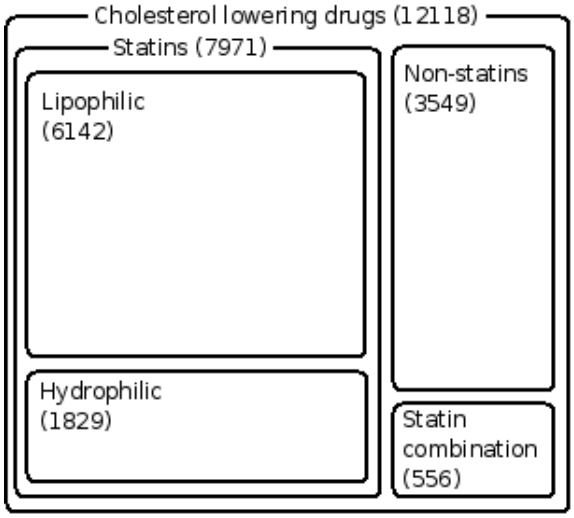

Figure 3-2: Distribution of comments in cholesterol lowering drug class. Numeric values are total number of reviews in each class.

3.2 Example Comments

The comments collected often consist of very detailed descriptions of their drug use and symptom progression. For example, one user who posted on People's Pharmacy shared the following:

> My father was perscribed lipitor in March of 2004, subsequently he developed muscle weakness and numbing and stopped taking it. The weakness did not go away, he got progressively weaker and was recommended to see a neurologist. In September of 2004 the neurologist diagnosed him with ALS ... He died in March of 2005, one month after his birthday and less than one year after taking lipitor.

The above is quite typical of comments posted online, whether on forums or in response to articles relating to statins. They are written in natural language, with a variety of sentence structures, misspellings, or grammar mistakes. Acronyms such as "ALS" (which stands for amyotrophic lateral sclerosis) abound. At the same time, these anecdotes allow users to share more relevant information than can be anticipated by structured forms.

3.3 Spelling Correction

We performed spelling correction on the entire corpus of user comments as a preprocessing step for all NLP tasks, with the goal of correcting words of medical interest that were misspelled frequently by many users. Collected data were first tokenized and case-normalized, and stop words were removed, following a commonly used stopword list [24]. Comments were then processed with automatic spelling correction as described below.

We began with a unique list of all unigrams composed only of the characters a-z. These 20,601 words were first sorted by likelihood of being misspelled based on the log ratio of unigram probabilities between the DrugReports corpus and the Google n-gram corpus[10]. The Google n-gram corpus is a collection of unigrams up to 5-grams with counts collected from public Web pages, and thus contains a wider vocabulary than conventional corpora.

For a given word w, we can define $c_g(w)$ as the count of w in the Google n-gram corpus, and $c_d(w)$ as the count in the DrugReports corpus. Words that have a high ratio of unigram probabilities are either more likely to be misspelled, because they have low or zero $c_g(w)$, or more likely to be medically relevant with a higher $c_d(w)$.

Upon manual inspection, we set a threshold cutoff for the unigram probability ratio at 0.20, resulting in a list of 17,199 unique words. We then further pruned the list of potentially misspelled words by eliminating those that satisfied any of the following conditions:

1. $c_g(w) > 1,000,000$

2. $c_d(w) > 120$

3. w appears in comments from only one site.

4. w appears in an external corpus that is unlikely to contain misspellings.

[10] `http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html`

The count thresholds were manually chosen to eliminate all frequent words that were not misspellings. Words that appeared only on one website (of the nine sites scraped) were removed because they are inherently uninteresting; often these were usernames or repeating character sequences. We also removed words that appeared in a set of commonly used external corpora[11] - the Brown corpus, Project Gutenberg Selections, the Genesis corpus, the Australian Broadcasting Commission corpus, the Reuters corpus, the Wordlist lexicon, and health articles and documents from Google Health, NIH, WebMD, Wikipedia, and iGuard. These published texts were chosen because they are less likely to contain misspellings.

The filtered list contained 3,025 candidate misspelled words. Proposed corrections were automatically generated for these words based on near-miss match to words that appeared at least 8 times in the DrugReports corpus (single-letter substitution, insertion, deletion; two letters inverted). In the case of multiple matches, the word with the highest unigram was chosen. Implausible corrections were discarded after manual inspection, resulting in a final count of 2,678 spelling correction rules. These were then applied to the entire corpus.

[11] `http://code.google.com/p/nltk/wiki/Corpora`

Chapter 4

Automatic Discovery of Side Effects: Focus on Cholesterol-Lowering Drugs

We explore the use of the corpus of patient-provided drug reviews in discovering drug adverse reactions. Patient-provided medical drug experiences can supplement drug adverse reaction findings and address the issue of the large language gap between patients and technical medical documents [93].

Previous work has been conducted to extract drug side effects from text, for example, mining drug package inserts to link drugs to side effects [45] or detecting infectious disease outbreaks by monitoring online news reports [12]. These studies have generally been concerned with technical text. Self-reported data poses a greater NLP challenge because of misspellings, ungrammaticality, and shorthand. While little extensive research has been conducted on patient-reported comments, we can compare with electronic health records, written unedited by clinicians to document patient conditions, that have as high as 10% incidence of misspellings [65]. Studies have also raised the problem of mapping terms in consumer health texts to concepts in UMLS; Divita [18] found that MetaMap Transfer had only a 53% success rate at matching terms in free text to concepts in UMLS. It is possible that patient-provided comments are even more difficult to analyze because, without any medical training,

non-clinicians are more likely to misspell and misuse words, and employ more creative use of language.

Leaman et al. [48] attempt to account for unexpected vocabulary by using the UMLS lexicon, further supplemented with a few colloquial terms, to detect adverse reactions from self-reported online posts. One of their observations was that the frequency of side effects in user comments was highly correlated with their documented frequency as provided by the FDA. Their study is the only one that we are aware of that performs textual analysis of online patient-provided comments.

In this chapter, we use several popular statistical NLP techniques to detect biases in word distributions when comparing reviews of statin drugs with reviews of other cholesterol-lowering drugs. We focus on these drugs because they are widely prescribed and have diverse side effects. We will begin with a review of the research literature reflecting known or suspected side effects associated with cholesterol-lowering drugs. We will then describe the set of statistical NLP techniques we used to detect likely associations between particular drug classes and particular health issues. We verify that many of our extracted associations align with observations from the literature.

4.1 Side Effects of Cholesterol-lowering Drugs: Brief Literature Review

In this section, we briefly review some of the literature on associations between cholesterol-lowering drugs and certain side effects. We will focus our discussion on the important class of HMG coenzyme A reductase inhibitors (statins) which have become increasingly prescribed as very effective agents to normalize serum cholesterol levels. The most popular of these, atorvastatin, marketed under the trade name, Lipitor, has been the highest revenue branded pharmaceutical for the past 6 years[1]. The official Lipitor web site lists as potential side effects mainly muscle pain and weakness and digestive problems. However, several practitioners and researchers have identified

[1]http://www.drugs.com/top200.html

suspected side effects in other more alarming areas, such as heart failure, cognition and memory problems, and even severe neurological diseases such as Parkinson's disease and ALS (Lou Gehrig's disease).

4.1.1 Statin Drugs

It is widely acknowledged that statin drugs cause muscle pain, weakness, and damage [32, 56], likely due in part to their interference with the synthesis of the potent antioxidant Coenzyme Q10 (CoQ10) [47]. CoQ10 plays an essential role in mitochondrial function to produce energy. Congestive heart failure is a condition in which the heart can no longer pump enough blood to the rest of the body, essentially because it is too weak. Because the heart is a muscle, it is plausible that heart muscle weakness could arise from long-term statin usage. Indeed, atorvastatin has been shown to impair ventricular diastolic heart performance [72], and low cholesterol levels were also found to be associated with greater 12-month mortality risk in patients with chronic heart failure [62]. Furthermore, CoQ10 supplementation has been shown to improve cardiac function [57, 86].

The research literature provides plausible biological explanations for a possible association between statin drugs and neuropathy [73, 94]. A recent evidence-based article by Cable [8] found that statin drug users had a high incidence of neurological disorders, especially neuropathy, parasthesia, and neuralgia, and appeared to be at higher risk to the debilitating neurological diseases, ALS and Parkinson's disease. His study was based on careful manual labeling of a set of self-reported accounts from 351 patients. A mechanism for such damage could involve interference with the ability of oligodendrocytes, specialized glial cells in the nervous system, to supply sufficient cholesterol to the myelin sheath surrounding nerve axons. Higher serum cholesterol levels have been correlated with prolonged survival in patients diagnosed with ALS [19]. Sim et al. [74] showed that statin drugs lead to recruitment of large numbers of glial progenitor cells to mature into oligodendrocytes, likely because of a reduced efficiency of the pre-existing oligodendrocytes. Genetically-engineered mice with defective oligodendrocytes exhibit visible pathologies in the myelin sheath which

manifest as muscle twitches and tremors [67].

Cholesterol depletion in the brain would be expected to lead to pathologies in neuron signal transport, due not only to defective myelin sheath but also to interference with signal transport across synapses [81]. Cognitive impairment, memory loss, mental confusion, and depression were significantly present in Cable's patient population [8]. Wagstaff et al. [84] conducted a survey of cognitive dysfunction from AERS data, and found evidence of both short-term memory loss and amnesia associated with statin usage. Golomb et al. [29] conducted a study to evaluate evidence of statin-induced cognitive, mood or behavioral changes in patients. She concluded with a plea for studies that "more clearly establish the impact of hydrophilic and lipophilic statins on cognition, aggression, and serotonin." It is anticipated that lipophilic statins would be more likely to cross the blood-brain barrier and therefore induce more neurological problems.

