USE OF TEXT DATA IN IDENTIFYING AND PRIORITIZING POTENTIAL DRUG REPOSITIONING CANDIDATES

by

Majid Rastegar-Mojarad

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

May 2019

ABSTRACT

USE OF TEXT DATA IN IDENTIFYING AND PRIORITIZING POTENTIAL DRUG REPOSITIONING CANDIDATES

by
Majid Rastegar-Mojarad

The University of Wisconsin-Milwaukee, 2019
Under the Supervision of Professor Susan McRoy

New drug development costs between 500 million and 2 billion dollars and takes 10-15 years, with a success rate of less than 10%. Drug repurposing (defined as discovering new indications for existing drugs) could play a significant role in drug development, especially considering the declining success rates of developing novel drugs. In the period 2007-2009, drug repurposing led to the launching of 30-40% of new drugs. Typically, new indications for existing medications are identified by accident. However, new technologies and a large number of available resources enable the development of systematic approaches to identify and validate drug-repurposing candidates with significantly lower cost. A variety of resources have been utilized to identify novel drug repurposing candidates such as biomedical literature, clinical notes, and genetic data. In this dissertation, we focused on using text data in identifying and prioritizing drug repositioning candidates and conducted five studies.

In the first study, we aimed to assess the feasibility of using patient reviews from social media to identify potential candidates for drug repurposing. We retrieved patient reviews of 180 medications from an online forum, WebMD. Using dictionary-based and machine learning approaches, we identified disease names in the reviews. Several publicly available resources were used to exclude comments containing known indications and adverse drug effects. After manually reviewing some of the remaining comments, we implemented a rule-based system to identify

beneficial effects. The dictionary-based system and machine learning system identified 2178 and 6171 disease names respectively in 64,616 patient comments. We provided a list of 10 common patterns that patients used to report any beneficial effects or uses of medication. After manually reviewing the comments tagged by our rule-based system, we identified five potential drug repurposing candidates. To our knowledge, this was the first study to consider using social media data to identify drug-repurposing candidates. We found that even a rule-based system, with a limited number of rules, could identify beneficial effect mentions in the comments of patients. Our preliminary study shows that social media has the potential to be used in drug repurposing.

In the second study, we investigated the significance of extracting information from multiple sentences specifically in the context of drug-disease relation discovery. We used multiple resources such as Semantic Medline, a literature-based resource, and Medline search (for filtering spurious results) and inferred 8,772 potential drug-disease pairs. Our analysis revealed that 6,450 (73.5%) of the 8,772 potential drug-disease relations did not occur in a single sentence. Moreover, only 537 of the drug-disease pairs matched the curated gold standard in the Comparative Toxicogenomics Database (CTD), a trusted resource for drug-disease relations. Among the 537, nearly 75% (407) of the drug-disease pairs occur in multiple sentences. Our analysis revealed that the drug-disease pairs inferred from Semantic Medline or retrieved from CTD could be extracted from multiple sentences in the literature. This highlights the significance of the need for discourse-level analysis in extracting the relations from biomedical literature.

In the third and fourth study, we focused on prioritizing drug repositioning candidates extracted from biomedical literature which we refer to as Literature-Based Discovery (LBD). In the third study, we used drug-gene and gene-disease semantic predications extracted from Medline abstracts

to generate a list of potential drug-disease pairs. We further ranked the generated pairs, by assigning scores based on the predicates that qualify drug-gene and gene-disease relationships. On comparing the top-ranked drug-disease pairs against the Comparative Toxicogenomics Database, we found that a significant percentage of top-ranked pairs appeared in CTD. Co-occurrence of these high-ranked pairs in Medline abstracts is then used to improve the rankings of the inferred drug-disease relations. Finally, manual evaluation of the top-ten pairs ranked by our approach revealed that nine of them have good potential for biological significance based on expert judgment.

In the fourth study, we proposed a method, utilizing information surrounding causal findings, to prioritize discoveries generated by LBD systems. We focused on discovering drug-disease relations, which have the potential to identify drug repositioning candidates or adverse drug reactions. Our LBD system used drug-gene and gene-disease semantic predication in SemMedDB as causal findings and Swanson's ABC model to generate potential drug-disease relations. Using sentences, as a source of causal findings, our ranking method trained a binary classifier to classify generated drug-disease relations into desired classes. We trained and tested our classifier for three different purposes: a) drug repositioning b) adverse drug-event detection and c) drug-disease relation detection. The classifier obtained 0.78, 0.86, and 0.83 F-measures respectively for these tasks. The number of causal findings of each hypothesis, which were classified as positive by the classifier, is the main metric for ranking hypotheses in the proposed method. To evaluate the ranking method, we counted and compared the number of true relations in the top 100 pairs, ranked by our method and one of the previous methods. Out of 181 true relations in the test dataset, the proposed method ranked 20 of them in the top 100 relations while this number was 13 for the other method.

In the last study, we used biomedical literature and clinical trials in ranking potential drug repositioning candidates identified by Phenome-Wide Association Studies (PheWAS). Unlike previous approaches, in this study, we did not limit our method to LBD. First, we generated a list of potential drug repositioning candidates using PheWAS. We retrieved 212,851 gene-disease associations from PheWAS catalog and 14,169 gene-drug relationships from DrugBank. Following Swanson's model, we generated 52,966 potential drug repositioning candidates. Then, we developed an information retrieval system to retrieve any evidence of those candidates co-occurring in the biomedical literature and clinical trials. We identified nearly 14,800 drug-disease pairs with some evidence of support. In addition, we identified more than 38,000 novel candidates for re-purposing, encompassing hundreds of different disease states and over 1,000 individual medications. We anticipate that these results will be highly useful for hypothesis generation in the field of drug repurposing.

To
My Parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Susan McRoy for her continuous support and help during my Ph.D. years. I would like to express my deepest gratitude to my mentor and advisor Prof. Hongfang Liu from Mayo Clinic for her a long-time support, mentorship, and help during developing my research questions, conducting studies and publishing scientific papers. Beside my advisors, I would like to thank my dissertation committee Dr. Rohit J. Kate and Dr. Jake Luo for their support, feedback, and for their participation in my committee.

I also want to thank my colleagues at Mayo Clinic and Marshfield Clinic for their support, feedback, help in advancing my study. First, my sincere thanks goes to Dr. Ravikumar Komandur Elayavilli, my colleague at Mayo Clinic, who helped and guided me a lot during preparing and publishing my research studies.

I would like to thank my colleagues and co-authors, Dr. Simon Lin, Dr. Scott J Hebbring, Dr. Zhan Ye from Marshfield Clinic, Dr. Jill M Kolesar from University of Wisconsin-Madison, Dr. Yanshan Wang, Dingcheng Li, Dr. Saeed Mehrabi, Dr. Liwei Wang, Dr. Sunghwan Sohn, Dr. Feichen Shen, and Dr. Sijia Liu from Mayo Clinic. Without their precious support it would not be possible to conduct these studies. Special thanks to Dr. Rashmi Prasad, Dr. Timothy Patrick, and Dr. Priya Nambisan at UW-Milwaukee for providing their support and guidance at the beginning of my Ph.D.

Last but certainly not least, I would like to thank my family: my parents, my brothers and sisters, and my friends for their love and support.

Chapter 1: Introduction

1.1 Drug repositioning

Drug development is an expensive and time-consuming process. Developing a new drug costs $500 million to $2 billion and takes 10-15 years [1]. In average, developing an oncology drug takes 13 years and could cost 1.8 billion [2]. The enormous cost and lengthy process of drug development underline the need for an alternative approach. A well-known approach to reducing the cost and time of new drug development is drug repositioning [3] which is also known as drug repurposing, re-profiling, re-tasking or therapeutic switching [4], [5] [1]. Drug repositioning is the process of identifying new indications for existing or failed drugs. Figure 1 shows time-lines for drug discovery and drug repositioning process and highlights their differences [4].

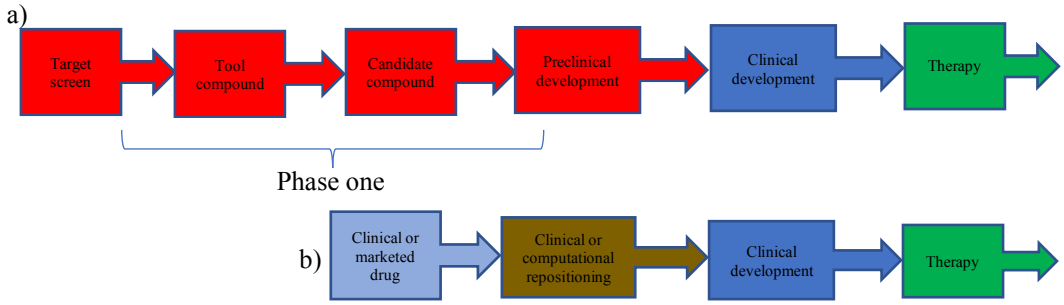

Figure 1: Comparing timelines of traditional drug discovery and drug repositioning [4] a) timeline in traditional drug discovery process b) timeline of drug repositioning

Drug repositioning could play a significant role in the drug development process. Since repositioning relies on previously approved drugs that already passed multiple toxicity tests in the first phase (Figure 1-a) of drug development process and their toxicity profiles are already known [2], drug repositioning candidates tend to be ready for clinical trials quickly and be reviewed by Food and Drug Administration (FDA) faster [5]. Drug repositioning can decrease the traditional

[1] Shameer et al. [4] provided an overview of definition for drug repositioning and related methods which repurpose pharmaceutical compounds.

timeline from 10-17 years to only 3-12 years [5]. The significance of drug repositioning would be more tangible knowing a small percentage of proposed drugs are able to pass the first phase successfully. In average less than 10% [6] of proposed drugs pass the first phase. This rate for oncology drugs is only 5% [2]. This low rate of success emphasizes the importance of drug repositioning in reducing the cost and risk of drug development. There have been many efforts on using repositioning which led to many successful discoveries.

1.2 Successful drug repositioning discoveries

One of the earliest drug repositioning cases is the use of reticulose for post-radiation effects in the 1950s [7], [8], [4]. Later in the 1980s, anti-malarial drugs were suggested as potential repurposing candidates for rheumatoid arthritis and connective tissue disease [9]–[11]. One famous example of drug repositioning is Pfizer's sildenafil (Viagra) which was repositioned from angina treatment to erectile dysfunction treatment in men in 1998 [12] and has recently been studied as a possible treatment for Age-related Macular Degeneration [13]. Thalidomide is another example of drug repositioning. Thalidomide first introduced as a sedative hypnotic [14] and then it was withdrawn because of severe side-effects. Later, Thalidomide was re-introduced as an anti-cancer agent and used for erythema nodosum leprosum and HIV wasting syndrome. In recent clinical trials, Duloxetine, a medication for depression, was found to be effective in the treatment of stress urinary incontinence in women [15]. Overall, from 2007-2009, approximately 30-40% of newly approved drugs were repurposed medications [16].

1.3 Drug repositioning approaches

Typically, a new indication for an available drug is identified only by chance. However, new data analytic technologies and a large number of available resources now enable the development of

systematic approaches to identify and assess drug repositioning candidates with considerably lower costs. Drug repositioning has been exhaustively studied, and various approaches have been used [16]–[18] to identify novel drug repositioning candidates. They have used clinical data [19], [20], genetic information [21]–[23], scientific literature [24]–[26], pathway information [27], chemical structure similarity [28], and databases of clinical side effects [29].

Several meta-analyses of drug repositioning allow us to categorize the different approaches. Dudley et. al. [30] reviewed computational methods for drug repositioning and categorized the methods into two classes: drug-based and disease-based, based on whether drug or disease perspective initiates the discovery. Wei et. al. [24] categorized the computational drug repositioning methods into literature-based and ontology-based. Grau and Serbedzija [17] named two types of drug repositioning: (1) identification of off-target drug actions and (2) identification of relevance of a known drug target to a new disease. In general, drug repositioning methods can be drug, disease or target oriented and can be further sub-classified into target-based, knowledge-based, signature-based, network-based and targeted-mechanism-based approaches [16], [31].

From an informatics perspective, we can categorize repositioning methods based on resources utilized to extract drug repositioning discoveries. These resources include scientific literature, clinical data, social media, and biological resources. The compound database PubChem [32] has been used in several drug-repurposing studies [33]. Hoehndorf et al [34] implemented a system that inferred novel associations between drugs and diseases by linking drug-gene associations in the PharmGKB database to phenotype studies and animal models of disease. Moriaud et al [35] presented a computational method that mined the Protein Data Bank [36] to identify drug repositioning candidates. Several studies [24], [25], [34], [35], [37], [38] considered literature mining for drug repurposing; this approach has been comprehensively reviewed elsewhere [39],

[40]. Another valuable resource is social media. Although patients mostly use medically oriented social media to describe adverse events associated with drugs [41], [42], their experiences may help others to conceive of new indications for existing medications if their descriptions also include beneficial effects. A well-known example is Zolpidem, an insomnia medication that, through social media and patient reviews, was subsequently used for brain injury [43]. Leaman at el [44] identified 157 beneficial effects from 3600 patient posts that could lead to drug repurposing. The accuracy of these reported beneficial effects in social media requires additional confirmation, but considering the value of drug repurposing and the huge amount of available social media data, it is worthwhile to study this type of information and investigate the possibility of identifying potential drug-repurposing candidates.

In the next chapter, we discuss related works in more detail.

1.4 Our contributions

This dissertation focuses on the use of text data in the drug repositioning process. We identify and assess challenges and opportunities in using text data in two main tasks of drug repositioning:

1) identifying potential candidates for drug repositioning (DR) and

2) validating (or prioritizing) the potential candidates.

The contributions in this dissertation consist of the following:

- Assessing the feasibility of using social media data in identifying potential drug repurposing candidates. Our hypothesis is that this imperfect resource could support drug repurposing by helping to identify unknown beneficial effects of drugs. This requires 1) gathering users' comments in social media 2) identifying mentions of medication names (including generic names,

synonyms, and brand names), instances of actual (rather than hypothetical) uses, and contrasting these uses with the original indications found in published sources such as DrugBank.) We examine two named entity recognition methods: 1) dictionary-based and 2) machine learning. We develop a rule-based system to tag sentences with potential uses for drug repositioning. To evaluate our system, we manually review the sentences which were tagged as positive to assess the possibility of the potential drug repositioning candidates to be a true finding. As there is not any gold standard for this task, we are not able to evaluate our system using more formal common metrics such as precision, recall, and F-measure.

- Comparing the number of relations that can be extracted from single sentences versus multiple sentences. Our goal is to assess the number of relations which relation extraction systems might be missing if they just focus on sentence-level analysis and to quantify the need for going beyond sentence-level extraction and implementing a relation extraction system capable of both sentence-level and discourse-level extraction. This study focusses on drug-disease relation extraction as a case study. It uses two available resources, Semantic Medline and Comparative Toxicogenomics Database (CTD), to infer drug-disease relations and then compare the coverage of these relations in sentence level extraction versus discourse level extraction. This study provides a means to compare the potential importance of considering both a) single sentences and b) multiple sentences in relation extraction systems for the task of identifying drug repositioning candidates from text data.

- Evaluating the effectiveness of two new methods to prioritize potential drug repositioning candidates identified by literature-based discovery (LBD) systems. In the first method, we consider LBD systems which use semantic predications to generate a list of potential candidates. In this study, we use predicates in semantic predications as the main metric to rank the candidates. We

assess our method by comparing it against curated databases such as CTD and expert judgment. In the second method, we train a binary classifier, using text surrounding the discoveries, to classify the discoveries into desired classes. To evaluate the performance of these classifiers, we calculate precision, recall, and F-measure.

- Evaluating the use of published biomedical literature and clinical trials in ranking potential DR candidates identified by Phenome-Wide Association Studies (PheWAS) [45]. We first generate a list of potential DR candidates using PheWAS. Then, we develop an information retrieval system to retrieve any evidence of those candidates appearing in the biomedical literature and clinical trials. We use this information to prioritize our discoveries. Like previous methods, we ask an expert to provide judgments to evaluate the performance of our ranking method.

Chapter 2: Background and Related Work

2.1 Introduction

In this chapter, we review related works focusing on identifying potential DR candidates using biomedical literature, social media and then the current studies on validation of potential DR candidates.

2.1.1 Use of biomedical literature in drug repositioning

In the past decade, advances in high throughput biotechnology have shifted biomedical research from individual genes and proteins to entire biological systems. To make sense of the large-scale data sets being generated, researchers must increasingly be able to connect with research fields outside of their core competence. In addition, researchers must interpret massive amounts of existing knowledge while keeping up with the latest developments. One way researchers cope with the rapid growth of scientific knowledge is to specialize, which leads to a fragmentation of the scientific literature. This specialization or fragmentation of literature is a growing problem in science, particularly in biomedicine. Researchers tend to correspond more within their fragments than with the field's broader community, promoting poor communication between specialties [46]. This is evidenced within the citations of such literature as authors tend to heavily cite those within their narrow specialties. Interesting and useful connections may go unnoticed for decades. This situation has created both the need and opportunity for developing sophisticated computer-supported methodologies to complement classical information processing techniques such as information retrieval [26].

Text mining can help bridge the gaps among subdisciplines by extracting new findings from the literature and presenting them to researchers based on their research interest, without requiring them to spend hours manually reading through new publications. There are systems, which go

beyond that and after extracting findings from the literature, use the findings to suggest and assess new hypotheses [47]. These systems support a process called Literature-Based Discovery (LBD) [26]. LBD strives to find novel connections or correlations between concepts by using scientific literature. LBD systems usually are comprised of the following steps:

1) Extracting findings (especially causal findings) from the scientific literature
2) Generating new hypotheses
3) Prioritizing and ranking the hypotheses

The first step usually carries out information extraction using natural language processing and machine-learning techniques. To generate new hypotheses, Swanson's ABC model [48] is a common and well-known approach. The ABC model, which we explain later, cross-references findings extracted from different articles and hypothesizes new findings. Validating these hypotheses, manually by experts, is expensive and time-consuming which makes the role of the third step within LBD systems extremely significant.

Similar to other domains, in the biomedical domain, a large pool of scientific literature is available that has been utilized by researchers to design and implement a variety of LBD systems which has led to valuable new discoveries [49]. In 1986, Don Swanson [50] introduced LBD and hypothesized that fish oil may have beneficial effects in patients with Raynaud's syndrome based on the following findings that he concluded from two separate groups of literature:

1) ***Raynaud's syndrome (A)*** patients have ***blood viscosity (B)*** disorder
2) ***Fish oil (C)*** can reduce ***blood viscosity (B)***.

Later the hypothesis was verified by clinical trials. This model is now known as Swanson ABC model. Swanson implemented the model in a software system, called ArrowSmith [51], to automate the hypothesis generation process. As the above example illustrates, the model contains

three entities/concepts, *A* as starting, *B* as linking, and *C* as target concept [52]. The discovery process starts with mining associations between *A* - *B* and *B* - *C* from literature, then by combining the associations with the same linking concept, a list of potential *A*-*C* associations is created. If there is not any prior mention about a relation between *A* and *C* in literature, then a hypothesis of association between *A* and *C* can be formulated which can be confirmed or rejected through human judgment, laboratory methods or clinical investigations. The *ABC* model can be implemented in two discovery manners, open and closed. In closed-discovery, the starting and target concepts are determined by the user and the LBD system only identifies linking concepts. In the open-discovery approach, the starting and target entities are not limited to any specific concept. This approach usually generates a long list of potential relations. As mentioned, to reject or confirm each of these relations, each candidate should be evaluated through clinical trials and laboratory experiments, which are expensive and time consuming. LBD systems can facilitate and accelerate this step by providing initial validation, using computational methods and existing knowledge, and narrowing down the list to a more reasonable number of candidates. Otherwise, it is not cost effective for other researchers or scientists to investigate the numerous numbers of candidates provided by open-discovery methods.