Wainwright et al. [85] provide compelling arguments for the diverse side effects of statins, and attribute them mainly to cholesterol depletion in cell membranes. Another study by Goldstein and Mascitelli [28] found that in cardiovascular patients, those taking statins are at a 9% higher risk of developing diabetes compared to those on a placebo. Statins have also been linked to decreased serotonin levels [14], and thus depression, as well as decreased testosterone [16], which may affect male sexual response.

4.1.2 Non-Statin Cholesterol-Lowering Drugs

The four main alternatives to statin drugs for improving lipid profile are fibrates, bile acid sequestrants (such as Questran and Welchol), nicotinic acid (niacin) derivatives and ezetimibe, which interferes with the absorption of cholesterol through the gut. The main side effect associated with niacin is the so-called "niacin flush." A biological explanation for its cause is provided in [33]. Patients taking ezetimibe can experience abdominal or back pain, diarrhea, joint pain, and sinusitis. Rare side effects include coughing, fatigue, sore throat, sexual dysfunction and viral infection[2]. A popular drug

[2]http://www.zetia.com/ezetimibe/zetia/consumer/index.jsp

combination is Vytorin, which contains simvastatin (a statin) combined with Zetia. Possible side effects are rash, pancreatic inflammation, nausea, headache, dizziness, gallstones, gallbladder inflammation, and swelling of the face, lips, tongue, and throat.

4.2 Data

We use data from drugs affecting the cardiovascular system, specifically those falling under ATC class C10, which includes all lipid modifying drugs. Statin drugs and other cholesterol-lowering drugs belong in this class. In addition, we collected data on drugs used to treat hypertension (ATC class C09), which serves as a fair corpus for comparison with cholesterol-lowering drugs, as it also affects the cardiovascular system.

The sites that these reviews were drawn from include all sites that contain labeled drug reviews, as seen in Table 3.1.

4.3 Methods

Our goal was to assess the usefulness of patient-reported free-text drug reviews in determining the side effects and areas of concern associated with certain drugs. We compared two mutually exclusive drug classes at one time, for example, statin drugs and other non-statin cholesterol lowering drugs. Such a comparison should highlight the side effects more associated with statin drugs than other drugs used for the same purpose of improving lipid profile. By comparing drugs within the same class, we can highlight features that distinguish two drugs that are used for the same purpose, thus controlling for patient preconditions.

We map our problem onto the general task of measuring association between two discrete random variables, X and Y. In our case, $P(X = x)$ is the probability of a term x being contained in any document. $P(Y = y)$ is the proportion of documents in a given class (e.g. statin). $P(x, y)$ is the probability that any given document is both in class y and contains term x. Terms can be n-grams with $n \leq 5$.

Association measures have been used extensively for collocation identification [11], sentence boundary detection [91] and word sense disambiguation purposes [63]. From an information-theoretic perspective, our problem maps well to the approach taken by [63] for word sense disambiguation by characterizing the co-occurrence of predicates with conceptual classes. We define the measures we use below, along with brief explanations of their adaptation to our problem.

4.3.1 Log Likelihood Statistic

Dunning's likelihood ratio test [20] is a statistical tool used to compare the homogeneity of two independent binomial distributions. It follows the χ^2 distribution with one degree of freedom, but unlike the χ^2 test, has the benefit of being robust to non-normal and low-volume data. We derive the likelihood ratio below.

Suppose a document has a probability p of containing the term x and we observe k documents of n total containing at least one instance of x. We can express the likelihood of this observation as the result of a repeated Bernoulli trial:

$$H(x) = p^k(1-p)^{n-k}\binom{n}{k} \tag{4.1}$$

With the log likelihood ratio (LLR), we compare the maximum values of the likelihoods of the null hypothesis (H_0) of there being a single probability p that explains both classes with the likelihood of two classes having different probabilities p_1 and p_2 of containing the term x (H_1). The likelihoods of these two hypotheses are expressed in Equations 4.2 and 4.3.

$$H_0(x) = p^{k_1+k_2}(1-p)^{n_1-k_1+n_2-k_2}\binom{n_1}{k_1}\binom{n_2}{k_2} \tag{4.2}$$

$$H_1(x) = p_1^{k_1}(1-p_1)^{n_1-k_1}\binom{n_1}{k_1}p_2^{k_2}(1-p_2)^{n_2-k_2}\binom{n_2}{k_2} \tag{4.3}$$

The log likelihood ratio is then defined as:

$$LLR(x) = \sum_{i \in \{1,2\}} k_i \log \frac{p_i}{p} + (n_i - k_i) \log \frac{1 - p_i}{1 - p} \tag{4.4}$$

where p and p_i are the values that maximize the likelihoods, i.e.:

$$p = \frac{k_1 + k_2}{n_1 + n_2}, p_i = \frac{k_i}{n_i}$$

To avoid division by zero and to compensate for sparse data, we used add-one smoothing scaled by the data set size.

Because the log likelihood statistic only tells us how unlikely it is that the two classes of documents have the same probability of containing the term x, we further define here a *class preference* measure, obtained by splitting the log likelihood ratio into two terms. The first term, defined in Equation 4.5, collects the terms associated with class 1. A symmetrical calculation can be made for class 2. The difference between these two terms is a measure of class preference.

$$A_1 = k_1 \log \frac{p_1}{p} + (n_2 - k_2) \log \frac{1 - p_2}{1 - p} \tag{4.5}$$

4.3.2 Pointwise Mutual Information

Commonly used in information theory, pointwise mutual information allows us to quantify the association between the two discrete random variables associated with outcomes x and y:

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \tag{4.6}$$

Furthermore, the ratio between $PMI(x, y_1)$ and $PMI(x, y_2)$ (i.e. the difference) can tell us which words are more closely associated with one class than another, much as the semantic orientation of words was calculated by Turney [82].

4.3.3 Set Operations

We also include two set operation based measures - Dice and Jaccard coefficients. Let D_x and D_y be two sets of documents containing the term x and relating to drug class y, respectively. Dice's coefficient calculates their similarity as follows:

$$Dice(x, y) = \frac{2|D_x \cap D_y|}{|D_x| + |D_y|} \tag{4.7}$$

The Jaccard coefficient is defined as:

$$Jaccard(x, y) = \frac{|D_x \cap D_y|}{|D_x \cup D_y|} \tag{4.8}$$

The preference of a term x for class y_1 over class y_2 can be found as a ratio between $Dice(x, y_1)$ and $Dice(x, y_2)$, or the Jaccard coefficients.

4.4 Results

Below, we will highlight some of the most interesting results that emerge from comparisons of various data sets.

4.4.1 Cholesterol-lowering vs Blood-pressure-lowering Drugs

Terms related to muscle pain and weakness and memory problems were far more common for the cholesterol-lowering drugs, as well as more unexpected words like *arthritis*, *joint pain* and *spasms*. Blood pressure drugs had a much more frequent appearance of words related to the cough associated with ACE inhibitors, such as *chronic cough*, *hacking*, *throat*, etc. *Sex drive* and *dizziness* were also prominent for blood pressure drugs. Selected terms can be found in Table 4.1.

Term	k_1	k_2	LLR	p-value
cholesterol	3108	91	3644.78	$0^\star$
arthritis	325	86	128.39	9.22E-30
spasms	212	56	83.42	6.63E-20
joint pain	560	293	63.78	1.39E-15
cough	66	2583	3644.78	$0^\star$
blood pressure	292	2556	2573.64	$0^\star$
throat	160	745	485.9	1.11E-107
hacking	3	219	299.32	4.63E-67
dizziness	376	821	226.14	4.14E-51
chronic cough	3	66	77.33	1.45E-18
sex drive	124	181	17.07	3.60E-05

Table 4.1: Selected words and phrases that distributed differently over cholesterol-lowering drug reviews and renin-angiotensin drug reviews. The log-likelihood ratio (LLR) and p-value are provided. k_1: cholesterol-lowering drugs. k_2: renin-angiotensin drugs. *Values are essentially 0 ($< 1E-300$).

4.4.2 Statins vs Non-statins

Within the cholesterol-lowering drug class, we compared the set of 7,971 statin reviews with 3,549 non-statin reviews. Table 4.2 shows the top 20 terms associated with statins, ranked by each of the association measures discussed in Section 4.3. Table 4.3 presents the terms for non-statin cholesterol-lowering drugs. The rankings from these measures exhibit high correlation with one another.

Gastrointestinal issues and rashes are common to patients taking other cholesterol-lowering drugs. These findings are in line with the expected side effects of niacin derivatives, fibrates, and ezetimibe, which dominate the non-statin reviews.

The drug names can be used as a reference against which to compare the other terms. The fact that *pain* appears between *lipitor* and *zocor* shows that pain is strongly associated with statins in the drug reviews. The list is highly dominated by unigrams because of data sparsity. Methods to better treat low count data may be an area of further investigation.

Table 4.4 highlights a few terms that are highly associated with either the statin or the non-statin class, ranked by the log likelihood ratio expressed in Equation 4.4. The class preference measure determines whether the term was more associated with

Rank	PMI Ratio	LL Ratio	Dice	Jaccard
1	lipitor	lipitor	lipitor	lipitor
2	short term memory loss	pain	zocor	pain
3	pain	zocor	simvastatin	zocor
4	short term memory	simvastatin	pain	muscle
5	zocor	muscle	crestor	simvastatin
6	muscle	crestor	memory	crestor
7	term memory loss	memory	muscle	cholesterol
8	simvastatin	loss	loss	loss
9	crestor	memory loss	walk	memory
10	memory loss	walk	cholesterol	legs
11	muscle pain	cholesterol	memory loss	walk
12	term memory	pravachol	legs	symptoms
13	cholesterol	legs	symptoms	taking
14	memory	pains	pains	drug
15	loss	left	left	pains
16	symptoms	symptoms	feet	muscle pain
17	legs	feet	statin	left
18	walk	walking	muscle pain	feet
19	pains	term memory	muscles	muscles
20	left	short term memory	walking	statin

Table 4.2: Twenty terms with highest class preference for statin drug reviews.

Rank	PMI Ratio	LL Ratio	Dice	Jaccard
1	niaspan	niaspan	niaspan	niaspan
2	flushing	flushing	flushing	flushing
3	tricor	tricor	tricor	tricor
4	zetia	aspirin	aspirin	itching
5	itching	itching	itching	zetia
6	aspirin	zetia	zetia	aspirin
7	welchol	welchol	welchol	welchol
8	low fat snack	fire	fire	fire
9	taking tricor	niacin	niacin	triglycerides
10	niaspan er	sunburn	triglycerides	niacin
11	niacin	snack	burning	burning
12	burning	triglycerides	flush	flush
13	triglycerides	flush	taking tricor	skin
14	fire	burning	sunburn	bedtime
15	sunburn	niaspan er	snack	reaction
16	baby aspirin	benadryl	bedtime	sunburn
17	flush	trilipix	skin	diarrhea
18	snack	gallbladder	reaction	woke
19	chronic diarrhea	bedtime	diarrhea	snack
20	night	applesauce	woke	bathroom

Table 4.3: Terms with high class preference for non-statin cholesterol-lowering drug reviews.

Term	k_1	k_2	LLR	p-value
memory loss	318	11	166.2	5.1E-38
muscle pain	864	196	89.0	3.9E-21
depression	335	56	58.4	2.1E-14
muscle weakness	257	62	21.3	4.0E-06
als	38	1	21.0	4.7E-06
hair loss	126	26	14.9	1.1E-04
diabetes	133	31	11.9	5.6E-04
heart failure	24	1	11.6	6.7E-04
parkinson's disease	19	1	8.4	3.8E-03
chronic diarrhea	3	44	84.2	4.6E-20
gall bladder	16	44	46.3	9.9E-12
rash	127	121	36.1	1.8E-09
severe itching	14	35	34.5	4.3E-09

Table 4.4: Selected words and phrases that distributed differently over statin and non-statin cholesterol lowering drug classes. The log-likelihood ratio (LLR) and p-value are provided. k_1 and k_2: number of statin and non-statin reviews containing the term, respectively. The upper set are far more common in statin drug reviews, whereas the lower set are more frequent in non-statin reviews.

statins or non-statin cholesterol lowering drugs. Many memory and muscle-related issues are more apparent with patients taking statins. The highly significant results for *diabetes* are in line with recent concern about the possibility that statins may increase risk to diabetes [31]. *Depression* also exhibits a significant bias towards statins. This effect may be attributable to their known interference with serotonin receptors [70]. *Heart failure* was also much more common in the statin drug branch, consistent with the findings of Silver et al. [72].

4.4.3 Gender Differences

We compared the reviews posted by males and females taking statin drugs. A large portion of the reviews collected were labeled with gender, with 2,770 female and 2,156 male reviews. While it is possible that gender-specific word choice may influence the term distributions, females clearly had more problems with neuromuscular disorders, including *muscle spasms*, *trouble walking* and *fibromyalgia*. This is in line with ob-

servations from the literature [34]. The prevalence of terms relating to libido among males is possibly due to the fact that statins interfere with testosterone synthesis from cholesterol [79]. Selected terms are shown in Table 4.5.

Term	k_1	k_2	LLR	p-value
sex drive	50	16	28.3	1.0E-07
libido	38	15	17.1	3.6E-05
soreness	69	44	13.9	1.9E-04
fibromyalgia	6	42	22.3	2.3E-06
cramps	139	264	15.7	7.6E-05
muscle spasms	11	38	9.8	1.7E-03
trouble walking	0	11	9.7	1.9E-03
arthritis	46	94	7.2	7.5E-03

Table 4.5: Selected words and phrases in the statin reviews that distributed differently over gender. k_1: male reviews. k_2: female reviews.

4.4.4 Lipophilic vs Hydrophilic Statins

For this comparison, we were most interested in the supposition that lipophilic statins may have a greater impact on the nervous system, particularly on oligodendrocytes, as discussed in Section 4.1. We consider statins with a positive lipophilicity to be *lipophilic*, and negative lipophilicity to be *hydrophilic*. Of the widely prescribed statins, atorvastatin (Lipitor) and simvastatin are both lipophilic, while rosuvastatin is hydrophilic [89]. Results were striking in that the severe neurological disorders, ALS and Parkinson's, occurred almost exclusively in comments associated with the lipophilic class. Selected terms can be found in Table 4.6.

4.5 Discussion

The results of these experiments show that corpus comparison methods can identify side effects and areas of concern that are more associated with one class of drugs

Term	k_1	k_2	LLR	p-value
tingling	278	47	14.61	1.32E-04
tremors	38	1	13.32	2.63E-04
parkinson's	29	0	13.01	3.10E-04
als	35	3	5.98	1.44E-02
neurological	16	0	6.55	1.05E-02

Table 4.6: Selected words that were more common in lipophilic than in hydrophilic statin reviews. k_1: lipophilic statin reviews. k_2: hydrophilic statin reviews.

than another. One initial concern was that it may be difficult to distinguish between patient preconditions and side effects using a bag-of-words approach. For example, a patient might state "I took Lipitor because I had *high cholesterol* but it caused *muscle aches*." However, by comparing drug classes used for the same purpose (e.g. of lowering cholesterol), we control for preconditions which should distribute evenly across both classes.

The highly ranked terms are those that not only appear frequently in one class, but also are more skewed to one class than another. A patient who takes statins, for example, is more likely to experience muscle pain than a patient who takes another cholesterol-lowering drug, such as niaspan, because the class preference of the term *muscle pain* is skewed toward statins. However, a patient taking statins is not necessarily more likely to experience *memory loss* than *muscle pain*, even though *memory loss* appears higher on the ranked list of terms that prefer statin drug reviews. What this means instead is that the skew in the two data sets on *memory loss* is greater than it is on *muscle pain*.

4.5.1 Limitations

While our study used only term and drug class co-occurrence, we believe further improvements can be made to side effect detection using parsing. For example, consider the term *heart failure*. In the context below, it is part of a general statement someone is making, based not on personal experience, but hearsay:

> ...statins are costly, marginally effective, and rife with adverse effects.

> Common side effects of statin drugs include muscle pain and weakness and liver problems. However, they are also linked with memory problems, heart failure, and increased risk of death...

This comment suggests potential side effects that the user did not personally experience. Whether the number of such comments significantly inflates the saliency of side effects should be further investigated. Even when a term does appear in the context of personal experience, it may be an existing precondition:

> I am a 58 year old male diagnosed with heart failure and afib in Jan 2004. I have been taking a combination of Lipitor, Topral, Hyzaar, Pacerone and Magnesium and Potassium supplements since then...

We want to distinguish between existing preconditions and cases of interest where the term is mentioned as a clear consequence of taking the drug, such as in the following comment:

> I haved been on Lipitor for a number of years with many of the side effects posted here. I have had Heart Failure fo a year now ... i am off lipitor an taking 400mg of coq10 per day. i am now in day seven an have slept in my own bed with my wife for the first time in a year. i am less restless, an have ha no recurrence of heart failure.

4.6 Summary

In this chapter, we have described a basic strategy of comparing word frequency distributions between two databases with highly similar topics – e.g., statin and non-statin cholesterol lowering therapies – as a means to uncover statistically salient phrase patterns. Our efforts focused on statin drugs, as these are a widely prescribed medication with diverse side effects. We uncovered a statistically significant association of statin drugs with a broad spectrum of health issues, including memory problems, neurological conditions, mood disorders, arthritis and diabetes, in addition to very common

complaints of muscle pain and weakness. Many of our findings are supported by the research literature on statins.

These experiments were inspired by the study conducted by Jeff Cable [8]. While he looked at only 350 reviews, he used careful manual analysis to deduce associated side effects. We looked at a much larger set of reviews (over 12,000), and used statistical NLP techniques for analysis. On the one hand, it is gratifying that both methods uncovered similar side-effect profiles on different data. On the other hand, it is disturbing that a drug class as widely prescribed as the statin drugs has such severe and sometimes life-threatening adverse reactions.

Chapter 5

Speech Recognition Experiments

As part of the drug reports system, users will have the ability to interact using natural language, making the system more engaging by better emulating interactions with human experts. We would like to allow the system to support queries beyond simple key word searching. Part of the challenge of applying speech recognition and language modeling techniques in the medical domain is the limited coverage that general lexica have for specialized words and pronunciations. General language and lexical models need to be updated to include drug and disease names, and their pronunciations. Recognition must also be robust to mispronunciations when users often do not know the right pronunciation, even when it is available. In this chapter, we present the results of preliminary experiments conducted to develop a language model for recognizing questions a user might ask relating to medical drugs and symptoms.

5.1 Collection of Spoken Questions Data

We collected spoken utterances relevant to the domain with Amazon Mechanical Turk[1] (AMT). AMT is a crowdsourcing tool has been used extensively by researchers to collect large amounts of data in a quick and cost-efficient manner, especially for natural language processing tasks. For example, it has been used to evaluate translation quality [9], annotate data [78], and transcribe spoken language [51].