There have been several LBDs implemented to generate hypotheses by following Swanson's paradigm, which has led to interesting and useful discoveries [53]. One of the most desirable findings in the biomedical domain is drug-disease relation which could shed light on side effects or new indications for an existing drug [54]. The latter one is known as drug repositioning which we introduced previously and has been receiving much attention from pharmaceutical companies and researchers. New uses connecting indomethacin and Alzheimer's disease [55], somatomedin c and arginine [56], anandamide and gastric cancer [57], psychiatric and somatic diseases [58],

hypogonadism and diminished sleep quality [59], NF-kappaB and autism [60] are some of these findings. LBD can be utilized to discover any type of relation such as disease candidate genes [61], drug-drug interaction [62], drug mechanism [63], adverse drug reaction [64].

All LBD systems mentioned above have utilized a similar paradigm, but their approaches for identifying starting, linking, and target concepts in text and extracting relationships between them vary. To identify desired concepts in literature, usually LBDs rely on named entity recognition systems [65] or medical subject heading (MeSH), terms assigned by experts to Medline abstracts [66] that describe the content of the associated abstract. BITOLA [61] is one of the LBD systems that uses assigned MeSH terms to each Medline abstract as concepts. After identifying concepts, LBD can be used to extract the relationship between the concepts (causal associations). Some common methods are: co-occurrence [61], association rules [67], term frequency–inverse document frequency [68], Z-Score [69], and mutual information [70]. Other approaches are available to identify associations between concepts and terms that do not co-occur with one another in the biomedical literature [71][72]–[74]. Yetisgen-Yildiz and Pratt [52] survey these approaches and Andronis et al. [54] reviews literature mining systems which have been applied to identifying potential drug repurposing candidates.

Another approach for generating causal associations is using semantic predication [75]. A semantic predication is a "*subject – predicate – object"* triple extracted from the literature. For example, in the domain of drug repositioning, *Subject* and *object* are biomedical concepts from the UMLS Metathesaurus [76] and *predicate* is a relation from the UMLS Semantic Network such as: *affects*, *causes*, *associated with*, *treats*, etc. Semantic predications for the biomedical domain have been extracted by a rule-based system called SemRep [77] and stored in a relational database called, SemMedDB [78]. The latest version of this database contains more than 84 million predications,

extracted from 25 million PubMed abstracts. This resource has been used in several studies to facilitate knowledge discovery [79]–[81]. Hristovski et al. [75] proposed a method to use semantic predication to enhance their LBD system, BITOLA. They concluded that using semantic predications instead of co-occurrences generates a smaller number of false positive discoveries and provides an explanation to support the findings. Ahlers et al [63] used semantic predications in a closed-LBD system to discover connections between antipsychotic agents and cancer. Cohen et al. [71] proposed using hyper dimensional computing in a LBD system based on semantic predications. They suggested a method, called predication-based semantic indexing, to build a sequence of semantic predications, which ultimately associate a drug to a disease as a novel therapy. Cameron et al [53] proposed an automatic subgraph creation method, based on semantic predications, to facilitate LBD. Workman and Stoddart [79] proposed using Semantic Medline as a source for building a decision support system for point of care. The Natural Language Processing Group at the Mayo clinic integrated semantic predications into a system, called Ask Mayo Expert (AME), to retrieve the most relevant literature to support the evidence-based clinical decision making process at point of care [80].

2.1.2 Use of Social media in Drug Repositioning

Social media provides a platform for patients to share their experiences with illnesses, medications, and also medical centers [82]. Patient posts, usually written in an informal language, contain hidden and valuable information. Owing to the massive amount of data derived from social media, computerized systems are needed to analyze and extract useful information from patient experience. Unlike scientific literature, these comments are usually written by non-expert users who do not have any obligation to follow proper grammar in their comments or report observations accurately. These differences make mining social media more complicated and challenging

compared to scientific literature. Nevertheless, there have been several attempts to extract knowledge from social media. Leaman et al [44] examined comments posted in a medical forum to identify reported adverse drug events. After manually annotating a corpus of patient posts, they used natural language processing methods to develop a system that extracted adverse drug reactions from the text. Chee et al [83] studied patient posts on Health and Wellness Yahoo! groups and applied common natural language processing methods to predict adverse drug events and identify medications that might require further scrutiny by the Food and Drug Administration. Freifeld et al [84] evaluated the correlation between adverse drug events reported in Twitter (where statements are limited to 140 characters) and spontaneous reports received by a regulatory agency. Rastegar et al [85] implemented a binary classifier to identify adverse drug reactions in tweets. Sharif et al [86] proposed a sentiment classification framework to detect adverse drug reactions in medical blogs and forums. Recently, Karimi et al [87] provided a corpus of 1321 medical forum posts on patient-reported adverse drug events, which allows researchers to develop and evaluate pharmacovigilance systems.

Although patients mostly use medically oriented social media to describe adverse events associated with drugs [41], [42] , their experiences may help others to conceive of new indications for existing medications if their descriptions also include beneficial effects. A well-known example is Zolpidem, an insomnia medication that, through social media and patient reviews, was subsequently used for brain injury [43]. Leaman at el [44] identified 157 beneficial effects, in 3600 patient posts that could lead to drug repurposing. The accuracy of these reported beneficial effects in social media may be questionable, but considering the value of drug repurposing and the huge amount of available social media data, it is worthwhile to study this type of information and investigate the possibility of identifying potential drug-repurposing candidates [88].

2.1.3 Validation and prioritizing potential drug repositioning candidates

As mentioned in the description of LBD above, one challenging and expensive task after identifying potential drug repurposing candidates is validation. The discoveries can be confirmed or rejected through human judgment, laboratory methods, or clinical investigations. The validation could be facilitated with automated ranking and prioritizing of the potential candidates. There have been several studies which proposed ranking algorithms, mostly as part of LBD systems. However there have been several attempts to propose effective ranking methods [69], [70], [89] but this area has been largely unexplored. Wren proposed an algorithm called average minimum weight (AMV) [70]. The algorithm calculates a weight for each potential discovery (A-C) based on the strength of A-B and B-C relations. The strength of each relation is calculated based on mutual information. The algorithm considers all possible B concepts that have a relation with A and C in the calculation. Another approach to rank the findings is proposed by Yetisgen-Yildiz and Pratt [69], [90]. They used the number of B concepts that link A to C as the indication of a strong correlation. The method, which called Linking Term Count with Average Minimum Weight (LTC-AMW), uses AMV in case of a tie. Swanson et al. [89] introduced a measure to rank the discoveries based on MeSH terms in literature called Literature Cohesiveness. AMV and LTC-AMV are generic and can be used in different LBD systems, but these algorithms do not consider semantic predicates in their calculation.

Chapter 3: Use of social media in drug repurposing

3.1 Introduction

Social media provides a platform for patients to share their experiences with illnesses, medications, and medical centers [82], [91]. Patient posts, usually written in an informal language, contain hidden and valuable information. Owing to the massive amount of data derivable from social media, computerized systems are needed to analyze and extract useful information from the patient's perspective. Unlike edited scientific literature, these comments are usually written by non-expert users who may not follow proper grammar in their comments or report complete or accurate observations. These differences make mining social media more challenging compared to scientific literature. Nevertheless, there have been several successful attempts to extract knowledge from social media. Leaman et al [44] examined comments posted in a medical forum to identify reported adverse drug events. After manually annotating a corpus of patient posts, they used natural language processing methods to develop a system that extracted adverse drug reactions from the text. Chee et al [83] studied patient posts on Health and Wellness Yahoo! groups and applied common natural language processing methods to predict adverse drug events and identify medications that might require further scrutiny by the Food and Drug Administration. Freifeld et al [84] evaluated the correlation between adverse drug events reported in Twitter (where statements are limited to 140 characters) and spontaneous reports received by a regulatory agency. Rastegar et al [85] implemented a binary classifier to identify adverse drug reactions in tweets. Sharif et al [86] proposed a sentiment classification framework to detect adverse drug reactions in medical blogs and forums. Recently, Karimi et al [87] provided a corpus of 1321 medical forum posts on patient-reported adverse drug events, which allows researchers to develop and evaluate pharmacovigilance systems.

Although patients have mostly used medically oriented social media to describe adverse events associated with drugs [41], [42], patients also sometimes report beneficial effects and their experiences may also help others to conceive of new indications for existing medications. A well-known example is Zolpidem, an insomnia medication that was subsequently used for brain injury [43]. The miracle of this medication discovered in 1994, when a patient who suffered severe brain injuries and went to coma, started to take Zolpidem as a sedative but it brought him back to life. Since then, Zolpidem has been waking up several brain-injured patients from a vegetative state. Leaman at el [44] identified 157 beneficial effects, in 3600 patient posts that could lead to drug repurposing. These reported beneficial effects in social media may not always be completely accurate (and possibly vulnerable to a placebo effect [44]), but considering the potential value of drug repurposing and huge amount of available social media data, it is worthwhile to study this type of information and investigate the possibility of identifying potential drug-repurposing candidates. In this chapter, we consider the feasibility of using social media data in identifying potential drug repurposing candidates. Our hypothesis is that this imperfect resource can be used to identify candidates for drug repurposing by revealing new beneficial effects of medications taken by patients and reported in social media. We collect patient reviews of 180 medications in an online forum and then use common Named Entity Recognition approaches, we identify beneficial or side effects reported by patients for each drug. We utilize public resources to distinguish known side effects and beneficial effects from potentially new ones and then using a rule-based system separate side effects from beneficial effects. To evaluate our approach, we manually evaluate the finding.

3.2 Data Sources

Data for this study is obtained from public resources to allow for replicability. We use four public resources: WebMD [92], DrugBank [93], SIDe Effect Resource (SIDER) [94], and Unified Medical Language System (UMLS) [76]. Below are brief descriptions of these resources and their uses in this research.

WebMD is a collection of Web-based health-related services provided by a US corporation, that consistently ranks as the top US health publisher in the United States [92]. It includes a forum specifically for patients to share their experiences with medications. The comments are entered as free text, and the length of comments is not subject to a character or word count limit. WebMD [92] allows users to score three different aspects of the medication in their reviews: (1) effectiveness, (2) ease of use, and (3) satisfaction. WebMD provides some basic information about the users such as age, sex, and duration of treatment. Figure 2 shows a screenshot of a WebMD review page. The patient comments from WebMD were the main material used in this study.

DrugBank is a bioinformatics and cheminformatics resource that provides drug information, such as indication, synonyms, gene target, drug interactions, and structure. This database can be used to identify the current indications of drugs and thus to identify mentions of potential new uses as distinct from the prior ones.

SIDER, developed by Kuhn et al [94], contains information about 1430 marketed medications and 5880 side effects (140,064 drug-side effect pairs) extracted from public documents and package inserts. The developers of SIDER retrieved adverse drug reaction and disease names from UMLS to generate a dictionary of side effects. One can use SIDER to identify known side effects of drugs mentioned in the comments.

UMLS [76] integrates medical terminology and coding standards to help researchers and developers create interoperable biomedical information systems. One can use UMLS resources to create a dictionary of disease names that might potentially be treated by repurposed drugs.

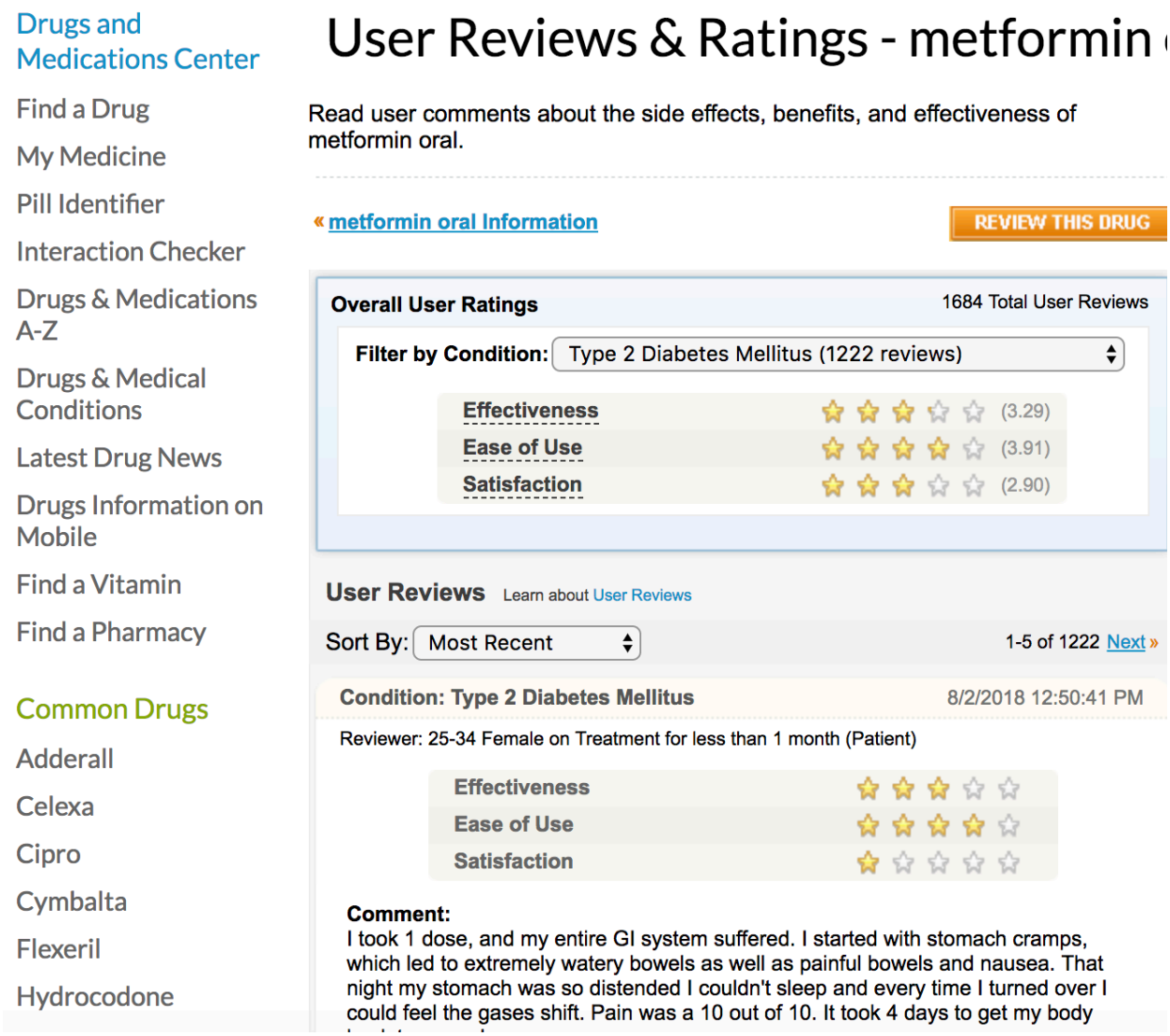

Figure 2: A screenshot of WebMD web page which allows users to post a comment about a medication and rate it based on effectiveness, ease of use, and satisfaction.

3.3 Method

3.3.1 Creating a Table of Medications

In the first step, we create a table of medications frequently mentioned in social media, the approved indications of these medications, and known side effects. For the list of medications, we generate a list of the top 180 most frequently searched medications on WebMD. Through DrugBank, we collect known and approved indications related to those medications. To locate the drugs in DrugBank, we search drug names, synonyms and brand name entries. In the next step, a list of known side effects for each drug is retrieved from SIDER. Finally, we retrieve all user-generated reviews posted in WebMD about these medications. (WebMD categorizes reviews for each medication, therefore distinguishing comments for each medication is a straightforward task and there is not any need to locate or identify drug names in the comments.)

3.3.2 Identification of Medication Effects within Consumer Reviews

We next identify beneficial or adverse effects within the posted reviews. Beneficial effects would be their impact on reducing some observed negative health condition, such as a rash. For simplicity, we refer to such negative conditions as "*diseases*" although they might be symptoms or observations related to a disease. Any mentions of disorders in the reviews are tagged by using two disease named entity recognition (NER) approaches.

In the first approach, which can be considered a dictionary-based approach, a list of disease names from UMLS is retrieved and a string-matching technique applied to identify any of the diseases mentioned in the comments. The dictionary-based approach misses mentions that are either semantic or grammatical variants of standard forms, so for these, a second approach was employed.

The second approach uses MetaMap [95]. We evaluate these two approaches by comparing disease names identified by each for the top ten most reviewed medications.

Since we are interested in newly reported uses, we discard reviews that contain only known adverse effects for related medication, using the table we created previously. We then manually review 100 of the remaining comments to develop a list of textual patterns used to report beneficial effects or indications. Using the textual patterns, we implement a rule-based system to tag the reported beneficial effects within the comments. We report the frequency of each textual patterns in the comments, before and after removing side effects and any previously known indications. In the final step and as the evaluation, the comments that contained any of the textual patterns are manually reviewed.

3.4 Results

3.4.1 Statistics of the medication table

The results of retrieving consumer generated posts for the top 180 most commonly searched drugs in WebMD (mean number of posts per drug was 358) yielded 64,616 separate posts. Within this set, Lisinopril (an angiotensin-converting enzyme inhibitor used to treat high blood pressure and heart failure) had the most comments (n=2931), whereas metoclopramide (used to treat gastric esophageal reflux disease) had the fewest comments (n=8). Table 1 shows the top 10 reviewed medications and includes the three most frequently named diseases in the respective comments. The dictionary created to capture all spelling variants of diseases provided in UMLS included 239,227 entries for 86,839 unique diseases.

Table 1: Most reviewed medications in WebMD and most frequently named diseases in reviews.

Drug name	Number of Reviews	Number of Disease names		Most frequent disease names	
		Dictionary-based	MetaMap	Dictionary-based	MetaMap
Lisinopril	2931	288	1135	Itch	Blood pressure
				High blood pressure	Cough
				Rash	Dry cough
Hydrocodone-acetaminophen	2684	320	987	Arthritis	Pain
				Itch	Back pain
				Chronic pain	Arthritis
Phentermine	1931	207	860	Dry mouth	Dry mouth
				Depression	Weight loss
				Obese	Blood pressure
Cymbalta	1651	320	1063	Depression	Depression
				Itch	Anxiety
				Fibromyalgia	Weight gain
Lexapro	1609	269	864	Depression	Depression
				Itch	Weight gain
				Panic attack	Anxiety
Effexor	1568	290	943	Depression	Depression

Drug name	Number of Reviews	Number of Disease names		Most frequent disease names	
		Dictionary-based	MetaMap	Dictionary-based	MetaMap
				Itch	Dizziness
				Panic attack	Anxiety
Tramadol	1404	261	826	Arthritis	Pain
				Fibromyalgia	Back pain
				Migraine	Dizziness
Trazodone	1305	226	701	Depression	Insomnia
				Dry mouth	Depression
				Chronic insomnia	Anxiety
Topamax	1191	271	840	Migraine	Migraine
				Gist	Headaches
				Memory loss	Tingling
Percocet	1125	245	713	Itch	Pain
				Chronic pain	Abuse
				Arthritis	Back pain

The dictionary-based NER approach identified 2178 disease names in the comments, whereas MetaMap identified 6171 disease mentions. Table 2 shows the 10 most commonly named diseases in the comments (after disambiguated terms and variations of diseases were removed manually).