[1] `www.mturk.com`

We collected the data in two stages. First, a task was created in which workers were asked to read an anecdote about a statin drug experience, and then come up with questions that the anecdote might answer. The anecdotes were drawn from snippets of comments collected online. An example prompt is shown in Figure 5-1, and sample anecdotes can be found in Appendix B.

Ask 2 questions about cholesterol related drug experiences

Imagine that there exists a large set of patient-reported anecdotes about medical drug experiences, specifically relating to cholesterol-lowering drugs (statins). Imagine also that a service is available that allows you to ask questions related to drug experiences and will provide you with a set of relevant anecdotes to browse.

Your task is to:

1. Read the following anecdote about a statin drug (or statin drugs).
2. Come up with two questions about the drug that might be answered by the anecdote.

Please remember:

- The questions must use standard English and spelling.
- The questions must relate to statin drugs or cholesterol-related health problems.
- Try to phrase the questions in a variety of different ways.

Figure 5-1: Prompt presented to Amazon Mechanical Turk workers to collect sample questions about cholesterol-lowering drug experiences.

In the second stage, speech data were collected from native speakers of American English by asking another group of turkers to read the questions posed earlier. The use of Amazon Mechanical Turk was a cost-effective way to collect speech data. Of the over 4500 utterances collected, only 40 were unusable due to recording noise or non-native pronunciation. Sample questions can be found in Appendix C.1.

In addition, turkers were asked to imagine that they were taking a new drug, and to come up with questions they would ask to a group of people who had experience taking that drug. From this task, we collected a set of less constrained questions in text format. Sample questions can be found in Appendix C.2.

From the AMT tasks, a total of 935 spoken questions relating to statins were collected. An additional 318 general drug-related questions were collected in text format only. Speech data were collected only for the statin questions because the speech recognition tasks were primarily focused on statins and cholesterol.

5.2 Methods

To perform the speech recognition, we used the SUMMIT speech recognizer developed in our group [95]. The SUMMIT recognizer works by composing a series of finite state transducers modeling the acoustic information, the context dependent phones, the pronunciation rules mapping phones to phonemes, the lexicon, and the grammar. In adapting the models to the medical domain, we made changes mainly to the lexicon, by adding pronunciations for words not found in the vocabulary, and developed a domain-specific trigram language model.

5.2.1 Trigram Language Model

An n-gram language model predicts the most likely word given a history of n words. This can be expressed as a probability:

$$P(w_i|w_{i-1}, w_{i-2}, \ldots, w_{i-n}) \tag{5.1}$$

The maximum likelihood estimation of these probabilities is based on the observed counts of these n-grams in the training corpus:

$$\begin{aligned} P_{ML}(w_i) &= \frac{count(w_{i-n}, \ldots, w_{i-2}, w_{i-1}, w_i)}{\sum_{w \in V} count(w_{i-n}, \ldots, w_{i-2}, w_{i-1}, w)} \\ &= \frac{count(w_{i-n}, \ldots, w_{i-2}, w_{i-1}, w_i)}{count(w_{i-n}, \ldots, w_{i-2}, w_{i-1})} \end{aligned} \tag{5.2}$$

where V is the vocabulary, or the set of unique words that appear in the training data. The language model used was based on trigrams, which is probably the most dominant language model used today.

5.2.2 Data Sparsity

Given that this project concerns a new domain, we face issues with sparse data. Maximum likelihood models often place too much emphasis on the training data given, and do not generalize well to unseen word sequences.

Smoothing

Smoothing techniques help to alleviate the problem of data sparsity by redistributing probability mass from observed n-grams to events that are unobserved in the training corpus. We used Kneser-Ney discounting, in which rare n-grams have probabilities that back off to lower-order n-grams. In a trigram model, rare trigram probabilities will back off to the probability of the bigram, based on how many contexts the word appears in.

Class N-gram Models

In addition to smoothing, we also used class n-grams to deal with the data sparsity problem. Selected words were assigned to each class, and n-gram probabilities were calculated using counts of class sequences. The class-based n-gram calculates word probabilities as follows:

$$\begin{aligned} & P(w_i|w_{i-1}, w_{i-2}) \\ = \quad & P(w_i|c(w_i)) \times P(c(w_i)|c(w_{i-1}), c(w_{i-2})) \end{aligned} \tag{5.3}$$

where $c(w)$ is the class that word w belongs to.

Using class n-grams allows us to easily incorporate semantic information into models based heavily on statistics. Furthermore, this allows us to better predict words that do not appear frequently in the training corpus, but that belong to the same class as more frequent words.

The classes used in training the class n-gram models were manually created by forming rules for words that were found to be significant in the corpus. Table 5.1 lists the classes used and some representative word members.

Table 5.1: Classes used for class n-gram training.

Class	Words
statins	lipitor, zocor, baycol, simvastatin, crestor, vytorin, lovastatin, tricor, pravachol
body parts	shoulder, arm, fingers, muscle, leg, tendon, thigh
symptoms	anxiety, numbness pain, tingling, soreness, fatigue, ache, exhaustion
diseases	parkinson's, polio, alzheimer's

Supplementary Training Data

The high cost of acquiring speech data for this new domain was a limiting factor on the amount of training data available for generating these language models. However, the language model training data does not need to come solely from the spoken questions collected. We also used text data to train the language models, including the comments that inspired the questions (665 utterances), the general drug questions (318 utterances), and the Michigan Corpus of Academic Spoken English (MiCASE) transcripts (96246 utterances), a general spoken English corpus containing transcripts from lectures, classroom discussions, and advising sessions, among other general speech activities [75].

5.3 Results and Discussion

Five-fold cross validation was performed and the word error rate (WER) in both the training and test sets were compared. The baseline recognizer simply trained a trigram language model on 80% of the data and was tested on the remaining 20%, achieving 44.84% WER. In table 5.2, we can see that using a class trigram model improved the recognizer to a 44.04% WER.

Class n-gram	WER (train)	WER (test)
no	26.46	44.84
yes	26.96	44.04

Table 5.2: The use of class n-grams slightly improves recognizer performance.

Next, the performance of class trigram models trained only on the training data was compared to language models trained with supplementary texts. Various combinations of supplementary texts were tested. For each supplementary text, I tested allowing only sentences with in-vocabulary words, and allowing all words, including those that were out of the vocabulary of the training questions (OOV words). Table 5.3 summarizes the findings.

Allow OOV	Add. corpus	WER (train)	WER (test)
yes	Drug comments	30.11	43.70
no	Drug comments	26.98	43.90
yes	Gen. questions	27.88	43.24
no	Gen. questions	26.92	43.86
yes	Gen. questions, Drug comments	30.08	43.02
no	Gen. questions, Drug comments	26.94	43.84
yes	Gen. questions, Drug comments, MiCASE	49.64	59.42
no	Gen. questions, Drug comments, MiCASE	28.98	46.66

Table 5.3: Word error rate for various training sets. Additional corpora were used to train the language model, including the comments about statins collected from online forums (and were then used to prompt turkers to ask questions), general medicine-related questions, and the MiCASE corpus.

The use of both additional drug-related questions and the comments which inspired the statin-related questions improved the performance of the recognizer. These additional corpora both add to the types of sentence structure on which the language model is trained. We may observe the same phrasing in general drug questions as those posed specifically regarding statins. The statin-related questions of interest may

also have been phrased in a manner similar to the comments that the turkers first read. With limited training data, these additional corpora help the language model generalize and perform with anywhere from a 0.34% to 1.02% decrease in WER.

When the MiCASE corpus was added, we observed a dramatic drop in recognition performance, because the language model is overwhelmed by irrelevant data, which does not aid in predicting words for statin-related questions. Notice that the performance improves when we limit the additional text to only in-vocabulary sentences in the case of the MiCASE corpus. The opposite effect is seen with the drug comments corpus and the general medicine questions corpora. Performance improvements in the recognizer are only seen when the additional training corpora contain sentences and sentence structure that relate to the recognition task.

Word error rates for the spoken question data were generally in the range of 40-50% for test data using language models trained on a subset of the data. The best performing training conditions used both a class n-gram and supplementary corpora of both the online patient comments regarding statins and the general medical questions, which resulted in nearly a 2% decrease in word error rates.

While the word error rates may seem high, the recognizer erred mostly on common words, or plurality. The ability of the recognizer to identify important words - drug names, symptoms - shows that it is still useful for our purposes of answering drug-related questions. Some of these recognition problems can likely be overcome by using a syntactic grammar to give higher probabilities to grammatical sentences, which is part of an on-going investigation.

5.4 Summary

We presented the preliminary experiments on recognition of spoken queries to the system. Methods to improve speech recognition through improved language modeling were explored. The use of class-based trigrams demonstrated an improvement over regular trigrams. Training on supplementary corpora related to statins and general drugs led to modest performance increases.

Chapter 6

Additional Preliminary Experiments

This chapter presents a series of additional experiments conducted with the DrugReports data. We begin with a comparison of term identification methods, then show the results from classification of the cholesterol-lowering drug reviews, and finally demonstrate the application of LDA to automatically cluster related terms.

6.1 Multi-word Term Identification

In this section, we present some common methods of term extraction and preliminary results. Term extraction is a process of automatically identifying multi-word units (MWUs), or a group of two or more words that form a meaningful phrase. It is a useful preprocessing step for tasks such as information retrieval to return relevant documents [59], natural language generation [77], and parsing [87]. In our research, it is used for topic identification with LDA, feature generation for classification, and parsing.