Table 2: Most frequently named diseases in reviews.

Dictionary-based		**MetaMap**	
Disease	**Count**	**Disease**	**Count**
Depression	5602	Pain	9990
Itch	3594	Depression	4921
Migraine	1610	Blood pressure	4016
Dry mouth	1269	Weight gain	3778
Infection	1218	Dizziness	3484
Panic attack	1174	Anxiety	3323
Rash	1086	Headache	2216
Arthritis	905	Nausea	1977
Fibromyalgia	850	Relief	1671
Mood swing	730	Dry mouth	1279

Of the 180 drugs, 164 (91.1%) were listed in DrugBank but only 74 (41.1%) were listed in SIDER. After filtering comments to remove text describing known indications and adverse drug events from the list of recognized disease names, the top three most frequently named "*new*" diseases from the text that remained are shown in Table 3 (note the overlap with Table 1).

Table 3: Most-reviewed medications in WebMD and most frequently named diseases in the reviews after removing known indications and adverse drug events.

Drug name	Number of Disease names		Most frequent new disease names	
	Dictionary-based	MetaMap	Dictionary-based	MetaMap
Lisinopril	280	1124	Itch	Blood pressure
			High blood pressure	Cough
			Rash	Dry cough
Hydrocodone-acetaminophen	320	987	Arthritis	Pain
			Itch	Back pain
			Chronic pain	Arthritis
Phentermine	195	834	Depression	Weight loss
			Obese	Blood pressure
			High blood pressure	Sleeping
Cymbalta	320	1063	Depression	Depression
			Itch	Anxiety
			Fibromyalgia	Weight gain
Lexapro	269	864	Depression	Depression
			Itch	Weight gain

Drug name	Number of Disease names		Most frequent new disease names	
	Dictionary-based	MetaMap	Dictionary-based	MetaMap
			Panic attack	Anxiety
Effexor	290	943	Depression	Depression
			Itch	Dizziness
			Panic attack	Anxiety
Tramadol	200	670	Fibromyalgia	Pain
			Chronic pain	Back pain
			Migraine	Headache
Trazodone	196	609	Chronic insomnia	Depression
			Migraine	Anxiety
			Fibromyalgia	Headache
Topamax	271	840	Migraine	Migraine
			Gist	Headaches
			Memory loss	Tingling
Percocet	245	713	Itch	Pain
			Chronic pain	Abuse
			Arthritis	Back pain

3.4.2 Statistics of Textual Patterns

We identified 18 textual patterns used (or can be used) to report beneficial effects and then counted the frequency of them in the reports. The frequencies of the top ten most common are shown in Table 4.

Table 4: Textual patterns to identify drug repositioning candidates.

Pattern	Count	Example drugs and comments[2]
I use * for	307	Methadone: I use this for diabetic neuopathy. works well with very little side effects.
		Percocet: I use this for M.S. pain
		Percocet: I use this med for peripheral neuropathy pain.
I use it for	42	Cymbalta: My use of Cymbalta is two fold. I use it for depression and fibromyalgia pain.
		Spironolactone: I use it for acne. Go figure it works
		Promethazine: I use it for gastroparesis. I also use it for sleep 4 or 5 times a month
It helps with	131	Nucynta: It helps with my pain from surgery
		Percocet: it helps with my back pain, better then any drug
		Klonopin: I like this medication it helps with my anxiety.
It help with	11	OxyContin: it help with muscle spasms
		Neurontin: i had drop foot and much pain. it help with the pain along with the 3 epidurals i receiveed in my spine.

[2] Consumer comments are shown exactly as they appeared on the WebMD site.

Pattern	Count	Example drugs and comments[2]
		Cymbalta: i started this medication years ago. not only did it help my depression, it help with my auto immune, muscle and nerve pain.
I take it	1,161	Nucynta: I take it for severe headache and neck pain from arthritis, bulging disks, and bone spur in my neck (cervical spine)
		Methadone: I take it for chronic pain it helps a lot
		Flexeril: I take it for muscle spasms related to fibromyalgia.
I take it for	91	Methadone: I take it for chronic pain it helps a lot
		Methadone: I take it for degenertive disk deteration in my neck.
		Hydrocodone-acetaminophen: i take it for my scholiosis of my back
It works for	258	Methocarbamol: It works for my muscle tension, but gives me a headache.
		Diazepam: it works for my pain weal good
		Tramadol: It works for my Arthritis Pain.
It is useful for	0	…
Useful for	18	Methadone: very useful for chronic and severe pain associated with fibromyalgia/rheumatoid arthritis.
		Effexor: I have been reading the reviews of this med. I have been using it for 1.5 yrs and has been very useful for my depression.
		Ultram: this med has been very useful for my hip and back pain.
Prescribed for	319	Percocet: I was prescribed for kidney stones. definately took the pain away and very high.

Pattern	Count	Example drugs and comments[2]
		Zoloft: I feel like the antidepressant is used in conjunction with my cymbalta which I am prescribed for both depression and fibromayalgia.
		Celebrex: I was prescribed for knee pain following surgery for torn muniscus.

Table 5 shows the frequency of the patterns after removing the comments that referred to only known side effects or indications.

Table 5: Frequency of common textual patterns after removing known indications and adverse drug effects.

Pattern	Count	Example drugs and comments[3]
I use * for	171	Flector: it's not so bad. I use them for stress headaches only if I have a mild headache
		Hydroxyzine: I use this drug for itching attacks and it works fast and effective for me.
		Elavil: I use this medication for restless leg syndrom
I use it for	23	Promethazine: I use it for gastroparesis.i also use it for sleep 4 or 5 times a month
		Amitriptyline: I use it for ic
		Seroquel: I m in love with seroquel its amazing! I use it for sleep and I wake up refreshed
It helps with	72	Neurontin: it helps with numbness in my legs and arms
		Neurontin: I was diagnosed with rsd in from a fall on the ice. It helps with controlling the pain;

[3] Consumer comments are shown exactly as they appeared on the WebMD site.

Pattern	Count	Example drugs and comments[3]
		Seroquel: although it helps with my depression I have gained over 50lbs
It help with	6	Oxycontin: it help with muscle spasms
		Hydrocodone-acetaminophen: it is ok I think and it help with my back pian.
		Neurontin: I had drop foot and much pain. It help with the pain along with the 3 epidurals I receiveed in my spine.
I take it	729	Methadone: I take it for chronic pain it helps alot
		Pristiq: I take it for depression and ptsd as well as for chronic pain from failed cervical fusion.
		Zoloft: I have taken it for three years almost and when I take it my depression worsens rather in the summer when I wouldnt take it I was the happiest
I take it for	48	Percocet: I take it for pain after a shoulder surgery and it works
		Buspar: I take it for stress.
		Effexor: I take it for depression.
It works for	155	Pristiq: I do not think it works for me makes me very consipated and I think it makes the back of my legs hurt in the muscle part.
		Metformin: I take it before bed no sideeffect so for taking one month hope it works for me yes I am scared
		Flexeril: back problems healed up then came right back. overall it works for a little while.
Useful for	13	Effexor: I have been reading the reviews of this med. I have been using it for 1.5 yrs and has been very useful for my depression.

Pattern	Count	Example drugs and comments[3]
		Hydrocodone-acetaminophen: this med. is useful for short term relief of pain.
		Ultram: this med has been very useful for my hip and back pain.
Prescribed for	0	…

A manual review of the remaining comments identified five drugs with potential for repurposing (see Table 6).

Table 6: Example comments suggesting the possibility of drug repositioning.

Medication	Indication	Adverse effect	Patient comments[4]
Methadone	Dry cough, drug withdrawal syndrome, opioid type drug dependence, and pain	Amenorrhea, phlebitis, sneezing, suffering, withdrawn, hypomagnesemia, urticaria, rhinorrhea, fever, spasm, …	I use this for diabetic **neuopathy**. Works well with very little side effects.
Elavil	Depression, chronic pain, irritable bowel syndrome, sleep disorders, diabetic neuropathy, agitation and insomnia, and migraine prophylaxis	None in SIDER	elavil is an old school antidepressant that is now considered a dirty drug because of its undesired side effects. one of the unintended side effects is to relax the skeletal muscle tissue. I use elavil off label to treat my **tmj**

[4] Consumer comments are shown exactly as they appeared on the WebMD site.

Medication	Indication	Adverse effect	Patient comments[4]
Spironolactone	Low-renin hypertension, hypokalemia, and Conn syndrome	Hyperkalemia, amenorrhea, urticaria, epidermal necrolysis, anaphylaxis, fever, toxic epidermal necrolysis, lethargy, nausea, …	I use it for **acne**. go figure it works
Strattera	Attention-deficit/hyperactivity disorder, alone or in combination with behavioral treatment	None in SIDER	I was prescribed this medication for slight adhd with off label **anxiety** help.
Viibryd	Acute episodes of major depression	None in SIDER	It even helps my **migraines** somewhat (maybe it will be off label in the future for migraine prophylaxis)

3.5 Discussion

3.5.1 Comparison of MetaMap Versus a Dictionary-Based Approach

MetaMap is a sophisticated tool that uses natural language processing and machine learning methods; thus, it is not surprising that it is more accurate than the dictionary-based approach. MetaMap, to some extent, addresses some general concerns such as disambiguation, misspelling, and word normalization, but none of these is addressed in the dictionary-based approach. For example, in the phrase "*My stomach and back hurts to sit, lay down, or stand,*" the dictionary-based approach would tag "*down*" as a disease because of overlap with the "*genetic disorder down*

syndrome." As Table 2 shows, MetaMap recognized about three times the number of disease names as compared to the dictionary-based approach. The main reason for this difference is word normalization in MetaMap. The dictionary-based approach is limited by its requirement for exact matches—for example, a dictionary that contains only "*dizzy*" would not detect "*dizziness*" as a relevant word. In contrast, MetaMap uses stemming and lemmatization to normalize words. The main advantage of dictionary-based mapping over MetaMap is its speed (the dictionary-based approach is considerably faster).

3.5.2 Using Patient Comments for Drug Repurposing

The reviews commonly described general disorders such as pain, itching, and headache. This is expected because comments usually are not authored by medical experts. We observed that patients more often tend to report adverse drug events instead of beneficial effects, as some of the previous studies reported a similar trend [44]. For example, in the corpus provided by Leaman et al [44], they annotated 157 beneficial effects in 3600 posts, while they found 1260 adverse drug events. Nevertheless, some patient comments contain beneficial effects of medication, which makes social media a useful resource for drug repurposing. This imbalanced distribution makes identifying beneficial effects more difficult, however, especially for training a classifier. Our results (see Tables 4, 5, and 6) suggest that an effective approach for this task would be to recognize the textual patterns that people used to report beneficial effects (eg, "*I use [drug] for [disease]*"). For example, in a review of Viibryd, a user mentioned, "*It even helps my migraines somewhat,*" clearly noting a beneficial effect of the drug, which could be captured by our rule-based system. Similar to other computational drug repurposing approaches, these findings need to be reviewed manually by experts and then confirmed or rejected by laboratory tests or clinical trials.

3.6 Limitations

There are some known limitations to this study. Analysis of the patient comments, which are written in an informal manner, obviously needs a system that can handle spelling and grammatical errors. The current implementation does not address these issues.

Our system covered only simple textual patterns, although the examples in Tables 4 and 5 highlight the need to decode complex patterns. A simple pattern-matching system obviously is insufficient for a statement such as "*I use it for nose allergies and it does not clear up my nostrils*." A system should be able to handle negation and coreference.

Another limitation of this study was that comments originated from only one forum. Other social media sites such as Yahoo! Answers, PatientsLikeMe [96], and even Twitter have similar information, which could be studied. In addition, using only one resource for known side effects and one for indication was another limitation. In Table 3, there are several known indications and adverse drug events, which highlight this limitation.

In this study, we were not able to evaluate our system using measures such as precision or recall because of the lack of an annotated corpus.

3.7 Summary

We assessed the feasibility of using social media to identify drug-repurposing candidates. After collecting patient reviews of medications from WebMD, we compared dictionary-based and MetaMap approaches to identify disorders mentioned in the reviews. Reviews describing known indications or known adverse drug events were excluded, and the remaining reviews were searched for textual patterns commonly used to report beneficial effects. Although the most commonly reported disorders were nonspecific (eg, pain, itching, headache), we nevertheless showed that

consumer comments contain beneficial effects of medication and have the potential to be used for drug repurposing. Our textual patterns were able to capture some beneficial effects, but there is a need for a more complex and sophisticated system to identify complex mentions of beneficial effects in social media, such as those involving negation or co-reference.

Chapter 4: Prioritizing potential drug repositioning candidates extracted from biomedical literature

4.1 Introduction

In this chapter, we discuss literature-based discovery (LBD) and how to prioritize potential drug repositioning candidates generated by LBD. This chapter consists of three separate studies. In the first one, we study the effect of discourse-level and sentence-level analysis for relation extraction, which is one of the main components in LBD systems. In the second and third studies, we propose and evaluate two prioritizing algorithms to rank potential drug repositioning candidates generated by LBD.

4.1.1 Discourse-level vs Sentence-level analysis for Relation Extraction

Information extraction (IE) aims to automatically extract information from text. To understand and extract information from text, IE systems should look at the text as a whole and not just individual sentences. Researchers in the discourse domain agree that usually sentences/clauses are not understood in isolation [97]. IE researchers have studied discourse-level analysis [98] for applications such as question answering and dialogue generation. One important aspect of discourse-level analysis is identifying relation between sentences (clauses), called discourse relations, such as contrast or explanation-evidence [97].

Two main tasks in IE are named entity recognition (NER) and relation extraction. In relation extraction, IE systems identify the relation between two or more entities where the entities could span multiple sentences. Identifying relations that span multiple sentences needs discourse-level analysis to interpret context dependent aspects of meaning such as coreference resolution.

The goal of the first study in this chapter is to assess the need for discourse-level analysis for relation extraction and evaluate the amount of information that IE systems would miss if they just focus on the sentence level. To perform this experiment, firstly, Semantic Medline [78] is used to

generate a list of potential drug-disease repositioning candidates. Then, we used Medline abstracts to find evidence of these drug-disease co-occurrences, using two methods, one based on sentences and the second based on discourse-level information.

4.1.2 Prioritization of LBD generated candidates

Once the potential novel findings are discovered by LBD systems, it is necessary to eliminate the false positives, and identify only true findings (novel discoveries). Distinguishing novel discoveries from the others is not a trivial task. Typically, the LBD method consists of two steps: 1) extracting and mining relations from the text, and 2) eliminating the false positives and identifying only the true relations. As a final step, however, it is also important to have a rigorous validation of the candidate relations before we proceeding to laboratory or clinical investigations, since these are not only expensive but also time consuming. The effectiveness of a LBD system, therefore, lies in its rigorous validation. Most prior studies lack such vigorous validation, including ranking of the generated candidates generated through LBD process. Though there are a few prior attempts [63], [71] in this direction, this area has been largely underexplored.

In the second part of the study, first we propose and evaluate the effectiveness of two predicate-based ranking methods for prioritizing potential drug repositioning candidates generated by LBD. Then we propose a prioritization method based on context.

We discuss the methods and results of each study separately.

4.2 Methods for Generating Candidates from single or multiple sentences

The study contains two steps: 1) generating a list of drug-disease pairs based on LBDs and 2) using the discoveries to evaluate the drug-disease relation extraction by comparing the extractions from a single sentence to those from multiple sentences.

4.2.1 Generating drug-disease pairs

In the first step, a list of potential relations between drugs and diseases is generated following Swanson's model [48]. As discussed in Chapter 2, according to Swanson's model if one scientific study notes a correlation between concept A (Starting concept) and concept B (Linking concept), and another study mentions a correlation between concept B and concept C (Target concept), then there might be a correlation between concept A and C. In our study, drug is the starting concept and disease is the target concept, a type of gene serves as a conceptual link between the two.

We use semantic predications from SemMedDB as a source of these two types of links [75]. For example, from the following two predications:

Flecainide (Drug) *INTERACTS WITH* SCN5A (Gene)

SCN5A (Gene) *ASSOCIATED WITH* Heart Failure (Disease)

the system generates *Flecainide-Heart Failure* as a potential new drug-disease pair. Figure 3 shows the architecture of the overall LBD process.

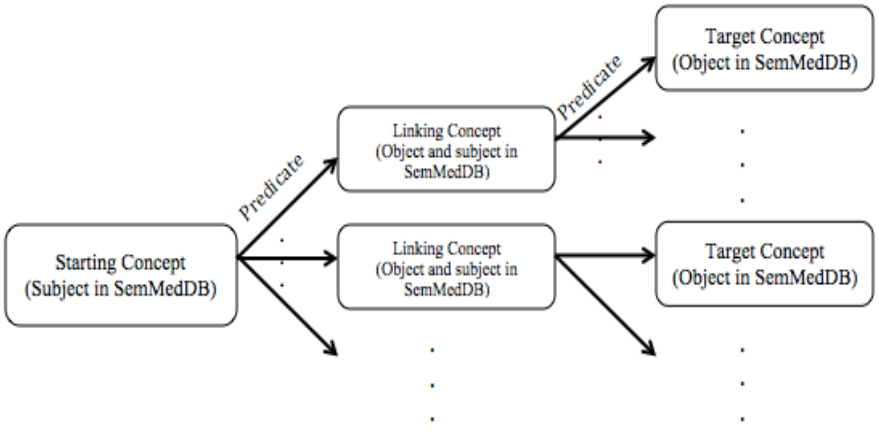

Figure 3: Architecture of our LBD system. In this system, starting concept is a drug, the linking concept is gene, and the target is a disease that leads to drug-disease discoveries. Our system uses Semantic predications as evidence of correlation between the concepts.

There are two major reasons for our choice of using semantic predication to generate LBDs

1) semantic predication takes biological meaning into consideration

2) the semantic type of the interaction and the contextual information about the interaction allow us to filter unnecessary predications such as: *NEGATIVE TREATS*, *NEGATIVE ASSOCIATED WITH*, etc.

To further refine the drug-disease pairs to be relevant for LBD, we used Timeline profiles to narrow down the drug-disease pairs. Any potential drug-disease pair that occurs before either the first occurrence of either the related drug-gene or gene-disease pair is excluded. Equation (1) clarifies this use of a Timeline profile:

(Eq 1)

where Y_{sp}(Drug-Gene) indicates the publications' date of all Medline abstracts that contain at least one semantic predication between Drug-Gene. For Drug-Disease pairs, we considered publications that contain at least one co-occurrence of the entities (Y_c). For example, assume four studies in 2002, 2004, 2007, and 2014 reported association between Drug A and Gene B; and two studies in 2007 and 2010 mentioned association between Gene B and Disease C. The left-hand side of equation would be:

Max (Min (2002, 2004, 2007, 2014), Min (2007, 2010)) = 20067

So, if Drug A and Disease C appeared together in any publication before 2007, we do not consider Drug A-Disease C. Figure 4 illustrates this example.