The methods shown below are easily applicable to any n-grams, however we only present detailed information for bigrams.

Rank	Bigram	Count	Rank	Bigram	Count
1	i have	10455	11	to be	3352
2	i am	8629	12	on the	3189
3	i was	6612	13	have been	3093
4	in the	6025	14	that i	3041
5	of the	5254	15	for the	2966
6	i had	5070	16	when i	2956
7	and i	4687	17	have a	2894
8	to the	3899	18	it was	2865
9	it is	3827	19	but i	2714
10	in my	3442	20	have to	2637

Table 6.1: Bigrams ranked by frequency.

6.1.1 Term Frequency

The simplest method of finding multi-word terms is by finding terms that appear the most frequently. Using this method, many uninteresting terms appear because they contain common words, as seen in Table 6.1. By simply filtering out stop words, we can improve the candidate bigrams, as shown in Table 6.2.

Rank	Bigram	Count	Rank	Bigram	Count
1	**side effects**	1736	11	go back	443
2	take care	1017	12	2 years	437
3	don't know	956	13	**fish oil**	419
4	years ago	946	14	**coq 10**	417
5	**blood pressure**	697	15	much better	412
6	**heart attack**	599	16	started taking	407
7	**muscle pain**	577	17	stopped taking	394
8	feel like	546	18	40 mg	380
9	year old	525	19	sounds like	379
10	**side effect**	486	20	every day	377

Table 6.2: Bigrams ranked by frequency with stop words removed.

6.1.2 Part of Speech Filter

Justeson and Katz [40] pass candidate terms through a part-of-speech filter to achieve a huge improvement. They suggest patterns with examples, which we list briefly in Table 6.3. The letters A, N, and P represent adjective, noun, and preposition, respectively.

Pattern	Example
AN	linear function
NN	regression coefficients
AAN	Gaussian random variable
ANN	cumulative distribution function
NAN	mean squared error
NNN	class probability function
NPN	degrees of freedom

Table 6.3: Example part of speech patterns for terminology extraction.

When we apply a manual part of speech filter to the stoplist filtered terms, we see much better results. The top ranked bigrams can be seen in Table 6.4. Other than temporal and measure terms, the top bigrams are all valid terms. The difficulty with this method is that many unknown words may not be recognized by a part of speech tagger.

Rank	Bigram	Count	Rank	Bigram	Count
1	**side effects**	1736	11	**blood sugar**	366
2	**blood pressure**	697	12	20 mg	356
3	**heart attack**	599	13	10 mg	354
4	**muscle pain**	577	14	3 months	351
5	**side effect**	486	15	**heart disease**	344
6	2 years	437	16	**acid reflux**	337
7	**fish oil**	419	17	**vitamin d**	337
8	40 mg	380	18	6 months	335
9	every day	377	19	last night	324
10	**high cholesterol**	375	20	2 weeks	324

Table 6.4: Bigrams passed through a part of speech pattern filter.

Passing through a character filter, that only allows the letters a-z, achieves much

better results, as seen in Table 6.5

Rank	Bigram	Count	Rank	Bigram	Count
1	**side effects**	1736	11	**vitamin d**	337
2	**blood pressure**	697	12	**acid reflux**	337
3	**heart attack**	599	13	last night	324
4	**muscle pain**	577	14	**statin drugs**	317
5	**side effect**	486	15	long time	296
6	**fish oil**	419	16	**chest pain**	294
7	**every day**	377	17	first time	262
8	**high cholesterol**	375	18	many people	259
9	**blood sugar**	366	19	**high blood**	252
10	**heart disease**	344	20	**blood work**	244

Table 6.5: Bigrams passed through a part of speech pattern filter and containing only letters a-z.

6.1.3 Association Measures

Purely statistical measures can be used to extract terms. Below, we define some commonly used association measures given a bigram, $[w_1, w_2]$.

Pointwise Mutual Information

Pointwise Mutual Information, defined in Equation 6.1, was first defined by Fano [22] and has been used by Church and Hanks [11] to find word association norms and Smadja et al. [76] to find collocations for translation purposes.

Highly ranked bigrams can be seen in Table 6.6, where bolded terms are valid multi-word units.

$$I(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{6.1}$$

Symmetrical Conditional Probability

Silva [71] introduced the Symmetrical Conditional Probability (SCP) of bigrams, which they showed to have the highest precision in detecting multi-word units when

Rank	Bigram	PMI	Rank	Bigram	PMI
1	**alpha lipoic**	17.0	11	panic resources	12.6
2	**carpal tunnel**	14.8	12	**ct scan**	12.3
3	**coenzyme q**	13.7	13	million dollars	12.2
4	**peripheral neuropathy**	13.5	14	**nurse practioner**	12.2
5	**stretching exercises**	13.3	15	**cell phone**	12.2
6	**horror stories**	13.0	16	**dark urine**	12.0
7	**tennis elbow**	12.8	17	play tennis	11.9
8	greatly appreciated	12.7	18	**cold turkey**	11.7
9	**contributing factor**	12.6	19	**sudden onset**	11.6
10	**law suit**	12.6	20	**medical profession**	11.6

Table 6.6: Bigrams ranked by pointwise mutual information.

compared to other measures such as PMI, Dunning's log likelihood statistic, and the Dice coefficient. The SCP measure is defined in Equation 6.2.

Highly ranked bigrams can be seen in Table 6.7, where bolded terms are valid multi-word units.

$$SCP(w_1, w_2) = \frac{p(w_1, w_2)^2}{p(w_1)p(w_2)} \tag{6.2}$$

Rank	Bigram	PMI	Rank	Bigram	PMI
1	**carpal tunnel**	9.1	11	**acid reflux**	1.7
2	**side effects**	6.0	12	**q -10**	1.6
3	**alpha lipoic**	5.3	13	**heart attack**	1.6
4	**fish oil**	4.8	14	panic resources	1.3
5	**coenzyme q**	2.8	15	**side effect**	1.1
6	**blood pressure**	2.7	16	greatly appreciated	1.1
7	**peripheral neuropathy**	2.6	17	years ago	1.1
8	**coq 10**	2.5	18	take care	1.1
9	**ct scan**	2.4	19	**memory loss**	1.0
10	**vitamin d**	1.8	20	year old	0.9

Table 6.7: Bigrams ranked by symmetric conditional probability.

6.1.4 Discussion

The results from the association measures (PMI and SCP) were quite similar, with both identifying about 15 valid multi-word units in the top 20. Though the filter method presented better results, it relies on a part of speech tagger, which may not be accurate for the out-of-vocabulary words common in the medical domain. Depending on the purpose of the MWU extraction task, different methods may be preferred. These methods are also valuable to generate a high quality list of MWUs for manual identification.

6.2 Side Effect Term Extraction

Related to the task of MWU identification is term extraction. We are especially interested in identifying side effect terms. While previous medical NLP research often relies on medical lexica such as those provided by UMLS or the FDA's COSTART corpus, we chose not to use these restrictive lexica because they have low coverage of colloquial side effect expressions.

We extracted side effects from the comments posted to Askapatient.com, which contains over 100,000 drug reviews, covering all drugs, and has labeled side effect data. Patients are able to submit drug reviews with an input for "side effects" where they could enter comments specifically related to side effects. Not all users used that area; some users entered free text. However, many users entered comma separated side effect terms, such as the comment below:

> Body aches, joint pain, decreased mobility, decreased testosterone and libido, difficulty getting out of bed in the morning, tingling and itchy hands,and decrease in overall strength.

Side effects were selected using regular expression (regex) string matching heuristics, including searching for comma-separated values. Qualifying terms such as *slight*, *overwhelmingly* and *extremely* were removed[1], and plural terms were consolidated.

[1]The entire list can be found in AppendixD

Terms that appeared at least 20 times were included. For a rough idea of the disparity, of the nearly 5,600 adverse effect terms found in the COSTART corpus[2], only 176 are shared with the 1,057 side effect terms we identified from the online drug reviews.

Some of the most common side effects are shown in Table 6.8. The terms in bold are not found in the COSTART corpus, and most are valid side effect terms. As we go further down the list, we see even less coverage of colloquial terms.

Rank	Side Effect	Count	Rank	Side Effect	Count
1	**weight gain**	5762	21	irritability	1248
2	headache	5689	22	weight loss	1219
3	nausea	5621	23	drowsiness	1129
4	**none**	4713	24	**night sweats**	1055
5	fatigue	4628	25	**memory loss**	995
6	depression	4562	26	acne	984
7	insomnia	3750	27	sleepiness	962
8	dizziness	3691	28	vomiting	899
9	anxiety	3592	29	confusion	884
10	dry mouth	3006	30	blurred vision	865
11	mood swings	2660	31	**feet**	860
12	constipation	2024	32	**no side effects**	849
13	**loss of appetite**	1795	33	**itching**	840
14	**tired**	1698	34	**moodiness**	820
15	bloating	1592	35	vivid dreams	807
16	hair loss	1525	36	sweating	779
17	tiredness	1361	37	lethargy	749
18	joint pain	1348	38	**dizzy**	726
19	hot flashes	1341	39	stomach pain	713
20	diarrhea	1308	40	weakness	671

Table 6.8: Side effects extracted from the Askapatient corpus. Bolded terms are not found in the COSTART corpus of adverse reaction terms.