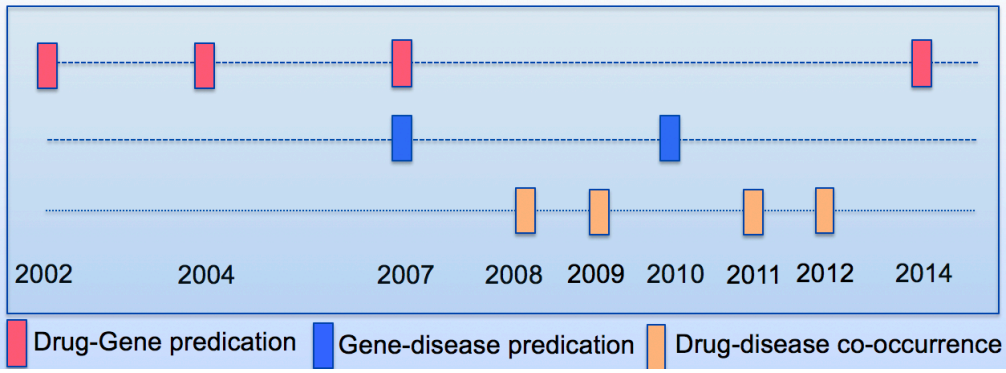

Figure 4: Timeline profile. This figure illustrates the timeline profile which we used to make the generated drug-disease pairs relevant to LBD. In this example, as drug-disease pair co-occurred in 2008 (for the first time), which is after both drug-gene and gene-disease relations mentioned in literature, is an acceptable pair in our study.

4.2.2 Evaluating sentence and discourse-level relation extraction

In the next step, we search Medline abstracts for any co-occurrence (evidence) of drug and disease pairs that were not filtered because of their place on the timeline. We categorize the drug-disease pairs with at least one evidence in Medline into two groups, depending on whether they occur in the same sentence or spanning multiple sentences

In order to assess the true validity of the drug-disease pairs, we compare the drug-disease pairs identified by the system against the chemical-disease associations in the Comparative Toxicogenomic Database (CTD) [99], a manually curated database of biological relations and associated PubMed citations. A pair is considered valid only if the Medline abstract matches a PubMed citation that contains the drug-disease pair.

Lastly, we perform an analysis to study whether discourse-level analysis would have a positive impact on the time lag since the reporting of causal pairs (drug-gene/gene-disease) and the appearance of drug-disease pair in the scientific literature.

4.2.3 Results

4.2.3.1 Retrieval of LBD relations

There were a total of 1,710 approved drugs extracted from DrugBank [93]. Using this list of drugs, from SemMedDB there were 4,096 unique drug-gene (A-B) pairs given the gene-related semantic predicates extracted from SemMedDB. For all the genes mentioned in A-B pairs we retrieved 2,741 gene-disease (B-C) relations from SemMedDB. With gene being the common link between Drug-Gene and Gene-Disease, we inferred 71,842 potential drug-disease (A-C) relations. The results of the study are illustrated in Figure 5.

4.2.3.2 Comparison of sentence level and discourse level

We found only 37,719 (52.05%) of the 71,842 drug-disease pairs with at least one literature evidence.

Timeline analysis (Eq1) further narrowed down the number of drug-disease pairs to 8,772 (23.25%) from 37,719. 6,450 drug-disease relation pairs (73.52%) out of 8,772 identified earlier transcend sentence boundaries, demanding the requirement of discourse-level analysis for textual extractions.

For the 2,322 pairs that co-occurred in at least one abstract at the sentence level, we found 89,805 total co-occurrences in the literature. Further composite analysis revealed that there was far more literature evidence across sentences than from a single sentence as shown in Figure 6.

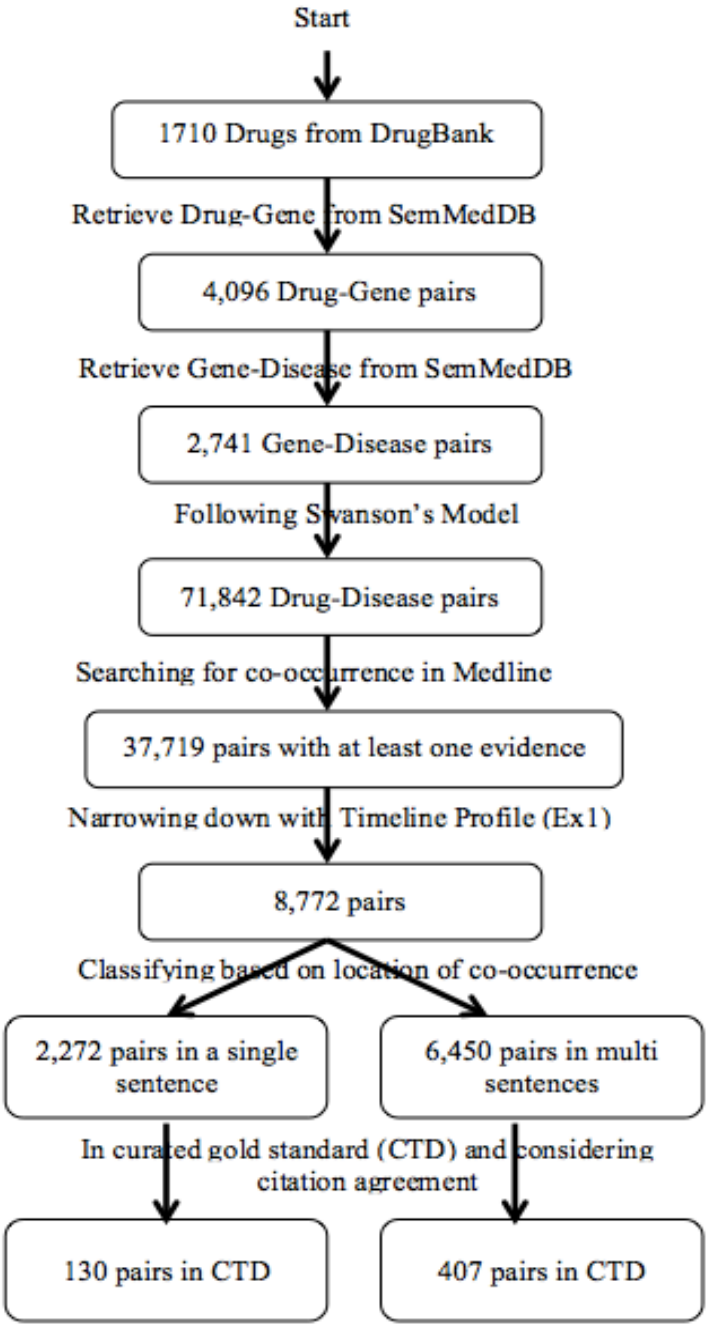

Figure 5: Results of the study.

For the 8,322 drug-disease pairs, only 537 (6.4%) of them matched the gold standard, CTD. Further analysis revealed that only 130 (24.20%) of the 537 that matched the gold standard occurred in a single sentence while the rest (75.80%) appeared across different sentences. Ignoring the match,

found in the literature citations we found 17,094 of the drug-disease pairs matched those in the CTD.

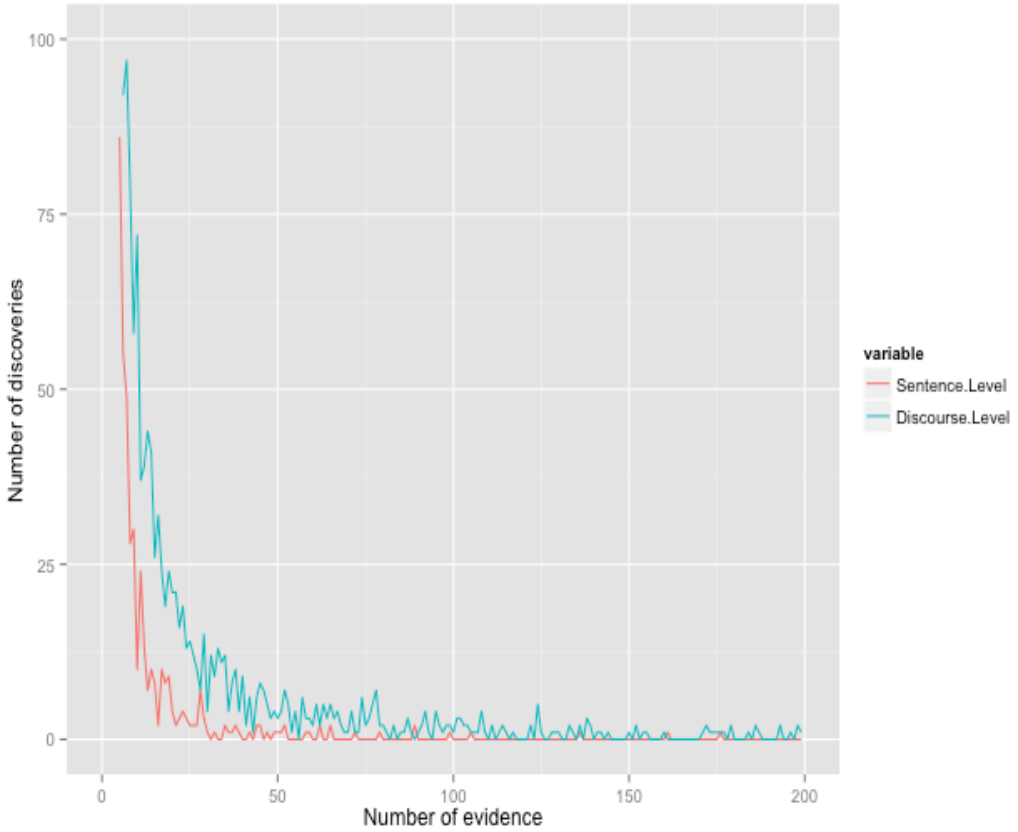

Figure 6: Comparison of frequencies of drug-disease relations' co-occurrence in a single sentence versus multiple sentences

4.2.3.3 Discourse-level analysis may impact Time lag of LBD

Figure 7 plots the cumulative percentage which represents a cluster of drug-disease pairs observed at a time zone at both the sentence level and discourse level. The trend shows that performing discourse-level analysis would significantly advance the identification of the drug candidate for a specific disease.

4.2.4 Discussion

In this study, we assessed a method for finding potential drug candidates based on co-occurrence of drug-gene pairs and gene-disease pairs within Medline Abstracts. The study also compared the candidates found when considering relations found within a single sentence versus those spanning multiple sentences and thus require some discourse-level analysis. We found that drug-gene and gene-disease relations quite often transcend clausal and sentence boundaries and hence demand mechanisms to connect information across such boundaries. Here we found the description of 23,268 potential drug-disease relations across sentence boundaries versus the 14,451 relations within sentences.

Across all experiments, we observed a consistent trend that drug-disease relations that occurred across sentences outnumbered the ones within sentences. Our evaluation against the curated CTD resource also reinforces this trend. The study also revealed that a discourse-level analysis would have significantly reduced the time lag between the scientific reporting of causal pairs (drug-gene/gene-disease) and drug-disease relational pair. This shows the need for advance discourse-level analysis approaches to extract information from literature in time and its ability to hasten the pace of discovery.

Another significant observation is the amount of potential false-positive drug-disease relations identified through literature mining. We observe a substantial reduction in the number of drug-disease relations when we compare the literature-based drug-disease pairs with those from curated resource CTD (Figure 5). The reduction among the discourse-level pairs is far greater (3.3 times) when compared to the ones from sentence level (2.5 times), which might reflect that the discourse level analysis provided more candidates.

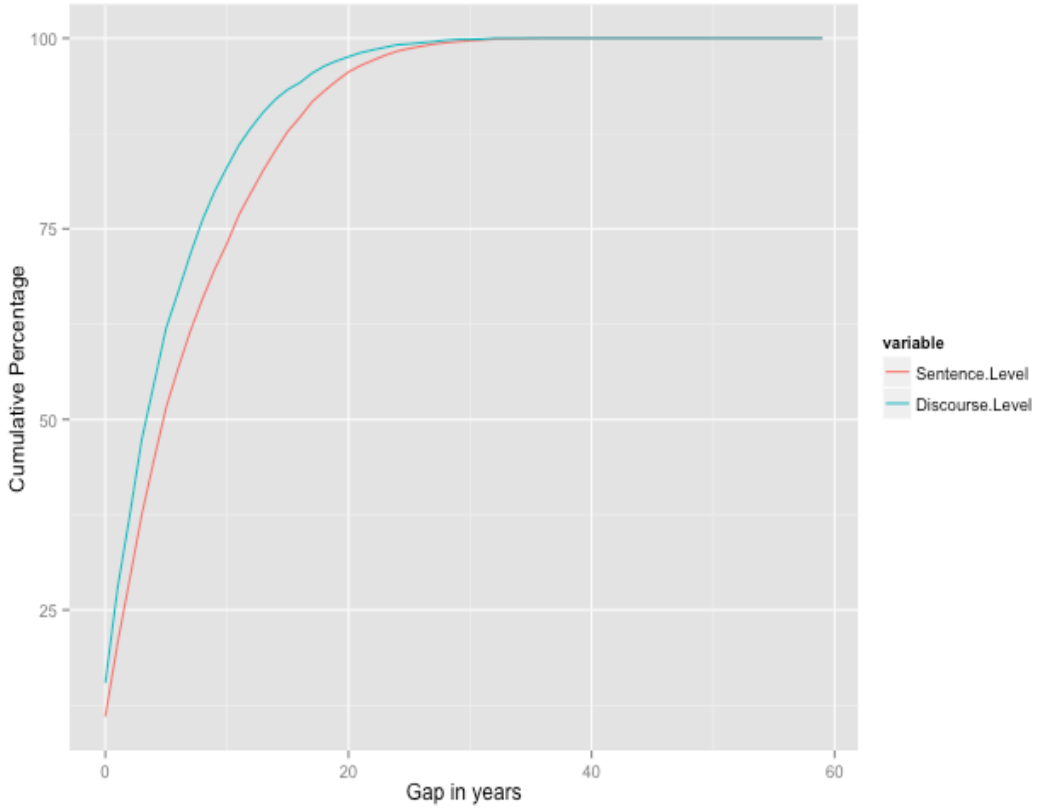

Figure 7: Comparing the time gap between the first co-occurrence of discovery and the causal pairs (Sentence level versus discourse level)

4.2.5 Summary

In this study, we investigated the extent of the need of discourse-level analysis for drug-disease relation extraction from biomedical literature. We used Semantic Medline to extract LBDs and then based on co-occurrence analysis, we collected any evidence of the discoveries in Medline abstracts. We categorized the evidence into two categories, sentence level and discourse level. From subsequent analysis, we infer that there is a potential to miss more than 70% of drug-disease relations when we extract information from just the sentence level. This clearly demonstrates the need for deeper discourse-level analysis, which may translate to significant improvement in the state of the art of NLP techniques.

4.3 Prioritization of potential discoveries generated by LBD

In the second part of this chapter, we study prioritizing drug repositioning candidates generated by LBD systems. We propose and evaluate two prioritization approaches; 1) predicate-based 2) context-based.

4.3.1 Predicate-based prioritization

In this study, we follow the same approach as in the previous study to build a LBD system which generates a list of potential drug-disease relations using semantic predications in SemMedDB. From the initial list, we explore a series of methods to eliminate erroneous extraction. We then assign a score to the remaining pairs (likelihood of being true drug-disease pair) and rank them. First, we remove pairs that were qualified by negated predicates such as "*did not inhibit*". We also remove pairs that were qualified by "*co-exists*" predicate, which does not semantically define a relationship between the pairs. For ranking, we use the occurrence of predicates that qualify the binary relationships (between drug-gene and gene-disease) as a feature to rank the final drug-disease relationships. For example, to rank this pair, *Strepsils - Chagas*, inferred from these semantic predications:

- Example 1) Strepsils (Drug), ***INTERACTS WITH***, CA2 (Gene)
- Example 2) CA2 (Gene), ***AUGMENTS,*** Chagas (Disease)

we use the predicate between drug-gene, "*INTERACTS WITH*", and the predicate between gene-disease, "*AUGMENTS*". The semantic predicates of both the drug-gene and gene-disease pairs (which we call intermediate predicates) play a determining role in qualifying a drug-disease pair. We attempt to find a meaningful co-relation between the predicates, one that qualifies drug-gene and gene-disease relationships, and the likelihood of generating a true drug-disease pair. The

importance of predicate to determine the relevance of drug-disease relations is even more important given the fact that the individual semantic predication can occur in more than one document. In order to assert a relationship that is inferred from two relationships from multiple documents, we propose that the predicate co-relation between the two relationships is one of the key factors. Besides, the relation and the predicates may have many-to-many relationships meaning that more than one predicate can qualify a relation between the two entities. For example, there is only one citation for the relationship between *Strepsils* and *CA2* (example 1), while there are six citations that contain the relationship between *CA2* and *Chagas* (example 2). Out of these six, three of them are "*ASSOCIATED WITH",* two of them are "*AUGMENTS"*, and one is "*AFFECTS"* relationship.

To assign a score to each predicate, we take advantage of the existing curated resources. There are numerous resources such as the UMLS and the Comparative Toxicogenomics Database (CTD) [99], which catalog drug-disease relationships. In this study, we use UMLS as the gold standard to evaluate the effect of intermediate predicates in generating a true drug-disease pair. To identify already known drug-disease relations in the list generated by the LBD system, we cross-reference the generated list of drug-disease pairs with UMLS drug-disease relations. We assign a score to each predicate based on how many times they generate a true drug-disease pair.

We propose two different ranking approaches based on two assumptions. In the first approach, we consider the predicate of drug-gene and gene-disease to be independent of each other, while in the second we consider the dependencies between the predicates of the two pairs.

4.3.1.1 Method of ranking based on predicate independence

Figure 8 shows the steps of calculating the independence scores for each drug-disease pair. In the first step, we use SemMedDB to retrieve drug-gene and gene-disease pairs and create a list of

potential drug-disease pairs. Then we cross-reference the drug-disease pairs with true drug-disease relations in UMLS. As our goal is to identify potential drug-repositioning candidates, we only consider drug-disease relations in UMLS where their type is "*May_be_treated_by*". As SemMedDB stores Concept Unique Identifier (CUI), assigned by UMLS to each biomedical entity, we use CUIs to cross-reference our list and UMLS. After identifying true drug-disease pairs in our list, we go back to identify drug-gene and gene-disease pairs which created those drug-disease pairs. Then we count how many times each intermediate predicate contributed to generating true drug-disease relations.

The score for a given drug-disease pair inferred from the individual pair (drug-gene (DG) and gene-disease (GD)) is calculated as per the equation 1.

$$Score = \sum_{i=1}^{n} \log\left(\frac{P_{pDG-U_i}}{P_{pDG-S_i}}\right) + \sum_{j=1}^{m} \log\left(\frac{P_{pGD-U_j}}{P_{pGD-S_j}}\right) \qquad (1)$$

Where:

n: the number of semantic predications between the drug-gene extracted from the literature

m: the same number for the gene-disease relationship

$P_{pDG\text{-}U}$[5]*:* Percentage frequency of the predicates in drug-gene which participated in creating drug-disease pairs appeared in UMLS

$P_{pGD\text{-}U:}$ Percentage frequency of the predicates in gene-disease which participated in creating drug-disease pairs appeared in UMLS

$P_{pDG\text{-}S}$[6]*:* Percentage frequency of drug-gene predicates in SemMedDB

[5] The first "P" stands for *percentage* and the second one stands for *predicate* and "DG" stands for "Drug-Gene".

[6] In this notation, "S" stands for "SemMedDB".

$P_{pGD\text{-}S}$: Percentage frequency of gene-disease predicates in SemMedDB

We use ($P_{pDG\text{-}S}$) and gene-disease ($P_{pGD\text{-}S}$) to normalize the percentage frequencies ($P_{pDG\text{-}U}$ and $P_{pDG\text{-}S}$).

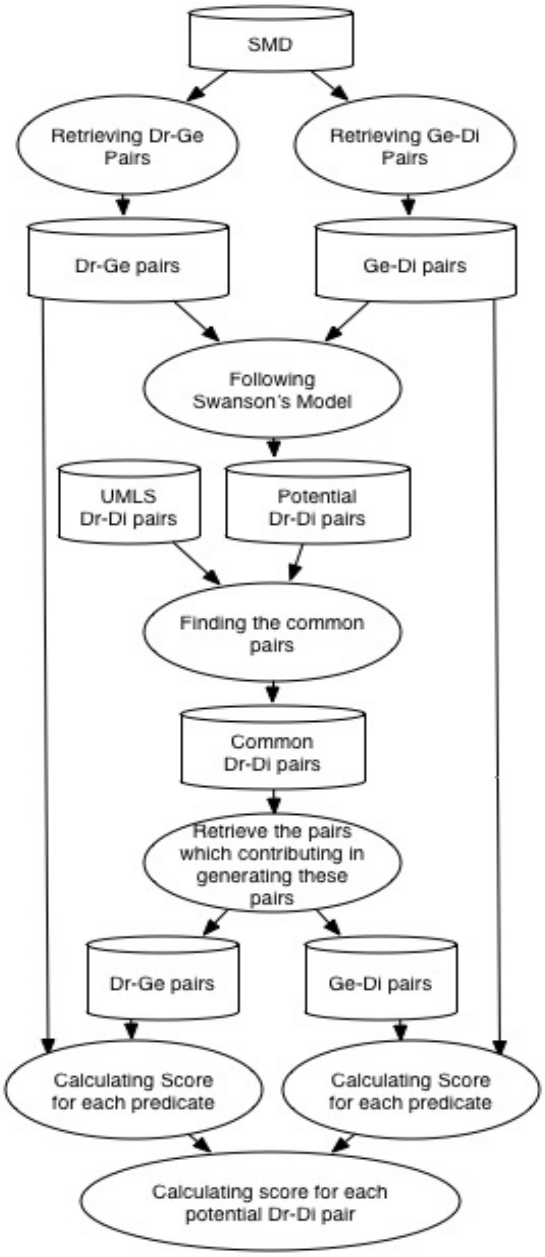

Figure 8: Steps of calculating independence scores

For example, consider the above drug-disease pair (*Strepsils - Chagas*). In order to calculate the score for the pair, we add the ratio of log scores of the individual predicates as outlined in equation (1). For this example, we add the score of the only predicate, "*INTERACTS WITH*" that defines the relationship between the drug-gene pair (Example 1) with the score of all six predicates between the gene-disease pair (Example 2). As mentioned before, more than one predicate may qualify a drug-gene/gene-disease pair, which we consider the summation of the ratio of log scores of all of them. At this point, we do not consider the semantic relatedness of the predicates while calculating their scores.

4.3.1.2 Method of ranking based on predicate inter-dependence

In the second ranking, we assume that the predicates of the drug-gene pair and the gene-disease pair are dependent on each other while estimating their relevance in pairing a drug with a disease. This changes the formula used to compute the final score. Here are the specific steps used to score the pairs:

1) Compute the Percentage Frequency of the combined predicates between the drug-gene and gene-disease pair ($P_{pDG\text{-}pGD}$). We limit this calculation to only those drug-disease pairs that are represented in UMLS drug-disease relations, which we indicate with this notation ($P_{pDG\text{-}pGD\text{-}U}$). For this computation, we count how many times each combined predicate appeared in true drug-disease pairs (generated by the LBD) and then calculate percentage frequency for them.

2) To normalize the percentages, we use the percentage frequency of the combined predicates from SemMedDB ($P_{pDG\text{-}pGD\text{-}S}$) as outlined in the following equation:

$$P_{pDG-pGD-S} = \frac{\#p_{DG} * \#p_{GD}}{\sum_{\substack{i=1 \\ j=1}}^{n,m} \left(\#p_{DG_i} * \#p_{GD_j}\right)} \qquad (2)$$

where *n* and *m* present the number of all different predicates between drug-gene and gene-disease, respectively. $\#p_{DG}$ shows the frequency of the drug-gene predicate in SemMedDB, and $\#p_{GD}$ shows the frequency of gene-disease.

3) Using the percentage frequency, we calculate the raw score for a given generated drug-disease pair as given in the following equation (3).

$$Score = \sum_{i=1}^{n} \log\left(\frac{P_{pDG-pGD-U_i}}{P_{pDG-pGD-S_i}}\right) \qquad (3)$$

where *n* presents the number of combinations which generate that drug-disease pair which is equal to the product of the number of different predicates between the drug-gene pair and the number of predicates between the gene-disease pair.

4.3.1.3 Validation and evaluation

To validate our ranking methods, we use two resources, CTD and Medline citations. After the scoring, the drug-disease pairs are sorted according to the scores, with the highest scoring pairs first. We then calculate and compare the percentage of high ranked and low ranked pairs that occur in the CTD or co-occurred in Medline abstracts.

We cross-reference the generated drug-disease with CTD and then using a statistical test, calculate the correlation between our ranking methods and being a true drug-disease pair (appearing in CTD). We also measure the correlation between the score assigned to each generated pair and the number of times that the pair co-occurred in Medline abstracts. As the last step of validation, an expert reviewed the top 10 ranked drug-disease pairs manually by conducting a web-based search

for each of these pairs and reviewed the relevant scientific literature to identify type of relation between the pairs.

4.3.1.4 Result

All drug-gene and gene-disease semantic predications were retrieved from SemMedDB. There were 19,993 drug-gene pairs (12,666 unique) and 59,945 gene-disease pairs (33,489 unique). When we applied Swanson's model to these pairs, it resulted in the generation of 653,108 potential drug-disease pairs (245,102 unique). Of the roughly 245,000 possible pairs, we found that about 0.5% (N=1,204) of the generated pairs appeared in UMLS. These 1,204 pairs were used to calculate percentage frequency related to each drug-gene and gene-disease predicate.

Figure 9 shows the percentage of high and low ranked drug-disease pairs, which co-occurred in Medline abstracts, for the both, inter-dependence and independence, methods. In this figure, the Y-axis shows the percentage of pairs that co-occurred in Medline and the X-axis shows the number of top ranked pairs. Using a T-test, we found that the inter-dependence method (shown in red) was significant P value < 2.2e-16.

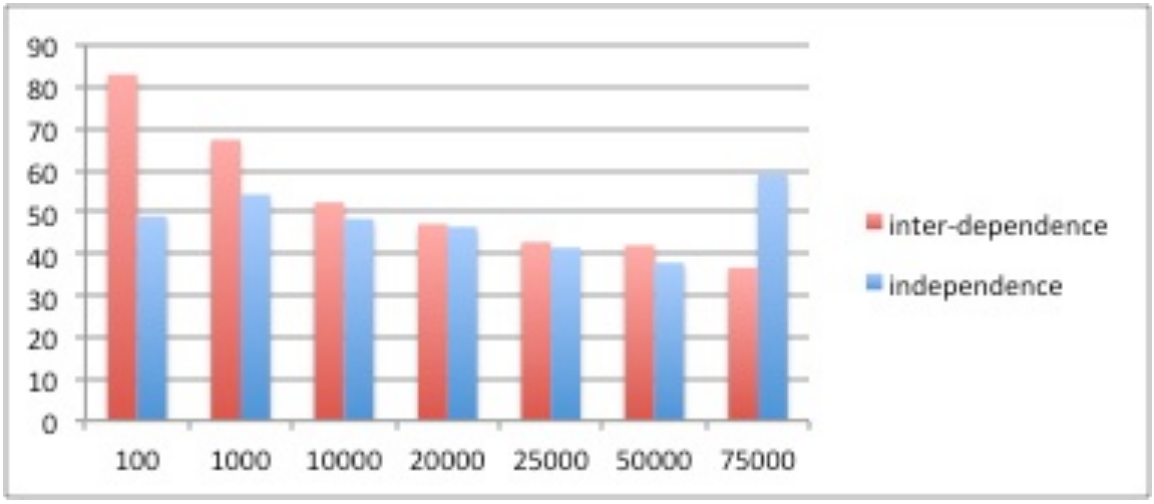

Figure 9: Comparison of the percentage of high and low ranked drug-disease pairs co-occurred in Medline abstracts.

Figure 10 shows the results of calculating the percentage of appearance of just the high and low ranked drug-disease pairs in CTD.

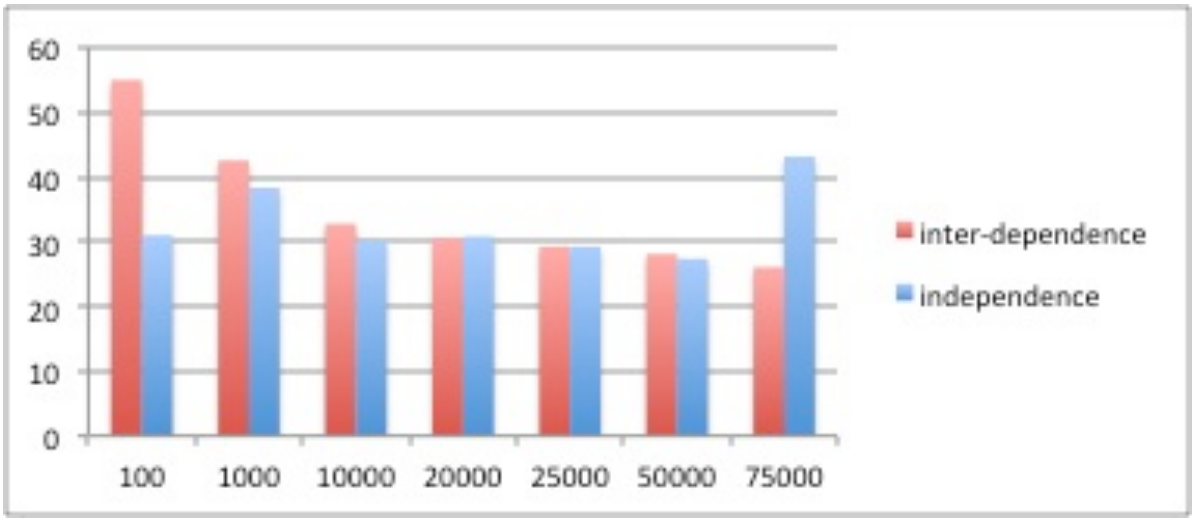

Figure 10: Comparison of the percentage of high and low ranked drug-disease pairs appeared in CTD.

Table 7 includes the result of our manual investigation of top ten ranked drug-disease pairs.

Table 7: Top ten ranked drug-disease pairs

Drug	Disease	Type	Reference
Omalizumab	Asthma	Treatment	[100]
Nifedipine	Tetanus	Treatment	Wikipedia
Nifedipine	Ischemia	Treatment	[101]
Omalizumab	Dermatitis, atopic	Treatment	[102]
Nifedipine	Heart failure	Treatment	[103]
Nifedipine	Renal tubular disorder	Relation	[104]
Calan	Hypertensive disease	Treatment	Not found

Airol	Asthma	-	Not found
Ezetimibe	Coronary heart disease	Treatment	[105]
Cyclosporine	Asthma	Treatment	[106]

4.3.1.5 Discussion

In the prioritization study, we found that the interdependence-based rankings of drug-disease pairs (especially the top ranked pairs) identified through LBD were much more likely to be supported by published evidence than the pairs ranked using the independent ranking approach. Figure 9 shows that 82% of the top 100 drug-disease pairs, ranked using inter-dependence approach had supporting literature-based evidence. These pairs were found to co-occur within a single abstract in Medline. However, there is a noticeable decline in the percentage of pairs as the ranking goes below 100. We observed that pairs ranked using independent ranking approach had relatively lower co-occurrence evidence in the biomedical literature.

We observed a similar trend when we evaluated the confidence levels of the top ranked pairs identified using both approaches against the CTD. Figure 10 further confirms the distinct advantage of the inter-dependence ranking over the independence ranking. Finally, manual evaluation of top ten pairs ranked by inter-dependence approach revealed that the pairs have some biological significance based on expert judgment. This indicates that the inter-dependence method would be useful for identifying biologically relevant drug-disease pairs. Moreover, nine out of ten top ranked drug-disease pairs were found to belong to DRUG-TREATS-DISEASE relationship category.

There are two main limitations in this study. First, we did not have a gold standard of drug-disease treatment pairs to evaluate the performance of our approaches. Second, there is an inherent

limitation both in the choice of resource (choice of CTD as a resource) and the measure (literature co-occurrence) to evaluate the confidence levels of top ranked drug-disease pairs identified by the system. CTD though a manually curated resource does not annotate the type of relationship between the drug and disease. Hence while evaluating our system against CTD we ignored the semantic predications extracted by the system, which would have resulted in loss of valuable information. Alternatively, we relied on document level co-occurrence in literature as a measure to validate drug-disease relationship. Document level co-occurrence of a relation is not a strong indicator for a valid drug-disease relation. There are also limitations in our ranking methods. Using some rigorous statistical validation may further refine the notion of semantic predication as evidence for relation between biomedical entities for LBD.

4.3.1.6 Summary

In this study, we proposed and evaluated two methods for ranking and prioritizing potential drug-repositioning discoveries extracted from literature. We used drug-gene and gene-disease predications, extracted by SemRep, to generate potential drug-disease pairs. The predicates between dug-gene and gene-disease pairs are used to rank the generated drug-disease pairs. Our results showed using combination of drug-gene and gene-disease predicates can be a metric to rank more likely true drug-repositioning candidates higher in the list.

4.3.2 Context-based prioritization

In the second study related to prioritization, we propose a new method, called context-based prioritization. This method utilizes text surrounding A-B and B-C relations (causal associations/findings), to prioritize discoveries generated by LBD systems.

While our previous method (Study 2 in this chapter) looked for evidence in the form of sematic predications, in this approach, we explore a ranking method that utilizes text surrounding causal findings to train a binary classifier that categorizes drug-disease pairs into the following two classes:

a) Positive class (Likely a true drug repositioning candidate)

b) Negative class (Unlikely a true candidate)

The classifier uses two sentences, which drug-gene and gene-disease relations are extracted from, to classify generated drug-disease pairs into one of these two classes.

4.3.2.1 Method of context-based prioritization

Like the two previous studies in this chapter, we use drug-gene and gene-disease semantic predications to generate potential drug-disease pairs. Like previous studies, we follow Swanson's model in our LBD system (more details in section 2.1).

After generating the list of drug-disease pairs, we search these relations in SemMedDB. We consider a drug-disease pair as a positive instance, if there is any semantic predication in SemMedDB that shows treat relationship (predicate) between that drug and disease. The rest of drug-disease pairs in the list, are considered as negative instances.

We randomly select 5000 drug-disease pairs (2500 positive and 2500 negative instances). We retrieve two sentences from Medline abstracts for each of the pairs:

a) one sentence which contains the corresponding causal drug-gene pair

b) one sentence which contains the corresponding causal gene-disease pair

To address the possible time gap between appearing causal pairs (drug-gene and gene-disease) and evidence of drug-disease pair in literature (discussed in section 2.1), we only select the drug-disease pairs with causal relationships published before 2009. To generate some of drug-disease pairs, more than one gene can play role of linking entity. For those drug-disease pairs, we randomly select only one gene.

In the next step, we use these 5000 pairs and the sentences, corresponding to their causal pairs, to train a classifier.

4.3.2.1.1 Features and learning models

We generate two separate sets of features for each of causal pair, however they contain similar features. For each causal pair and corresponding sentences, the following features, determined after some preliminary analysis, are used in the classifier:

1. Words that appeared at least 10 times in the causal sentences
2. Bi-grams that appeared at least 5 times in the causal sentences
3. Tri-grams that appeared at least 5 times in the causal sentences
4. Predicate between the entities

To find the optimal learning model, we compare several common models such as: Random Forest, Naïve Bayes, Support Vector Machine, Decision tree (J48 algorithm) and Rule-based (JRip algorithm). We implement the classifier in Java and use Weka training models [107].

4.3.2.1.2 Ranking

After training, the classifier is used to rank generated drug-disease relations by the LBD system. For each drug-disease relation, we take the three following steps:

1) Classify all pairs of causal sentences, which might lead to a relation (some drug and disease pairs can be inferred by more than one gene and also each drug-gene or gene-disease pair can be extracted from more than one sentence)

2) Count the pairs of causal sentences classified as positive (positive instances) for each drug-disease pair

3) Rank the drug-disease pairs based on number of positive instances (in descending order). In case of a tie, the linking term count [90] is used to rank the pairs (in descending order).

As the aim of LBD is to discover novel findings, we only rank potential novel drug-disease relations and remove already known relations. We consider a drug-disease relation as already known relation, if there is at least one semantic predication that shows relationship between them.

4.3.2.2 Evaluation

We evaluate the classifier and the ranking approach separately.

4.3.2.2.1 Classifier

To evaluate performance of the classifier and find the best learning model, the common metrics, precision, recall, and F-Measure are used.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

$$F - Measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

The classifier is evaluated using 10-fold cross validation.

4.3.2.2.2 Ranking method

To evaluate the ranking method, we generate a test dataset containing drug-disease pairs where their causal drug-gene and gene-disease pairs appeared after 2009 in the literature. We rank these pairs using the classifier and use the ranked list to evaluate our method. Firstly, we study distribution of known drug-disease relations (pairs with at least one evidence predication) in the list.

Secondly, we calculate the two following correlations (using a t-test) between actual class labels (based on evidence predications in SemMedDB) and the following:

1) Number of instances classified as positive
2) Difference of number of instances classified as positive and negative

Thirdly we compare our ranking method with the LTC [69] method. For this comparison, we rank the pairs in the test dataset using both methods and compare the distribution of true pairs in the ranked pairs.

In addition to drug repositioning, we assess our method for two other tasks

1) identifying novel adverse drug reaction (ADR)

2) identifying existence of relation between drug and disease (regardless of relation type).

To train the classifier for identifying ADR, drug-disease pairs augmented with "*predisposes*", "*disrupts*", "*complicates*", and "*causes*" predicts are considered as positive instances and the

others as negative. For latter one, we use CTD [99] to label instances as positive or negative, instead of using the SemMedDB.

4.3.2.3 Results

We retrieved all drug-gene and gene-disease semantic predications from SemMedDB, which yielded 19,993 drug-gene and 59,945 gene-disease pairs. Following Swanson's model, 245,102 unique potential drug-disease relations were generated. Table 8 shows the performance of the classifiers trained using different training models for the drug repositioning task. Table 9 illustrates the performance of the classifier for identifying adverse drug events and Table 10 shows the classifier performance for identifying the relation between drug and disease, regardless of relation types. All these tables show the results of 10-fold cross validation.

Table 8: The performance of the classifier for drug repositioning task

Training Model	Precision	Recall	F-Measure
Random Forest	*0.785*	*0.781*	*0.782*
Naïve Bayes	0.688	0.685	0.686
SVM	0.708	0.712	0.709
J48	0.709	0.714	0.711
JRip	0.686	0.677	0.681

Our calculation showed that there are strong positive correlations between the actual class label and the a) number of positive pairs of causal sentences (p-value = 2.2e-16) and the b) difference between the number of positive and negative pairs (p-value = 1.005e-07).