6.3 Review Classification

Unsupervised document classification is an important task previously applied to a wide range of text such as technical abstracts, news stories, and spam e-mails. We

[2]http://hedwig.mgh.harvard.edu/biostatistics/files/costart.html

perform the classification task on the cholesterol-lowering drug reviews, classifying reviews as either a statin review or non-statin review. As each drug class has different tendencies for specific side effects, we can train a document classification model to classify an unlabeled drug review into a specific drug class using these terms as learning features. Our findings both validate the utility of the side effects for identifying the drug class and offer a useful technique for automatic assignment of unlabeled reviews. These experiments were conducted jointly with JingJing Liu, a fellow graduate student.

6.3.1 Methods

We use a Support Vector Machine [39] classifier to classify comments based on the drug class. We compared 7,971 reviews on statin drugs with 3,549 reviews on non-statin drugs using ten-fold cross validation. As a baseline, we use a classification model trained on all the unigrams in the drug reviews. We compare this with a system that uses as features the words and phrases that are skewed in distribution between the two datasets, according to the log likelihood statistic. Given the list of terms ranked by log likelihood, we filtered out terms with p-value higher than 0.05 (equivalently, log likelihood lower than 3.85). 1,991 terms selected using this threshold cutoff were used to train the LLR classification model.

Obviously, the drug's name is a very strong indicator of the drug class, but has no information about side effects (e.g., a review containing the term *lipitor* is most likely to be a review related to statin drugs). Therefore, we conducted a second experiment where all drug names were removed from both the unigrams used in the baseline system and the terms used in the LLR system.

6.3.2 Results

Table 6.9 presents the experimental results on classification. BS represents the baseline system using all the unigrams in the reviews for model training. LLR represents our classification model trained on the 1,991 terms selected by the log likelihood

Feature Set	Accuracy	Precision	Recall	F-score
BS	84.4%	82.9%	97.8%	89.7
LLR	87.1%	86.3%	96.8%	91.2
BS - DN	78.4%	76.8%	98.6%	86.3
LLR - DN	80.1%	80.6%	95.6%	87.4

Table 6.9: Drug review classification performance. BS: baseline; LLR: log likelihood ratio; DN: drug names. Precision, recall, and F-score are for statin reviews.

method. BS - DN represents the baseline trained on unigrams without drug names. LLR - DN represents the LLR system trained on 1,959 terms learned by the log likelihood method with drug names removed. Experimental results show that the LLR system outperforms the baseline system in both settings (with or without drug names).

6.3.3 Discussion

As expected, without the drug name features, performance drops in both systems. However, even without drug names, the LLR system can still achieve over 80% precision on the classification task. This indicates that the drug classes can be predicted quite well based on their unique side effect profile, by exploiting the LLR-derived features.

The classification experiments presented can serve as a good starting point for identifying unlabeled patient reviews. While our experiments were conducted on labeled data from drug review sites, many patient comments on health forums also contain personal anecdotes about medical drugs. We can use those comments to supplement the drug reviews for a larger data set. For this application, the classification threshold should be adjusted to achieve higher precision.

6.4 Topic Modeling

Topic models are also a useful tool for processing large collections of documents by more efficiently representing text, and aid in discovering abstract concepts in text.

Methods in Latent Semantic Analysis (LSA) and LDA, which is a generalization of probabilistic LSA developed by Blei et al. [7], and currently one of the most used topic models, represents documents as a random mixture of topics, or word distributions. LDA has been employed in the biomedical domain to characterize the change in research focus over time in a bioinformatics journal [90]. We applied LDA to the corpus of cholesterol-lowering drug reviews to discover correlated terms.

6.4.1 Methods

We used the MALLET toolkit[3] to perform topic classification with LDA on the entire corpus. Because MALLET processes only unigrams, we preprocessed the raw text data by joining (via the device of underbars) common multi-word side effect terms, found as described in Section 6.2, as we are most interested in side effect classes.

6.4.2 Results and Discussion

A total of 100 latent topics were generated using LDA. While some of the automatically generated topics appeared somewhat arbitrary, several topics could be assigned a clear label associated with a side effect class, as illustrated in Table 6.10. Perhaps the most striking topic is one we have labeled as "neurological," which included *lipitor* (a lipophilic statin) in a class with *parkinson*, *neurologist*, *twitching* and *tremors*.

LDA generated many useful classes of side effects. These can be used to as features to improve classification [6], or associated with ratable aspects to generate text summaries [80].

[3]http://mallet.cs.umass.edu/

Topic	Terms
muscle aches	pain, left, arm, shoulder, neck, elbow, upper, shoulder, pain, hand, developed, neck pain, lift, sore, feels, blade, upper back, hurts, blades, arm pain
weakness	muscle pain, weakness, fatigue, extreme, general, muscle weakness, stiffness, symptoms,tiredness, joint, severe, malaise, muscle fatigue, difficulty walking, cq, extremities, dark urine, clear, stronger
mental problems	fatigue, depression, extreme, anxiety, insomnia, memory loss, weight gain, energy, mild, tiredness, short term memory loss, shortness of breath, exhaustion, muscle aches, night sweats, lethargy, mental, experiencing, confusion
neurological disorders	lipitor, husband, diagnosed, recently, suffered, yrs, disease, parkinson's, early, connection, mentioned, neurologist, diagnosis, result, tremors, prior, suggest, possibility, twitching
indigestion	stomach, gas, terrible, constipation, bloating, chest, back, chest pain, abdominal pain, back pain, stomach pain, heartburn, acid reflux, bad, chest pains, rib, sick, abdomen, indigestion
arthritis	knees, joint pain, arthritis, joints, hand, pain, joint, hands, fingers, hips, shoulders, stiff, painful, finger, elbows
skin problems	itching, rash, skin, itchy, itch, reaction, burning, hot flashes, red, hives, hot, relief, redness, cream, broke, allergic, area, benadryl, unbearable

Table 6.10: Examples of latent classes automatically discovered using LDA

Chapter 7

Conclusions and Future Work

In this work, we have presented a new corpus of online patient-provided drug reviews and described preliminary experiments in developing a speech-enabled online interface for patients who want to learn more about side effects and experiences with pharmaceutical drugs. Using statistical methods, we demonstrate that patient-provided text can be used both to confirm known side effects and to discover new side effects of cholesterol-lowering drugs. They are also useful for extracting and grouping colloquial side effect terms.

In our study of cholesterol-lowering drugs, we used several popular statistical NLP techniques to detect biases in word distributions when comparing reviews of statin drugs with reviews of other cholesterol-lowering drugs. We found a statistically significant association between statins and a wide range of disorders and conditions, including diabetes, depression, Parkinson's disease, memory loss, Lou Gehrig's disease, fibromyalgia and heart failure. A review of the research literature on statin side effects also corroborates our findings. These results show promise for patient drug reviews to serve as a data source for pharmacovigilance.

We also collected spoken data of questions regarding medical drugs and associated symptoms with transcriptions. Methods to improve speech recognition in the medical domain through language modeling were explored, and we obtained slight improvements using class-based trigrams and supplementary text training data.

Finally, we used statical measures and simple string matching to extract colloquial

side effect terms from the drug reviews. We found that many concepts are represented differently in patient vocabulary and medical lexica.

In the future, we plan to expand our methods to other drug classes, such as psychopharmaceuticals and acid reflux therapies. We also encountered many terms in our analysis that were biased toward one data set, but were not statistically significant. The data sparsity issue can be addressed by collecting more drug experience comments. Classification methods may also be used to identify unlabeled patient reviews to supplement the labeled comments. Future work will address some of the issues we encountered by better filtering comments for only personal experiences. Syntactic parsers can also be applied to demonstrate a clearer cause and effect relation between drugs and adverse reactions.

Ultimately, the results of these experiments will be used to help consumers decide which medicines to take, if any.

Appendix A

Hierarchy for Cholesterol Lowering Drugs

- statin
  - atorvastatin: lipitor, torvast
  - cerivastatin: baycol, lipobay
  - fluvastatin: lescol, lescol xl, canef, vastin
  - lovastatin: altocor, altoprev, mevacor
  - pravastatin: pravachol, selektine, lipostat
  - pitavastatin: livalo, pitava
  - rosuvastatin: crestor
  - simvastatin: zocor, lipex, ranzolont, simvador, velastatin
- statin combination
  - atorvastatin/amlodipine: caduet, envacar
  - ezetimibe/simvastatin: vytorin
  - niacin/lovastatin: advicor
  - niacin/simvastatin: simcor
  - pravastatin/fenofibrate
- bile acid sequestrant
  - cholestyramine: questran, questran light, prevalite

  - colesevelam: cholestagel, welchol
  - colestilan
  - colestipol: colestid
  - colextran: dexide

- fibrate
  - aluminium clofibrate
  - bezafibrate: bezalip
  - ciprofibrate: modalim, oroxadin
  - clinofibrate
  - clofibrate: atromid-s, atromid
  - clofibride
  - etofibrate: clofibrate/niacin
  - fenofibrate: tricor, trilipix, fenoglide, lipofen, lofibra, antara, fibricor, triglide
  - gemfibrozil: lopid, gemcor
  - ronifibrate
  - simfibrate

- niacin derivatives
  - niacin: nicotinic acid
    - slo-niacin
    - niaspan: niaspan er
  - acipimox: olbetam
  - nicotinamide: niacinamide, nicotinic acid amide

- cholesterol absorption inhibitor
  - ezetimibe: zetia, ezetrol

Appendix B

Anecdotes for AMT Question Collection

Below are sample anecdotes presented to workers on Amazon Mechanical Turk to collect questions that patients might ask that could be answered by these comments.