Table 9: The performance of the classifier for adverse drug event task

Training Model	Precision	Recall	F-Measure
Random Forest	*0.8646*	*0.8600*	*0.8622*
Naïve Bayes	0.7825	0.7819	0.7821
SVM	0.7281	0.7123	0.7201
J48	0.7471	0.7472	0.7471
JRip	0.7910	0.7562	0.7732

Table 10: The performance of the classifier for identifying relation between drug and disease task

Training Model	Precision	Recall	F-Measure
Random Forest	*0.8357*	*0.833*	*0.8343*
Naïve Bayes	0.6895	0.688	0.688
SVM	0.673	0.681	0.6769
J48	0.7699	0.7422	0.7557
JRip	0.6953	0.693	0.6941

Out of 245,102 drug-disease pairs, 6205 pairs met the test dataset criteria. We found out that 180 pairs in the test dataset were true drug-disease pairs (treatment relationship). Figure 11 shows the number of true drug-disease pairs in the ranked test dataset. Each interval shows the number of true pairs in that interval.

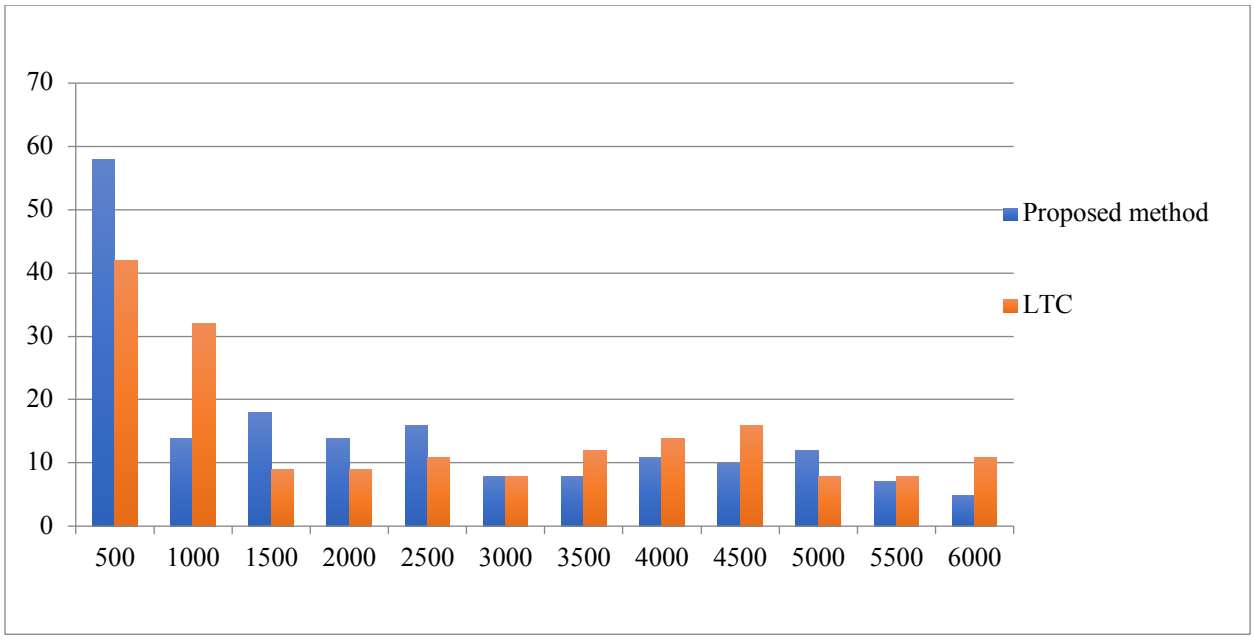

Figure 11: Number of true drug-disease pairs in different intervals of ranked test dataset

We compared true drug-disease pairs in top 100 pairs ranked by LTC and our method. There were 20 true drug-disease pairs in top 100 pairs ranked by our method and only 13 pairs in the pairs ranked by LTC. Figure 12 shows result of this comparison and intersection of true drug-disease pairs in top 100 pairs. We observed a similar trend in the top 500 pairs. Our method ranked the 58 true pairs in the top 500 pairs and LTC ranked 42 true pairs (38 pairs in common).

4.3.2.4 Discussion

Table 2, 3, and 4 show the performance of our classifier. The performance reported in Table 8 confirms that using the surrounding text of drug-gene and gene-disease causal relations can help determine the likelihood of being a true drug repositioning candidate for generated drug-disease pairs. Using the random forest model, the classifier obtained f-measure of 0.78, which is a reasonable performance for a classifier for a task of this complexity. The performance of the classifier for two other tasks, identifying adverse drug events and relation detection, also supports

the proposed approach. Random forest achieved the highest f-measure among the five learning models for those two tasks, 0.86 for adverse drug event (Table 9) and 0.83 for relation detection (Table 10). As two different public resources are used as silver standard datasets (SemMedDB and CTD) to train and test the classifiers for these tasks, the results demonstrate that the performance of the classifier is not dependent on the particular dataset.

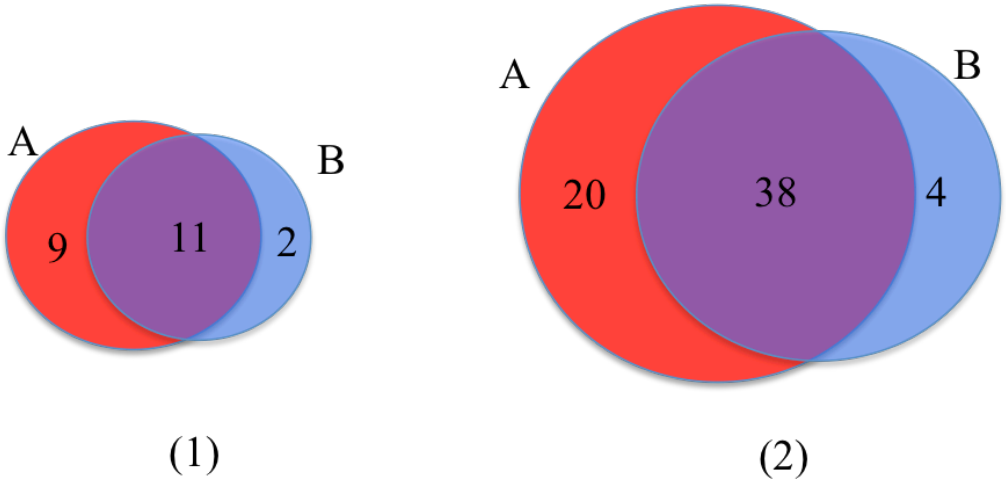

Figure 12: Comparing true drug-disease pairs in top pairs ranked by (A) our method and (B) LTC.

However, we found strong correlations between the actual class label and both the number of positive instances and the difference between the number of positive and negative instances, but the correlation between actual class label and earlier is stronger. These results support using the number of instances classified as positive in our ranking method.

After ranking drug-disease pairs in our test dataset, we studied the distribution of true drug repositioning pairs in the list and as Figure 11 shows, the top ranked pairs include most of the true pairs. We identified 32% (58 out of 181) of the true drug repositioning pairs in the top 8% (500 out of 6000) of ranked pairs. We plotted this distribution for both our ranking and LTC methods in Figure 11. The distributions showed that our approach ranked more true pairs in top of the list

comparing to LTC method. The next comparison illustrated intersection between top ranked true drug repositioning pairs by both methods. Part 1 in Figure 12 demonstrates that our method was able to rank 84% (11 out of 13) of true drug repositioning pairs ranked by LTC in top 100, in its own top 100 pairs. This number for top 500 pairs (part 2, Figure 12) was even higher 92% (38 out of 42). These two Venn diagrams clearly demonstrate that the majority of true pairs are in common between both methods, but our method was able to rank more true pairs in higher ranks of the list.

There are two main limitations in this study, which first one is the lack of access to a gold standard, the same as in the previous study. Another limitation is using a balanced dataset (50% positive and 50% negative cases) for training the classifiers. At this point, we have not trained/evaluated our classifiers using an unbalanced dataset.

However, the results show that our proposed method can play a role in prioritizing LBD findings, but still provide pharmaceutical companies with a long list of potential candidates. As future work, we plan to not only prioritize the candidates, but also identify and remove false positive candidates from the list.

4.3.2.5 Summary

In this study [108], we created and tested a new method to prioritize discoveries identified by our LBD system. Our LBD system, which is based on Swanson's model, uses semantic predications as causal findings to hypothesize new findings. To rank the generated hypotheses, we trained a binary classifier using information surrounding causal findings in literature. We trained our classifier for three different purposes, drug repositioning, adverse drug event, and drug-disease relation detection. Our classifier obtained reasonable f-measures for all the tasks (respectively, 0.78, 0.86, and 0.83). Our results showed that the proposed method performed better than one of previous methods in ranking true drug repositioning candidates at the top of the list.

Chapter 5: Using biomedical literature and clinical trials to prioritize potential candidates

5.1 Introduction

In chapter 4, we studied two ranking methods for potential drug repositioning candidates generated by LBD systems. In this chapter, we study ranking the candidates generated by information from a different source, phenome-wide associations studies (PheWAS). The goal of this chapter is to assess this method of using text data in prioritizing non-LBD discoveries.

PheWAS provide evidence for the association of genetic variants with a wide spectrum of human disorders[109]. The PheWAS strategy relies on electronically available phenotypic data collected from patient cohorts. PheWAS is similar to a genome-wide association study (GWAS), but whereas a GWAS asks *"What genetic variants are associated with a disease?"*, a PheWAS asks *"What diseases are associated with a genetic variant?"*. PheWAS, similar to GWAS, provides us a list of gene-disease associations which we utilize to generate a list of potential discoveries.

5.1.1 Related works of Phenome-Wide Association Studies

In 2010, Denny et al. [45] introduced PheWAS and demonstrated an approach to discover gene-disease associations using genetic data coupled to longitudinal electronic medical records. Later, Nature Biotechnology published a study by Denny et al. [110] in which they conducted a comprehensive PheWAS on 3,144 single-nucleotide polymorphisms (SNPs) that had been previously associated with a variety of phenotypes by GWAS. This PheWAS, like many other PheWAS published [109]–[111] used International Classification of Diseases version 9 (ICD9) codes extracted from electronic medical record systems in large patient cohorts to define case-control groups for many phenotypes. In the United States, ICD9 coding is primarily used for billing and can have variable effectiveness for describing discrete phenotypes. Regardless, Denny et al. [110] demonstrate that for many of the GWAS SNPs, PheWAS was able to rediscover expected SNP- disease associations while also identifying novel associations [110]. This suggests that

PheWAS results may also provide new opportunities to identify candidates for drug repositioning. However, it is also likely that false positives exist in the reported PheWAS data, especially for those SNP-disease associations with moderate evidence of association. Independent PheWAS may prove highly effective when replicating novel findings. Regardless, to balance accuracy while still allowing for discovery and hypothesis generation, one might choose a loose P-value threshold from PheWAS data.

5.1.2 GWAS vs PheWAS in drug repositioning

As mentioned before, drug repositioning is the process of discovering new indications for existing drugs [112]. Critical to drug repositioning is the initial identification of candidate drug-disease relationships. Genetic-based association studies, including GWASs, have proven to be effective for generating hypotheses related to drug repurposing [113], [114]. GWASs can identify disease susceptibility genes that are targets for existing drugs used to treat different conditions. For example, a large GWAS implicated flavopiridol, a CDK4 inhibitor and anti-cancer agent, as a possible drug to repurpose for the treatment of rheumatoid arthritis [115]. However, GWAS data is partly limited by the number of diseases that can be linked to a gene. A unique advantage of the PheWAS approach is the ability to measure genetic associations with thousands of diseases simultaneously, which may be ideal when identifying pleiotropic effects. As a result, PheWAS may prove to be an effective alternative to GWAS when identifying susceptibility loci. Importantly, PheWAS may further expand the horizon for discovery and prioritization of candidates for drug repositioning [116].

5.2 Method

We seek to demonstrate the potential use of PheWAS information for drug repositioning by cross-referencing PheWAS data with biomedical databases.

5.2.1 Generating Candidates

We begin by extracting all SNP-disease associations from the recently reported PheWAS by Denny et al. [110]. We filter the association based on P-value. We use a loose P-value threshold ($P \leq 0.05$). Only those SNPs that are mapped directly to a gene according to dbSNP, which includes 2 kb upstream and 500 bp downstream, are considered in the study. In the next step, we retrieve all drugs from DrugBank and their direct and indirect gene targets and generate a list of drug-gene relations. These drug target data have been previously applied to other in silico drug-screening studies [117]–[119]. Then, we look for all sets of transitive pairs, A-B, B-C where A-B is a disease-gene pair from PheWAS and B-C is a gene-drug pair from DrugBank. (Figure 13).

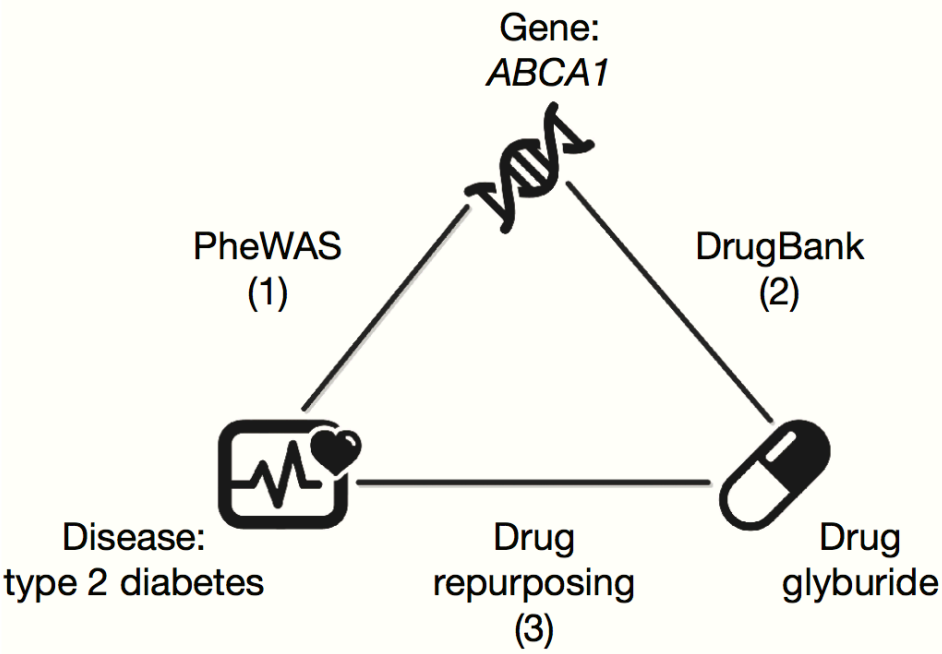

Figure 13: The three steps in the discovery process. (1) PheWAS associations [108] to connect diseases to genes; (2) DrugBank analysis to connect genes to drugs; and (3) drug repurposing targets that connect drugs with diseases. Highlighted is an example where this process rediscovered glyburide as an indication to treat type 2 diabetes.

5.2.2 Validation and Ranking Candidates

To measure the relevance of the generated drug-disease pairs identified using PheWAS data under a loose P-value cutoff, we assess the co-occurrence of the drug and disease terms in Medline abstracts. For this purpose, MetaMap [95], is used to extract disease names from the PheWAS phenotypes and then we scan Medline abstracts for the paired drug-disease terms using a text search engine, Apache Lucene.

To validate our method, we generate a list of drug-disease pairs (the same number of pairs as our method generated) where both drug and disease names are randomly selected from a list of drugs and diseases. Then we follow the same approach to identify any co-occurrence of these pairs in biomedical literature. Finally, we compare the percentage of the pairs in both lists which have at least one co-occurrence in literature. We repeat the process 1000 times and look for any significant difference in the percentages.

For comparison purposes, the process is also performed using gene-disease pairs driven solely by GWAS data extracted directly from the GWAS catalog [120].

To assess the novelty of the drug-disease pairs and to better rank candidates for drug repositioning, we further cross-reference all pairs with the clinical trial registry (clinicaltrials.gov). We categorize the pairs into four categories, including:

- known/rediscovered
- strongly supported (some support in the literature and clinical trial registry)
- likely (some support in the literature or clinical trial registry)
- novel (no evidence in the databases)

5.3 Result

We extracted all 212,851 SNP-disease associations from PheWAS catalog [110]. Only 1,501 SNPs were mapped directly to a gene according to dbSNP, which allowed us to identify 48,488 unique gene-disease relationships. Using gene-drug relations in DrugBank, we were able to identify 52,966 drug-disease pairs.

Of the 52,966 drug-disease pairs, 14,816 (28%) were supported in the literature. After permuting drug-disease pairs 1,000 times and cross-referencing random drug-disease pairs with Medline abstracts, we identified a total of 3,270 ± 57 drug-disease pairs (6.2%; Figure 14-a) that co-occurred in literature, which suggests that PheWAS data significantly enriches information for drug-disease pairs supported by evidence. Of the 52,966 drug-disease pairs, 127 pairs were 'rediscovered' according to original indications listed in DrugBank (Table 11). As an example, the PheWAS catalog reported an association between SNP rs2515629 and 'type 2 diabetes' (P = 0.000732). A search of dbSNP maps rs2515629 to the gene ABCA1, which is mentioned in DrugBank as a gene target for glyburide, a drug used to treat type 2 diabetes. Therefore, this process appropriately identified type 2 diabetes as an indication for glyburide (Figure 13). By contrast, using the GWAS approach, the number of drug-disease pairs identified was 7,945 of which 3,087 (38.8%) co-occurred in at least one abstract (Table 11). When the same permutation procedure was established for the GWAS data, a total of 489 ± 21 pairs (6.2%) appeared at least once in the abstracts (Figure 14-b).