- My doctor recommended CoEnzyme Q10 after I complained about muscle pain from Simvastatin. CoEnzyme Q10 works tremendously. I started with the lowest dosage, 50mg, once per day and I haven't needed to raise the dosage. The pain was gone. Recently I needed to go off all vitamins, supplements for a medical test. Within 2 days of being off CoEnzyme 10, the pain returned. Looking forward to taking it again after the test.
- I am on a 80mg regimen of lipitor. I am experiencing severe leg cramps and my legs have lost all muscle tone and are turning into sticks. Does this sound like it is lipitor related? My doctor mentioned a CK test would this definitely show something if it is?
- I have been diagnosed with severe arthritis for over ten years and told I need a hip replacement. I knew until then I'd just have to tolerate the groin/thigh pain. Well I started taking Lipitor and after about 6 months, I was in unbearable pain, particularly both my thighs and buttocks and groin area. My doc took me off the Lipito and in two weeks, my pain was lessened 50% or more - the right side

not helped so much as that is where the “bad hip is”.

- Started taking simvastatin 40 mg and within 2 wks pain started in my neck and thighs. The pain has gotten worse in my thighs, so I am going to stop med. and see what happens. I have been this med. for 3 months.
- My aunt is 82 years old, has had heart valve surgery a few years ago and is on Zocor. she is currently hospitalized with severe pain in the upper back area. Nothing seems to helo and oain killers make her hallucinate. Does anyone think this pain could be Zocor related?
- My husband started on Lovastatin in 2006. He started to notice weakness in his right arm. This weakness progressed to the point that he saw his MD in June 2007 thinking he had a pinched nerve. After a couple of MRI’s which did not show a pinched nerve, he was referred to a neurologist who gave him a diagnosis of ”possible ALS”. In August 2007 on his 60th birthday, a second opinion confirmed the diagnosis of ALS. Since that time, my husband has progressed from weakness in his right arm to complete loss of function in his arms, very weak leg muscles and difficulty breathing. The doctors are now encouraging us to enter him into hospice care.
- I have been on 40mg Simvistatin for 3 years. The only problems have been muscle twinges in one shoulder that has failed to heal over time as most muscle twinges do. In fact, the source of pain seems to be growing or spreading, which is worrying.
- I have been experiencing a considerable amount of pain in my legs and feet as mentioned in previous posts by other people. I am on Lipitor and all of the tendons in my arms and legs seem to be inflamed. All of this came upon me slowly after starting Lipitor. I was once on Celebrex but discontinued use due to stomach bleeding episodes. I now take Mobic. I am now under the care of a “Pain Management” group.

Appendix C

Sample Questions Collected Using AMT

C.1 Cholesterol Lowering Drugs

- Are leg cramps a normal side effect of Lipitor?
- Could Lipitor be causing the numbness in my feet?
- Does Vytorin cause exhaustion?
- How long does it take to get your strength back after stopping statins?
- If I start taking Lipitor and have side effects, are there other drugs I can take?
- Is there any association between statin drug use and kidney problems?
- What are the long term effects of Lipitor?
- What other drugs can I try if I don't like Zocor?
- Will discontinuing Zocor alleviate the muscle pain?

C.2 General Medication

- How soon can I drive after taking my Ambien?
- If I have to skip a dose of Nexium, how quickly will my acid reflux return?
- Will Yasmin hurt the baby if I get pregnant?
- Will taking this medication affect the use of other meds I am taking?

- If I take prednisone for more than 2 weeks, can I stop it suddenly?
- Can Nexium cause diarrhea?
- What are the differences between Lexapro and Celexa?
- Are there particular drugs to avoid while on Ramipril?
- If I have bad kidneys, can I take Advil?

Appendix D

Qualifying Terms Excluded from Side Effects

almost	intermittent	serious
always	major	severe
complete	massive	slight
constant	mild	slightly
extreme	minor	some
extremely	occasional	still
general	overall	terrible
horrible	overwhelming	very
intense	possible	

Bibliography

[1] Lipitor drug interactions, lipitor side effects and lipitor patient reviews. `http://www.iguard.org/medication/Lipitor.html`, 2010. [Online; accessed 7-Jan-2010].

[2] James Allen, George Ferguson, Nate Blaylock, Donna Byron, Nathanael Chambers, Myroslava Dzikovska, Lucian Galescu, and Mary Swift. Chester: towards a personal medication advisor. *J. of Biomedical Informatics*, 39(5):500–513, 2006.

[3] M. Angell. Drug companies & doctors: a story of corruption. *The New York Review of Books*, 56(1):8–12, 2009.

[4] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe. Extraction of adverse drug effects from clinical records. *Studies in health technology and informatics*, 160:739, 2010.

[5] D.W. Bates, R.S. Evans, H. Murff, P.D. Stetson, L. Pizziferri, and G. Hripcsak. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115, 2003.

[6] D.M. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] J. Cable. Adverse Events of Statins - An Informal Internet-based Study. *JOIMR*, 7(1), 2009.

[9] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics, 2009.

[10] F. M. Chowdhury and A. Lavelli. Disease Mention Recognition with Specific Features. *ACL 2010*, 2010.

[11] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[12] N. Collier, S. Doan, A. Kawazoe, R.M. Goodwin, M. Conway, Y. Tateno, Q.H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, et al. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940, 2008.

[13] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 201–207. Association for Computational Linguistics, 2000.

[14] J. Cott. Omega-3 Fatty Acids and Psychiatric Disorders. *Alternative therapies in women's health*, 1:97–104.

[15] K.P. Davison, J.W. Pennebaker, and S.S. Dickerson. Who talks? The social psychology of illness support groups. *Social psychology of illness support groups. American Psychologist*, 55:205–217, 2000.

[16] L. De Graaf, A. Brouwers, and WL Diemont. Is decreased libido associated with the use of HMG-CoA-reductase inhibitors? *British journal of clinical pharmacology*, 58(3):326–328, 2004.

[17] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[18] G. Divita, T. Tse, and L. Roth. Failure analysis of MetaMap transfer (MMTx). In *Medinfo 2004: Proceedings Of THe 11th World Congress On Medical Informatics*, page 763. Ios Pr Inc, 2004.

[19] J. Dorstand, P. Kühnlein, C. Hendrich, J. Kassubek, A.D. Sperfeld, and A.C. Ludolph. Patients with elevated triglyceride and cholesterol serum levels have a prolonged survival in amyotrophic lateral sclerosis. *J Neurol*, in Press:Published online Dec. 3 2010, 2010.

[20] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[21] G. Eysenbach. SARS and population health technology. *Journal of Medical Internet Research*, 5(2), 2003.

[22] R.M. Fano and WT Wintringham. Transmission of information. *Physics Today*, 14:56, 1961.

[23] Bruce Fireman, Joan Bartlett, and Joe Selby. Can Disease Management Reduce Health Care Costs By Improving Quality? *Health Aff*, 23(6):63–75, 2004.

[24] C. Fox. A stop list for general text. In *ACM SIGIR Forum*, volume 24, pages 19–21. ACM, 1989.

[25] C.C. Freifeld, K.D. Mandl, B.Y. Reis, and J.S. Brownstein. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150, 2008.

[26] R. Gaizauskas, G. Demetriou, and K. Humphreys. Term recognition and classification in biological science journal articles. In *In Proc. of the Computional Terminology for Medical and Biological Applications Workshop of the 2 nd International Conference on NLP*. Citeseer, 2000.

[27] J. R. Glass, T. J. Hazen, and I. L. Hetherington. Real-time telephone-based speech recognition in the jupiter domain. In *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, pages 61–64, Washington, DC, USA, 1999. IEEE Computer Society.

[28] M.R. Goldstein and L. Mascitelli. Statin-induced diabetes: perhaps, it's the tip of the iceberg. *QJM*, Published online, Nov 30, 2010.

[29] B.A. Golomb, M.H. Criqui, H. White, and J.E. Dimsdale. Conceptual foundations of the UCSD Statin Study: a randomized controlled trial assessing the impact of statins on cognition, behavior, and biochemistry. *Archives of internal medicine*, 164(2):153, 2004.

[30] Q. Gu, CF Dillon, and VL Burt. Prescription drug use continues to increase: us Prescription drug data for 2007-2008. *NCHS data brief*, (42):1, 2010.

[31] J. Hagedorn and R. Arora. Association of Statins and Diabetes Mellitus. *American journal of therapeutics*, 17(2):e52, 2010.

[32] J. Hanai, P. Cao, P. Tanksale, S. Imamura, E. Koshimizu, J. Zhao, S. Kishi, M. Yamashita, P.S. Phillips, V.P. Sukhatme, et al. The muscle-specific ubiquitin ligase atrogin-1/MAFbx mediates statin-induced muscle toxicity. *Journal of Clinical Investigation*, 117(12):3940–3951, 2007.

[33] J. Hanson, A. Gille, S. Zwykiel, M. Lukasova, B.E. Clausen, K. Ahmed, S. Tunaru, A. Wirth, and S. Offermanns. Nicotinic acid–and monomethyl fumarate–induced flushing involves GPR109A expressed by keratinocytes and COX-2–dependent prostanoid formation in mice. *The Journal of clinical investigation*, 120(8):2910, 2010.

[34] K. Hedenmalm, G. Alvan, P. Ohagen, and M-L Dahl. Muscle toxicity with statins. *Pharmacoepidemiology and Drug Safety*, 19:223231, 2010.

[35] L. Hirschman, A.A. Morgan, and A.S. Yeh. Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259, 2002.

[36] A. Hliaoutakis, K. Zervanou, E.G.M. Petrakis, and E.E. Milios. Automatic document indexing in large medical collections. In *Proceedings of the international workshop on Healthcare information and knowledge management*, pages 1–8. ACM, 2006.