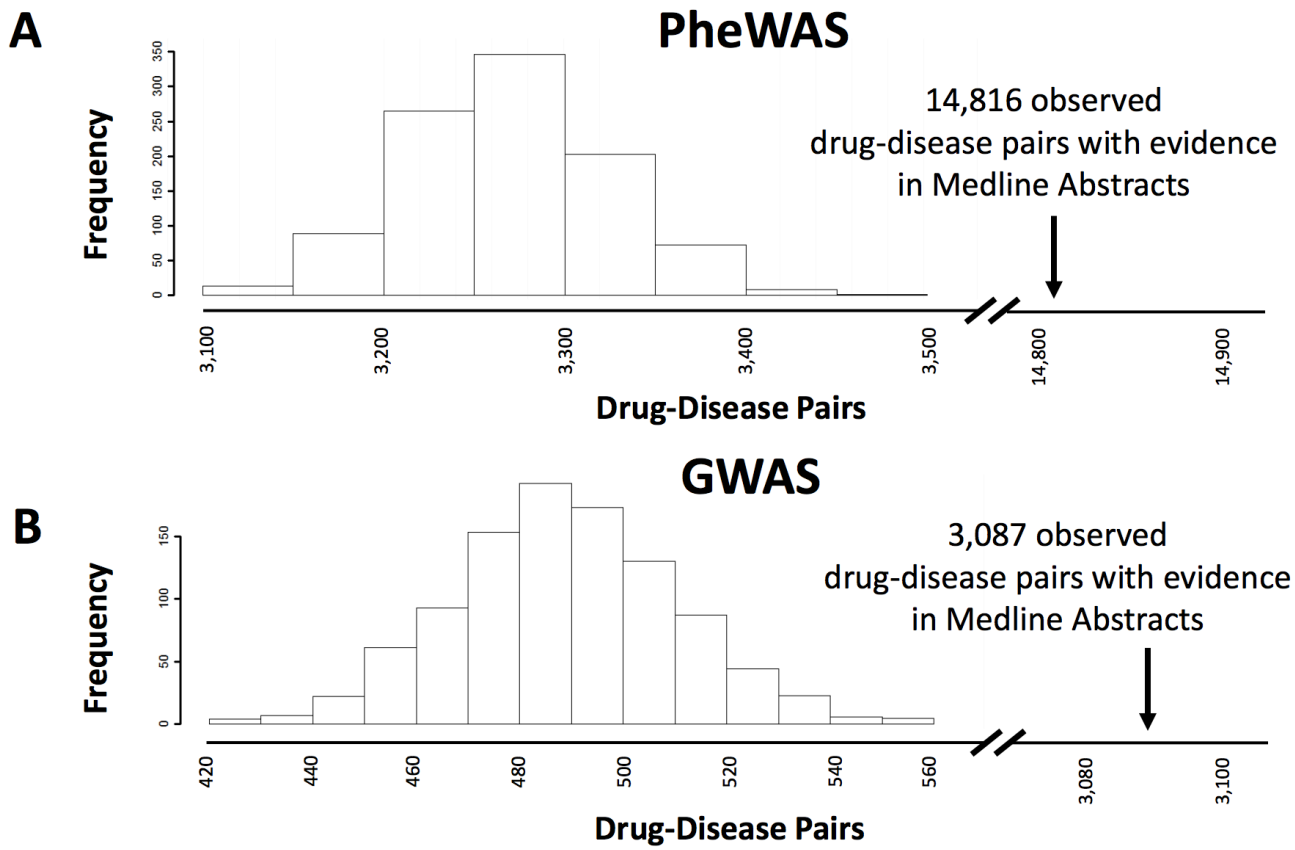

Figure 14: Co-occurrence distribution plotted for 1000 randomly permuted drug-disease pairs identified in Medline Abstracts for (A) PheWAS- and (B) GWAS-derived data. For each dataset, the number observed drug-disease pairs with evidence in Medline Abstracts is also highlighted. This contrast demonstrates that there is a significant difference between the observed drug-disease pairs in Medline Abstracts from those randomly permuted.

The resulting distribution of the 52,966 PheWAS-derived pairs and 7,945 GWAS-derived pairs into the four categories of discovery, are shown in Table 11.

Table 11: Comparison of PheWAS- and GWAS-based approaches to drug repositioning

	GWAS	PheWAS
Drug-Disease pairs	7945	52,966
Drug indication type		
Known/rediscovered	140 (1.8%)	127 (0.2%)

Strongly supported	908 (11.4%)	2,583 (4.9%)
Likely	2,060 (25.9%)	12,221 (23.1%)
Novel (candidates for drug repositioning)	4,837 (60.9%)	38,035 (71.8%)

Some examples from each category driven by PheWAS data are illustrated in Table 12. For example, PheWAS results indicate that SNP rs2736100 in the gene TERT is associated with diabetes (P = 0.00029). Zidovudine, a drug prescribed to treat HIV/AIDS, is a reverse transcriptase inhibitor and may inhibit TERT activity. This example suggests that zidovudine might be repositioned to treat diabetes, a not unreasonable assertion as increased telomerase activity has been reported to be associated with increased diabetes complications [121]. It should be noted that the co-occurrence of drug-disease pairs may also be the result of adverse drug events [122].

Table 12: Examples of Drug-Disease pairs identified from PheWAS data

Status	**Drug**	**Disease Indication**	**PheWAS SNP**	**PheWAS *P*-value**	**Associated Gene**	**Medline Citations/ Clinical Trial Citations (count)**
Known	Paclitaxel	breast cancer	rs242557	0.041	*MAPT*	3003/321
Known	Glyburide	type II diabetes	rs2515629	0.00073	*ABCA1*	240/12
Strongly supported	Dexamethasone	rheumatoid arthritis	rs4795067	0.050	*NOS2*	4271/48
Strongly supported	Everolimus	breast cancer	rs17036350	0.040	*MTOR*	221/46
Likely	Verapamil	vaginal cancer	rs216013	0.011	*CACNA1C*	1144/0
Likely	Chlorpromazine	liver cancer	rs11214606	0.033	*ARVCF*	423/0
Novel	Porfimer	hyper-cholesterolemia	rs6511720	2.5E-6	*LDLR*	0/0
Novel	Zidovudine	diabetes	rs2736100	0.00029	*TERT*	0/0

5.4 Discussion

This study offers many promising new candidates to explore. Complete results from this study are available to the scientific community for further experimental validation and clinical study [22], [23]. For example, an investigator interested in developing a new drug for breast cancer would have 18 candidates, including oxcarbazepine and disopyramide, for investigation. Similarly, new indications may be identified for existing medications. On the basis of our analysis, docetaxel has 88 potential new indications, including psoriasis, migraine and osteoporosis.

Because the SNPs selected for the PheWAS were derived from reported GWAS, it is not surprising that several drug-disease pairs identified by PheWAS overlap those identified by GWAS. Specifically, there were 161 overlapping drug-disease pairs between GWAS and PheWAS data; 70 were initially identified as 'novel' (no evidence in the databases). The 161 overlapping pairs may be an under- representation of the true overlap as phenotypic terms used by the GWAS may differ from those defined by the PheWAS, especially GWAS that assessed quantitative traits. For example, the PheWAS identified a link between LDLR, a target for Porfimer [93], and *"hyper-cholesterolemia,"* whereas GWAS connected LDLR to "*cholesterol*" (Table 11).

As mentioned previously, we chose a loose P-value threshold for PheWAS data for discovery and hypothesis generation. On the basis of permutation analysis, many of these associations may be real, but these are undoubtedly mixed with false positives. We investigated the effect of various P-value thresholds. When using a P-value threshold of ≤ 0.001, approximately 250 drug-disease pairs with evidence were observed. This more stringent P-value threshold may decrease false positives but also dramatically reduce the number of potential candidates. Interestingly, as P values decreased below $1.0 \times 10–5$, the proportion of drug-disease pairs with evidence in the literature increased (Figure 15). Additional replication of PheWAS findings should

improve the association results. Furthermore, incorporating other biomedical databases, such as SemMedDB [78] and the Drug-Gene Interaction Database [123], in combination with manual curation, may also significantly benefit the prioritization process.

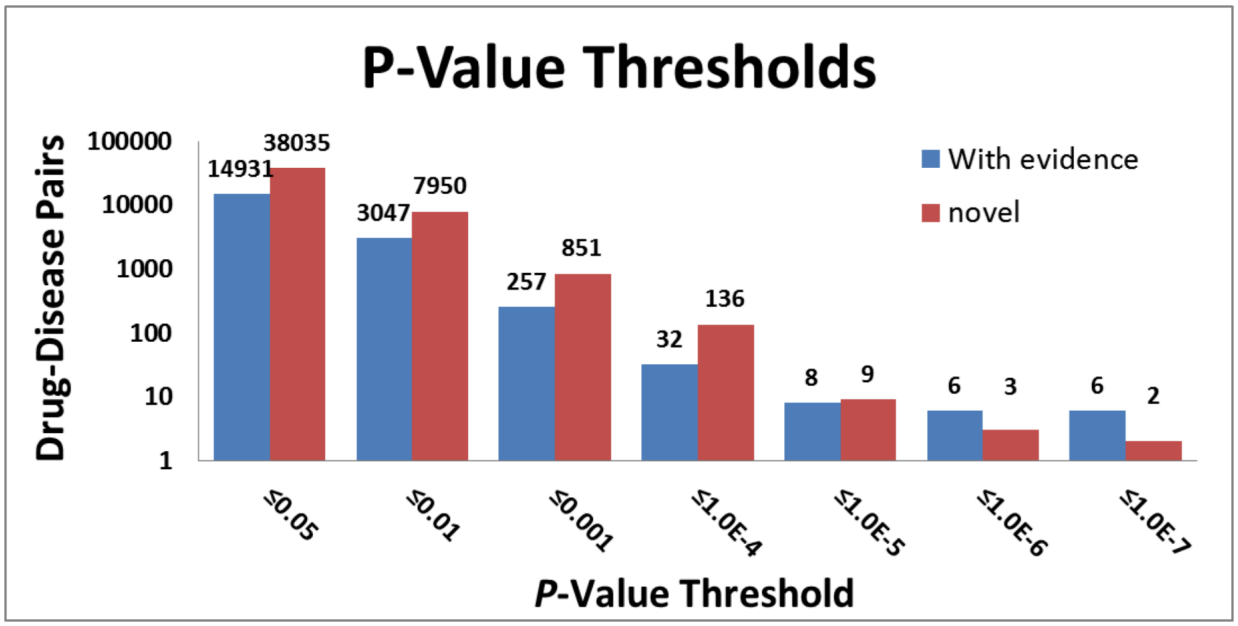

Figure 15: Number of drug-disease pairs by P-value threshold for those pairs with and without evidence according to Medline Abstracts, Clinical Trial Registry, or DrugBank.

Even with many drug-disease pairs identified through this raw screen, other factors of course influence whether candidates for drug repurposing may be suitable for commercial development. Economic factors may limit some drug candidates. Pharmaceutical companies may not support drug repositioning due to insufficient market potential driven by disease frequencies, competing medications, and/or intellectual property protection. In addition to economic constraints, some agents may require reformulation due to therapeutic index, toxicity profile and/or agent type (topical agents versus systemic agents). A second level 'screen-out' approach would further help identify candidates for repositioning.

5.5 Summary

The conclusions of this study are two-fold. First, we suggest that the utilization of PheWAS data provides a robust approach for identifying new drug candidates for repurposing. Second, we illustrate how to use biomedical literature and clinical trials to rank non-LBD drug repositioning candidates. In this study, we identified nearly 14,800 drug-disease pairs with some evidence of support in biomedical literature or clinical trials. In addition, we identified more than 38,000 novel candidates for re-purposing, encompassing hundreds of different disease states and over 1,000 individual medications. We anticipate that these results will be highly useful for hypothesis generation in the field of drug repurposing.

Chapter 6: Conclusion

6.1 Introduction

The drug development process is a prolonged and expensive process. In this thesis, we examined drug repositioning as a popular alternative approach to reducing the cost and time needed for developing a new drug. We explored the opportunity of identifying potential drug repositioning candidates by leveraging publicly available text-based resources. We conducted five studies and utilized reviews posted by patients to WebMD, abstracts from the biomedical literature, clinical trials, and PheWAS data as the main resources for generating and prioritizing the potential drug repositioning candidates. We developed a rule-based system and studied the feasibility of using patient reviews in drug repositioning. To our knowledge, this is the first study to consider using social media data for this purpose. We were able to detect several beneficial effects of medications reported by patients, which highlighted the value of social media in drug repositioning.

In three studies, we focused on literature-based discovery. First, we presented a method to compare the relative importance of using either isolated sentences or multiple sentences comprising a discourse to identify potential candidates. We found that nearly three quarters of the findable candidates would require mining candidates using multiple sentences. In the other two studies, we proposed statistical methods to prioritize the potential candidates. In the first method, we utilized predicates in semantic predications extracted from Medline abstracts to rank the candidates. Our results illustrated that predicates can be a metric to rank more likely true drug-repositioning candidates higher in the list, however this approach is limited to LBD systems that utilize semantic predication for discovery. The other method utilized the text surrounding discoveries to train several classifiers and used the classifiers to score and rank the potential candidates. Our analysis showed the effectiveness of this method as well. These two studies illustrated that by using

computational methods and existing knowledge sources, we can rank and narrow down the long lists of the potential candidates generated by LBD systems.

In the fifth study, we utilized text data to prioritize and categorize potential drug repositioning candidates generated by a non-LBD system. In this study, we used Phenome-Wide Association Studies to generate a list of potential candidates and then utilized text from the biomedical literature and clinical trials to categorize and prioritize the candidates. The results of this study showed that PheWAS can be used to generate new hypotheses for drug repositioning and that biomedical literature and clinical trials can be used to prioritize the candidates and filter out already known candidates.

6.2 Comparison

In this section, we will compare the systems presented in this thesis based on the number of potential drug repositioning candidates generated by each. We presented three main systems to generate the potential candidates: 1) LBD using semantic predication, introduced in chapter 4 2) a PheWAS based system, presented in chapter 5 and 3) a GWAS based system, introduced in chapter 5 as well. Table 13 and figure 16 show the number of potential candidates generated by each system, and the percentage of the candidates categorized as known, supported in literature (based on co-occurrence of drug and disease), and novel.

Table 13: Comparing the systems based on the number of generated candidates

System	Total number	Already known	Supported in literature	Novel
LBD using sematic predication	245,102	1,204 (0.49%) (in UMLS)	53,922 (22%)	189,976 (77.5%)
PheWAS based	52,966	127 (0.2%) (in DrugBank)	14,816 (28%)	38,035 (71.8%)
GWAS based	7,945	140 (1.8%) (in DrugBank)	2,963 (37.3%)	4,837 (60.9%)

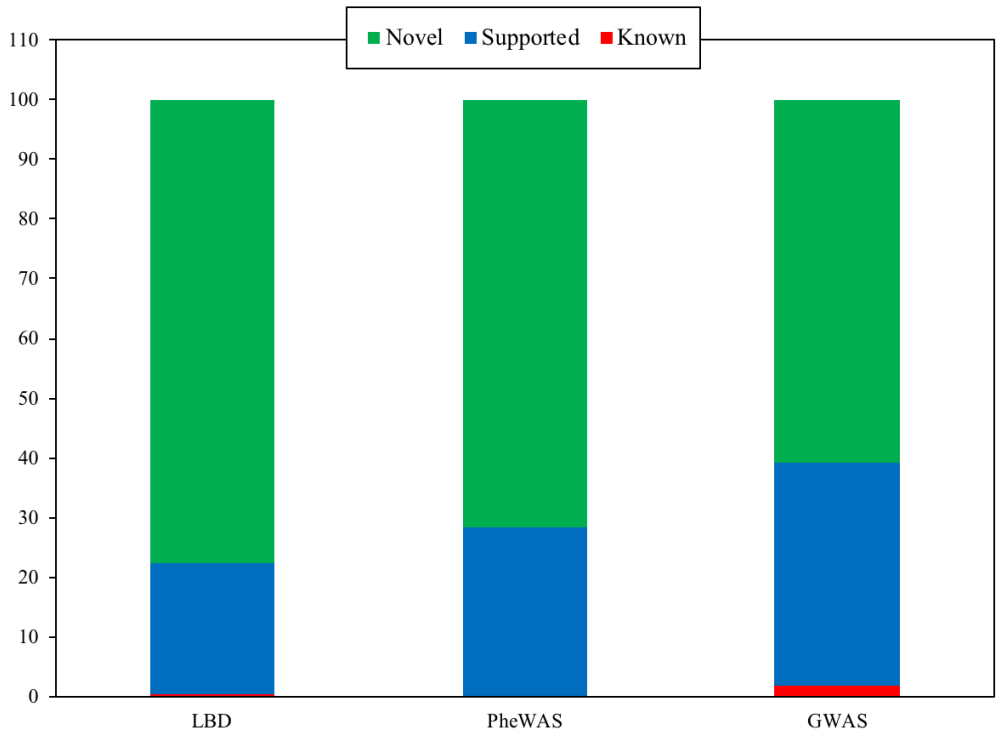

Figure 16: Comparing the systems based on the number of drug repositioning candidates

By examining the results in Table 13, it is clear that the LBD system generated more potential drug repositioning candidates than the others, but it provided the smallest percentage of candidates supported by evidence from the literature. The GWAS-based system generated only 7,945 candidates with 37.3% supported by the literature, which suggests a lower false positive rate, but has a limitation of generating the smallest number of novel candidates. Figure 14 indicates that for all three systems, the biggest proportion of the candidates are novel and thus there is a need to prioritize them and remove false positive candidates.

6.3 Future work

In this thesis, we assessed methods for leveraging three important text-based resources, social media, biomedical literature, and clinical trials. We did not include one of the most important text-based resources now available, clinical notes, because they are not public. Exploring clinical notes as resource to identify and prioritize the potential candidates is a potentially valuable extension to this work, as they might provide the earliest evidence of off-label uses of prescribed medications.

For each study described here, we mentioned the limitations which could be considered as potential future work. For example, for the social media study, as we only explored the feasibility of using social media for drug repositioning, and only considered data from one source, many extensions are possible such as: exploring the other websites (tweeter, Facebook, …), creating an annotated dataset and developing statistical methods, using state-of-the-art NLP algorithms to address known issues with consumer-generated text, such as mis-spellings and ungrammaticality. To address the need for better methods for prioritizing candidates and removing more false positives, the next step would be to explore incorporating other biomedical databases, such as SemMedDB [78] and the Drug-Gene Interaction Database [123], in combination with manual curation.

References

[1] C. P. Adams and V. V. Brantner, “Estimating the cost of new drug development: is it really 802 million dollars?,” *Health Aff. Proj. Hope*, vol. 25, no. 2, pp. 420–428, Apr. 2006.

[2] S. Kato *et al.*, “Challenges and perspective of drug repurposing strategies in early phase clinical trials,” *Oncoscience*, vol. 2, no. 6, pp. 576–580, 2015.

[3] E. L. Tobinick, “The value of drug repositioning in the current pharmaceutical market,” *Drug News Perspect.*, vol. 22, no. 2, pp. 119–125, Mar. 2009.

[4] M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal, “Computational drug repositioning: from data to therapeutics,” *Clin. Pharmacol. Ther.*, vol. 93, no. 4, pp. 335–341, Apr. 2013.

[5] T. T. Ashburn and K. B. Thor, “Drug repositioning: identifying and developing new uses for existing drugs,” *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–683, Aug. 2004.

[6] J. Gilbert, P. Henske, and A. Singh, *Rebuilding Big Pharma’s business model*. In Vivo, 2003.

[7] F. F. Yonkman, “[New drugs for old uses and new uses for old drugs],” *J. - Mich. State Med. Soc.*, vol. 58, no. 6, pp. 913–918, Jun. 1959.

[8] A. J. Atkinson, M. J. Finkel, J. J. Burns, G. H. Hitchings, B. A. Kemp, and S. R. de Dennis, “Panel on public service drugs and new uses for old drugs,” *Clin. Pharmacol. Ther.*, vol. 18, no. 5 Pt 2, pp. 659–662, Nov. 1975.

[9] “Rheumatologists gauge new uses of old drugs,” *Hosp. Pract. Off. Ed*, vol. 16, no. 8, pp. 37, 40-40A, 40D passim, Aug. 1981.

[10] E. Harris, “Antidepressants: old drugs, new uses,” *Am. J. Nurs.*, vol. 81, no. 7, pp. 1308–1309, Jul. 1981.

[11] M. A. Mathewson, “New uses for old drugs: vasodilators (CEU home study),” *Crit. Care Update*, vol. 9, no. 11, pp. 7–13, Nov. 1982.

[12] N. Novac, “Challenges and opportunities of drug repositioning,” *Trends Pharmacol. Sci.*, vol. 34, no. 5, pp. 267–272, May 2013.

[13] D. J. Coleman *et al.*, “Treatment of Macular Degeneration with Sildenafil: Results of a Two-Year Trial,” *Ophthalmol. J. Int. Ophtalmol. Int. J. Ophthalmol. Z. Augenheilkd.*, vol. 240, no. 1, pp. 45–54, 2018.

[14] S. Kumar, T. E. Witzig, and S. V. Rajkumar, “Thalidomide as an anti-cancer agent,” *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 160–174, Jun. 2002.