[37] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[38] A.K. Jha, G.J. Kuperman, J.M. Teich, L. Leape, B. Shea, E. Rittenberg, E. Burdick, D.L. Seger, M.V. Vliet, and D.W. Bates. Identifying adverse drug events. *Journal of the American Medical Informatics Association*, 5(3):305, 1998.

[39] T. Joachims. Making large scale SVM learning practical. 1999.

[40] J.S. Justeson and S.M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01):9–27, 1995.

[41] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8. Association for Computational Linguistics, 2002.

[42] Andreas Kellner. Initial Language Models for Spoken Dialogue Systems. volume 1, pages 185–188, 1998.

[43] S. Kogan, Q. Zeng, N. Ash, and RA Greenes. Problems and challenges in patient information retrieval: a descriptive study. In *Proceedings of the AMIA Symposium*, page 329. American Medical Informatics Association, 2001.

[44] G. Kolata and N. Singer. Good news and bad from a heart study. November 15 2010.

[45] M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1), 2010.

[46] Ronilda C. Lacson, Regina Barzilay, and William J. Long. Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization. pages 541–555, 2006.

[47] P.H. Langsjoen and A.M. Langsjoen. The clinical use of HMG CoA-reductase inhibitors and the associated depletion of coenzyme Q _ {10}. A review of animal and human publications. *Biofactors*, 18(1):101–111, 2003.

[48] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. *ACL 2010*, page 117, 2010.

[49] Jingjing Liu and Stephanie Seneff. Review sentiment scoring via a parse-and-paraphrase paradigm. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 161–169, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[50] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson. RxNorm: prescription for electronic drug information exchange. *IT professional*, pages 17–23, 2005.

[51] M. Marge, S. Banerjee, and A.I. Rudnicky. Using the Amazon Mechanical Turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE, 2010.

[52] A.T. McCray, R.F. Loane, A.C. Browne, and A.K. Bangalore. Terminology issues in user access to Web-based medical information. In *Proceedings of the AMIA Symposium*, page 107. American Medical Informatics Association, 1999.

[53] C. Medawara, A. Herxheimer, A. Bell, and S. Jofre. Paroxetine, Panorama and user reporting of ADRs: Consumer intelligence matters in clinical practice and post-marketing drug surveillance. *The International Journal of Risk and Safety in Medicine*, 15(3):161–169, 2002.

[54] G.B. Melton and G. Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457, 2005.

[55] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and JF Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 3:128–144, 2008.

[56] M.G. Mohaupt, R.H. Karas, E.B. Babiychuk, V. Sanchez-Freire, K. Monastyrskaya, L. Iyer, H. Hoppeler, F. Breil, and A. Draeger. Association between statin-associated myopathy and skeletal muscle damage. *Canadian Medical Association Journal*, 181(1-2):E11, 2009.

[57] S.L. Molyneux, C.M. Florkowski, A.M. Richards, M. Lever, J.M. Young, and P.M. George. Coenzyme Q10; an adjunctive therapy for congestive heart failure? *Journal of the New Zealand Medical Association*, 122:1305, 2009.

[58] E. Mykhalovskiy and L. Weir. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. *Canadian journal of public health*, 97(1):42–44, 2006.

[59] A. Peñas, F. Verdejo, J. Gonzalo, et al. Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics*, volume 2001. Citeseer, 2001.

[60] J.F.E. Penz, A.B. Wilcox, and J.F. Hurdle. Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182, 2007.

[61] L. Plaza, M. Stevenson, and A. Dıaz. Improving Summarization of Biomedical Documents using Word Sense Disambiguation. *ACL 2010*, page 55, 2010.

[62] M. Rauchhaus, A.L. Clark, W. Doehner, C. Davos, A. Bolger, R. Sharma, A.J.S. Coats, and S.D. Anker. The relationship between cholesterol and survival in patients with chronic heart failure. *Journal of the American College of Cardiology*, 42(11):1933–1940, 2003.

[63] P. Resnik. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57. Washington:, 1997.

[64] L. Rojas-Barahona, S. Quaglini, and M. Stefanelli. HomeNL: Homecare Assistance in Natural Language. An Intelligent Conversational Agent for Hypertensive Patients Management. *Artificial Intelligence in Medicine*, pages 245–249, 2009.

[65] P. Ruch, R. Baud, and A. Geissb
"uhler. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine*, 29(1-2):169–184, 2003.

[66] Alexander Rudnicky. Language modeling with limited domain data. In *Proceeding of the 1995 ARPA Workshop on Spoken Language Technology*, pages 66–69. Morgan Kaufmann, 1995.

[67] G. Saher, B. Brügger, C. Lappe-Siefke, W. Möbius, R. Tozawa, M.C. Wehr, F. Wieland, S. Ishibashi, and K.A. Nave. High cholesterol level is essential for myelin membrane growth. *Nature neuroscience*, 8(4):468–475, 2005.

[68] Stephanie Seneff and Joseph Polifroni. Dialogue management in the mercury flight reservation system. In *ANLP/NAACL 2000 Workshop on Conversational systems*, pages 11–16, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[69] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.

[70] S. Shrivastava, T.J. Pucadyil, Y.D. Paila, S. Ganguly, and A. Chattopadhyay. Chronic Cholesterol Depletion Using Statin Impairs the Function and Dynamics of Human Serotonin1A Receptors. *Biochemistry*, 49(26):5426–5435, 2010.

[71] J.F. da Silva and G.P. Lopes. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Sixth Meeting on Mathematics of Language*, 1999.

[72] M.A. Silver, P.H. Langsjoen, S. Szabo, H. Patil, and A. Zelinger. Effect of atorvastatin on left ventricular diastolic function and ability of coenzyme Q10 to reverse that dysfunction. *The American journal of cardiology*, 94(10):1306–1310, 2004.

[73] C. Silverberg. Atorvastatin-induced polyneuropathy. *Annals of Internal Medicine*, 139(9):792, 2003.

[74] F.J. Sim, J.K. Lang, T.A. Ali, N.S. Roy, G.E. Vates, W.H. Pilcher, and S.A. Goldman. Statin treatment of adult human glial progenitors induces PPARγ-mediated oligodendrocytic differentiation. *Glia*, 56(9):954–962, 2008.

[75] R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales. The michigan corpus of academic spoken english. 2002.

[76] F. Smadja, K.R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[77] F.A. Smadja and K.R. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 252–259. Association for Computational Linguistics, 1990.

[78] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[79] R.D. Stanworth, K.S. Channer, D. Kapoor, and T.H. Jones. Statin therapy is associated with lower total but not bioavailable or free testosterone in men with type 2 diabetes. *Diabetes Care*, 32:541–546, 2009.

[80] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, pages 111–120. ACM, 2008.

[81] J. Tong, P.P. Borbat, J.H. Freed, and Y.K. Shin. A scissors mechanism for stimulation of SNARE-mediated lipid mixing by cholesterol. *Proceedings of the National Academy of Sciences*, 106(13):5141, 2009.

[82] P.D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[83] C.S. van der Hooft, M.C.J.M. Sturkenboom, K. van Grootheest, H.J. Kingma, and B.H.C. Stricker. Adverse drug reaction-related hospitalisations: a nationwide study in The Netherlands. *Drug Safety*, 29(2):161–168, 2006.

[84] L.R. Wagstaff, M.W. Mitton, B.M. ARVIK, and P.M. Doraiswamy. Statin-associated memory loss: analysis of 60 case reports and review of the literature. *Pharmacotherapy*, 23(7):871–880, 2003.

[85] G. Wainwright, L. Mascitelli, and M.R. Goldstein. Cholesterol-lowering therapy and cell membranes. stable plaque at the expense of unstable membranes? *Arch Med Sci*, 5:3, 2009.

[86] K.A. Weant and K.M. Smith. The Role of Coenzyme Q10 in Heart Failure (September). *The Annals of pharmacotherapy*, 2005.

[87] E. Wehrli. Parsing and collocations. *Natural Language ProcessingNLP 2000*, pages 272–282, 2000.

[88] Fuliang Weng, Andreas Stolcke, and Ananth Sankar. Hub4 language modeling using domain interpolation and data clustering. In *in Proceedings of the DARPA Speech Recognition Workshop*, pages 147–151, 1997.

[89] C.M. White. A review of the pharmacologic and pharmacokinetic aspects of rosuvastatin. *The Journal of Clinical Pharmacology*, 42(9):963, 2002.

[90] H. Wu, M. Wang, J. Feng, and Y. Pei. Research Topic Evolution in Bioinformatics. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pages 1–4. IEEE, 2010.

[91] C.C. Yang, J.W.K. Luk, S.K. Yung, and J. Yen. Combination and boundary detection approaches on Chinese indexing. *Journal of the American Society for Information Science*, 51(4):340–351, 2000.

[92] Q. Zeng, S. Kogan, N. Ash, and R.A. Greenes. Patient and clinician vocabulary: How different are they? *Studies in health technology and informatics*, pages 399–403, 2001.

[93] Q. Zeng, S. Kogan, N. Ash, RA Greenes, and AA Boxwala. Characteristics of consumer terminology for health information retrieval. *Methods of information in medicine*, 41(4):289–298, 2002.

[94] P.E. Ziajka and T. Wehmeier. Peripheral neuropathy and lipid-lowering therapy. *Southern medical journal*, 91(7):667, 1998.

[95] Victor Zue, James Glass, Michael Phillips, and Stephanie Seneff. The mit summit speech recognition system: a progress report. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 179–189, Morristown, NJ, USA, 1989. Association for Computational Linguistics.