[15] W. H. Jost and P. Marsalek, “Duloxetine in the treatment of stress urinary incontinence,” *Ther. Clin. Risk Manag.*, vol. 1, no. 4, pp. 259–264, Dec. 2005.

[16] G. Jin and S. T. C. Wong, “Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines,” *Drug Discov. Today*, vol. 19, no. 5, pp. 637–644, May 2014.

[17] D. Grau and G. Serbedzija, “Innovative Strategies for Drug Repurposing,” *Drug Discov. Dev.*, 2005.

[18] J. T. Dudley, T. Deshpande, and A. J. Butte, “Exploiting drug-disease relationships for computational drug repositioning,” *Brief. Bioinform.*, vol. 12, no. 4, pp. 303–311, Jul. 2011.

[19] H. Xu *et al.*, “Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 22, no. 1, pp. 179–191, Jan. 2015.

[20] Y. Wang *et al.*, “Clinical information extraction applications: A literature review,” *J. Biomed. Inform.*, vol. 77, pp. 34–49, 2018.

[21] P. Sanseau *et al.*, "Use of genome-wide association studies for drug repositioning," *Nat. Biotechnol.*, vol. 30, no. 4, pp. 317–320, Apr. 2012.

[22] M. Rastegar-Mojarad, Z. Ye, J. M. Kolesar, S. J. Hebbring, and S. M. Lin, "Opportunities for drug repositioning from phenome-wide association studies," *Nat. Biotechnol.*, vol. 33, no. 4, pp. 342–345, Apr. 2015.

[23] S. Moosavinasab *et al.*, "'RE:fine drugs': an interactive dashboard to access drug repurposing opportunities," *Database J. Biol. Databases Curation*, vol. 2016, 2016.

[24] C.-P. Wei, K.-A. Chen, and L.-C. Chen, "Mining Biomedical Literature and Ontologies for Drug Repositioning Discovery," in *Advances in Knowledge Discovery and Data Mining*, V. S. Tseng, T. B. Ho, Z.-H. Zhou, A. L. P. Chen, and H.-Y. Kao, Eds. Springer International Publishing, 2014, pp. 373–384.

[25] M. Rastegar-Mojarad, R. K. Elayavilli, D. Li, R. Prasad, and H. Liu, "A new method for prioritizing drug repositioning candidates extracted by literature-based discovery," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, pp. 669–674.

[26] H. Liu and M. Rastegar-Mojarad, "Literature-Based Knowledge Discovery," in *Big Data Analysis for Bioinformatics and Biomedical Discoveries*, CRC Press, 2016, pp. 233–248.

[27] J. Li and Z. Lu, "Pathway-based drug repositioning using causal inference," *BMC Bioinformatics*, vol. 14, no. S16, pp. 1–10, Oct. 2013.

[28] F. Tan *et al.*, "Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity," *Mol. Biosyst.*, vol. 10, no. 5, pp. 1126–1138, May 2014.

[29] L. Yang and P. Agarwal, "Systematic Drug Repositioning Based on Clinical Side-Effects," *PLOS ONE*, vol. 6, no. 12, p. e28025, Dec. 2011.

[30] J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug-disease relationships for computational drug repositioning," *Brief. Bioinform.*, vol. 12, no. 4, pp. 303–311, Jul. 2011.

[31] H. Mei, T. Xia, G. Feng, J. Zhu, S. M. Lin, and Y. Qiu, "Opportunities in systems biology to discover mechanisms and repurpose drugs for CNS diseases," *Drug Discov. Today*, vol. 17, no. 21–22, pp. 1208–1216, Nov. 2012.

[32] Y. Wang *et al.*, "PubChem BioAssay: 2014 update," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D1075–D1082, Jan. 2014.

[33] T. Cheng, Y. Pan, M. Hao, Y. Wang, and S. H. Bryant, "PubChem applications in drug discovery: a bibliometric analysis," *Drug Discov. Today*, vol. 19, no. 11, pp. 1751–1756, Nov. 2014.

[34] R. Hoehndorf, A. Oellrich, D. Rebholz-Schuhmann, P. N. Schofield, and G. V. Gkoutos, "Linking PharmGKB to phenotype studies and animal models of disease for drug repurposing," *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 388–399, 2012.

[35] F. Moriaud *et al.*, "Identify drug repurposing candidates by mining the protein data bank," *Brief. Bioinform.*, vol. 12, no. 4, pp. 336–340, Jul. 2011.

[36] F. Moriaud *et al.*, "Computational fragment-based approach at PDB scale by protein local similarity," *J. Chem. Inf. Model.*, vol. 49, no. 2, pp. 280–294, Feb. 2009.

[37] D. K. Rajpal, X. A. Qu, J. M. Freudenberg, and V. D. Kumar, "Mining emerging biomedical literature for understanding disease associations in drug discovery," *Methods Mol. Biol. Clifton NJ*, vol. 1159, pp. 171–206, 2014.

[38] L. B. Tari and J. H. Patel, “Systematic drug repurposing through text mining,” *Methods Mol. Biol. Clifton NJ*, vol. 1159, pp. 253–267, 2014.

[39] R. Xu and Q. Wang, “Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing,” *BMC Bioinformatics*, vol. 14, p. 181, Jun. 2013.

[40] S. N. Deftereos, C. Andronis, E. J. Friedla, A. Persidis, and A. Persidis, “Drug repurposing and adverse event prediction using high-throughput literature analysis,” *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 3, no. 3, pp. 323–334, Jun. 2011.

[41] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, “Novel data-mining methodologies for adverse drug event discovery and analysis,” *Clin. Pharmacol. Ther.*, vol. 91, no. 6, pp. 1010–1021, Jun. 2012.

[42] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, “Social Media Mining for Drug Safety Signal Detection,” in *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, New York, NY, USA, 2012, pp. 33–40.

[43] S. Boggan, “The ‘miracle’ treatment that’s bringing the brain-damaged back to life,” *The Guardian*, 12-Sep-2006.

[44] 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,” in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Stroudsburg, PA, USA, 2010, pp. 117–125.

[45] J. C. Denny *et al.*, “PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations,” *Bioinforma. Oxf. Engl.*, vol. 26, no. 9, pp. 1205–1210, May 2010.

[46] R. Azondekon, Z. J. Harper, F. R. Agossa, C. M. Welzig, and S. McRoy, "Scientific authorship and collaboration network analysis on malaria research in Benin: papers indexed in the web of science (1996-2016)," *Glob. Health Res. Policy*, vol. 3, p. 11, 2018.

[47] M. Weeber, H. Klein, L. T. W, J. Berg, and D. R. S. Has, "Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migrainemagnesium discoveries," *J Am Soc Inf Sci Tech*, pp. 548–557, 2001.

[48] D. R. Swanson, "Migraine and magnesium: eleven neglected connections," *Perspect. Biol. Med.*, vol. 31, no. 4, pp. 526–557, 1988.

[49] M. C. Ganiz, W. M. Pottenger, and C. D. Janneck, "Recent advances in literature based discovery," *J. Am. Soc. Inf. Sci. Technol. JASIST Submitt.*, 2005.

[50] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986.

[51] N. R. Smalheiser and D. R. Swanson, "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses," *Comput. Methods Programs Biomed.*, vol. 57, no. 3, pp. 149–153, Nov. 1998.

[52] M. Yetisgen-Yildiz and W. Pratt, "A new evaluation methodology for literature-based discovery systems," *J. Biomed. Inform.*, vol. 42, no. 4, pp. 633–643, Aug. 2009.

[53] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider, "Context-driven automatic subgraph creation for literature-based discovery," *J. Biomed. Inform.*, vol. 54, pp. 141–157, Apr. 2015.

[54] C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis, "Literature mining, ontologies and information visualization for drug repurposing," *Brief. Bioinform.*, vol. 12, no. 4, pp. 357–368, Jul. 2011.

[55] N. R. Smalheiser and D. R. Swanson, “Indomethacin and Alzheimer’s disease,” *Neurology*, vol. 46, no. 2, p. 583, Feb. 1996.

[56] D. R. Swanson, “Somatomedin C and arginine: implicit connections between mutually isolated literatures,” *Perspect. Biol. Med.*, vol. 33, no. 2, pp. 157–186, 1990.

[57] W. Dong, Y. Liu, W. Zhu, Q. Mou, J. Wang, and Y. Hu, “Simulation of Swanson’s literature-based discovery: anandamide treatment inhibits growth of gastric cancer cells in vitro and in silico,” *PloS One*, vol. 9, no. 6, p. e100436, 2014.

[58] R. Vos *et al.*, “Finding potentially new multimorbidity patterns of psychiatric and somatic diseases: exploring the use of literature-based discovery in primary care research,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 21, no. 1, pp. 139–145, Feb. 2014.

[59] C. M. Miller *et al.*, “A Closed Literature-Based Discovery Technique Finds a Mechanistic Link Between Hypogonadism and Diminished Sleep Quality in Aging Men,” *Sleep*, vol. 35, no. 2, pp. 279–285, Feb. 2012.

[60] I. Petrič and B. Cestnik, “Predicting future discoveries from current scientific literature,” *Methods Mol. Biol. Clifton NJ*, vol. 1159, pp. 159–168, 2014.

[61] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, “Using literature-based discovery to identify disease candidate genes,” *Int. J. Med. Inf.*, vol. 74, no. 2–4, pp. 289–298, Mar. 2005.

[62] J. D. Duke *et al.*, “Literature Based Drug Interaction Prediction with Clinical Assessment Using Electronic Medical Records: Novel Myopathy Associated Drug Interactions,” *PLoS Comput. Biol.*, vol. 8, no. 8, Aug. 2012.

[63] C. B. Ahlers, D. Hristovski, H. Kilicoglu, and T. C. Rindflesch, "Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms," *AMIA. Annu. Symp. Proc.*, vol. 2007, pp. 6–10, 2007.

[64] N. Shang, H. Xu, T. C. Rindflesch, and T. Cohen, "Identifying plausible adverse drug reactions using knowledge extracted from the literature," *J. Biomed. Inform.*, vol. 52, pp. 293–310, Dec. 2014.

[65] R. R. V. Goulart, V. L. Strube de Lima, and C. C. Xavier, "A systematic review of named entity recognition in biomedical texts," *J. Braz. Comput. Soc.*, vol. 17, no. 2, pp. 103–116, Mar. 2011.

[66] B. Vincent, M. Vincent, and C. G. Ferreira, "Making PubMed searching simple: learning to retrieve medical literature through interactive problem solving," *The Oncologist*, vol. 11, no. 3, pp. 243–251, Mar. 2006.

[67] D. Hristovski, J. Stare, B. Peterlin, and S. Dzeroski, "Supporting discovery in medicine by association rule mining in Medline and UMLS," *Stud. Health Technol. Inform.*, vol. 84, no. Pt 2, pp. 1344–1348, 2001.

[68] R. K. Lindsay and M. D. Gordon, "Literature-based discovery by lexical statistics," *J. Am. Soc. Inf. Sci.*, pp. 574–587, 1999.

[69] M. Yetisgen-Yildiz and W. Pratt, "Using statistical and knowledge-based approaches for literature-based discovery," *J. Biomed. Inform.*, vol. 39, no. 6, pp. 600–611, Dec. 2006.

[70] J. D. Wren, "Extending the mutual information measure to rank inferred literature relationships," *BMC Bioinformatics*, vol. 5, no. 1, p. 145, Oct. 2004.

[71] T. Cohen, D. Widdows, R. W. Schvaneveldt, P. Davies, and T. C. Rindflesch, "Discovering discovery patterns with predication-based Semantic Indexing," *J. Biomed. Inform.*, vol. 45, no. 6, pp. 1049–1065, Dec. 2012.

[72] M. D. Gordon and S. Dumais, "Using Latent Semantic Indexing for Literature Based Discovery," *J Am Soc Inf Sci*, vol. 49, no. 8, pp. 674–685, Jun. 1998.

[73] R. J. Cole and P. D. Bruza, "A Bare Bones Approach to Literature-Based Discovery: An Analysis of the Raynaud's/Fish-Oil and Migraine-Magnesium Discoveries in Semantic Space," in *Discovery Science*, A. Hoffmann, H. Motoda, and T. Scheffer, Eds. Springer Berlin Heidelberg, 2005, pp. 84–98.

[74] T. Cohen, R. Schvaneveldt, and D. Widdows, "Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections," *J. Biomed. Inform.*, vol. 43, no. 2, pp. 240–256, Apr. 2010.

[75] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin, "Exploiting Semantic Relations for Literature-Based Discovery," *AMIA. Annu. Symp. Proc.*, vol. 2006, pp. 349–353, 2006.

[76] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. suppl 1, pp. D267–D270, Jan. 2004.

[77] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *J. Biomed. Inform.*, vol. 36, no. 6, pp. 462–477, Dec. 2003.

[78] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "SemMedDB: a PubMed-scale repository of biomedical semantic predications," *Bioinforma. Oxf. Engl.*, vol. 28, no. 23, pp. 3158–3160, Dec. 2012.

[79] M. J. Cairelli, C. M. Miller, M. Fiszman, T. E. Workman, and T. C. Rindflesch, "Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox," *AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp.*, vol. 2013, pp. 164–173, 2013.

[80] M. Rastegar-Mojarad, D. Li, and H. Liu, "Operationalizing Semantic Medline for meeting the information needs at point of care," presented at the AMIA Clinical Research Informatics Summit, 2015.

[81] M. Rastegar-Mojarad, R. Komandur Elayavilli, D. Li, and H. Liu, "Assessing the Need of Discourse-Level Analysis in Identifying Evidences for Drug-Disease Relations in Scientific Literature," presented at the Medinfo, 2015.

[82] M. Rastegar-Mojarad, Z. Ye, D. Wall, N. Murali, and S. Lin, "Collecting and Analyzing Patient Experiences of Health Care From Social Media," *JMIR Res. Protoc.*, vol. 4, no. 3, p. e78, Jul. 2015.

[83] B. W. Chee, R. Berlin, and B. Schatz, "Predicting Adverse Drug Events from Personal Health Messages," *AMIA. Annu. Symp. Proc.*, vol. 2011, pp. 217–226, 2011.

[84] C. C. Freifeld *et al.*, "Digital drug safety surveillance: monitoring pharmaceutical products in twitter," *Drug Saf.*, vol. 37, no. 5, pp. 343–350, May 2014.

[85] M. Rastegar-Mojarad, R. Komandur Elayavilli, Y. Yu, and H. Liu, "Detecting signals in noisy data - can ensemble classifiers help identify adverse drug reaction in tweets?," in *Proceedings of the social media mining shared task*, 2016.

[86] H. Sharif, F. Zaffar, A. Abbasi, and D. Zimbra, "Detecting adverse drug reactions using a sentiment classification framework," 2014.

[87] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "Cadec: A corpus of adverse drug event annotations," *J. Biomed. Inform.*, vol. 55, pp. 73–81, Jun. 2015.

[88] M. Rastegar-Mojarad, H. Liu, and P. Nambisan, "Using Social Media Data to Identify Potential Candidates for Drug Repurposing: A Feasibility Study," *JMIR Res. Protoc.*, vol. 5, no. 2, Jun. 2016.

[89] D. R. Swanson, N. R. Smalheiser, and V. I. Torvik, "Ranking indirect connections in literature-based discovery: The role of Medical Subject HEADINGS (MeSH)," *J AM SOC Inf. SCI TECHNOL*, vol. 57, pp. 1427–1439, 2006.

[90] W. Pratt and M. Yetisgen-Yildiz, "LitLinker: Capturing Connections Across the Biomedical Literature," in *Proceedings of the 2Nd International Conference on Knowledge Capture*, New York, NY, USA, 2003, pp. 105–112.

[91] S. McRoy, M. Rastegar-Mojarad, Y. Wang, K. J. Ruddy, T. C. Haddad, and H. Liu, "Assessing Unmet Information Needs of Breast Cancer Survivors: Exploratory Study of Online Health Forums Using Text Classification and Retrieval," *JMIR Cancer*, vol. 4, no. 1, p. e10, May 2018.

[92] "WebMD Drugs & Medications - Medical information on prescription drugs, vitamins and over-the-counter medicines," *WebMD*. [Online]. Available: http://www.webmd.com/drugs/index-drugs.aspx.

[93] C. Knox *et al.*, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D1035-1041, Jan. 2011.

[94] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Mol. Syst. Biol.*, vol. 6, p. 343, Jan. 2010.

[95] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” *Proc. AMIA Symp.*, pp. 17–21, 2001.

[96] “PatientsLikeMe.” [Online]. Available: https://www.patientslikeme.com/.

[97] D. Marcu and A. Echihabi, *An Unsupervised Approach to Recognizing Discourse Relations*. 2002.

[98] P. W.-H. Johanna D Moore, “Discourse in Computational Linguistics and Artificial Intelligence.”

[99] A. P. Davis *et al.*, “The Comparative Toxicogenomics Database’s 10th year anniversary: update 2015,” *Nucleic Acids Res.*, Oct. 2014.

[100] R. C. Strunk and G. R. Bloomberg, “Omalizumab for Asthma,” *N. Engl. J. Med.*, vol. 354, no. 25, pp. 2689–2695, Jun. 2006.

[101] R. A. Kloner, “Nifedipine in Ischemic Heart Disease,” *Circulation*, vol. 92, no. 5, pp. 1074–1078, Sep. 1995.

[102] M. C. Fernández-Antón Martínez, V. Leis-Dosil, F. Alfageme-Roldán, A. Paravisini, S. Sánchez-Ramón, and R. Suárez Fernández, “Omalizumab for the treatment of atopic dermatitis,” *Actas Dermo-Sifiliográficas*, vol. 103, no. 7, pp. 624–628, Sep. 2012.

[103] C. V. Leier *et al.*, “Nifedipine in congestive heart failure: effects on resting and exercise hemodynamics and regional blood flow,” *Am. Heart J.*, vol. 108, no. 6, pp. 1461–1468, Dec. 1984.

[104] J. R. Diamond, J. Y. Cheung, and L. S. Fang, “Nifedipine-induced renal dysfunction. Alterations in renal hemodynamics,” *Am. J. Med.*, vol. 77, no. 5, pp. 905–909, Nov. 1984.

[105] C. M. Rotella, A. Zaninelli, C. Le Grazie, M. E. Hanson, and G. F. Gensini, "Ezetimibe/simvastatin vs simvastatin in coronary heart disease patients with or without diabetes," *Lipids Health Dis.*, vol. 9, p. 80, Jul. 2010.

[106] E. Nizankowska *et al.*, "Treatment of steroid-dependent bronchial asthma with cyclosporin," *Eur. Respir. J.*, vol. 8, no. 7, pp. 1091–1099, Jul. 1995.

[107] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor Newsl*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[108] M. Rastegar-Mojarad, R. K. Elayavilli, L. Wang, R. Prasad, and H. Liu, "Prioritizing Adverse Drug Reaction and Drug Repositioning Candidates Generated by Literature-Based Discovery," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, New York, NY, USA, 2016, pp. 289–296.

[109] S. J. Hebbring, "The challenges, advantages and future of phenome-wide association studies," *Immunology*, vol. 141, no. 2, pp. 157–165, Feb. 2014.

[110] J. C. Denny *et al.*, "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data," *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1102–1110, Dec. 2013.

[111] Z. Ye *et al.*, "Phenome-wide association studies (PheWASs) for functional variants," *Eur. J. Hum. Genet.*, vol. 23, no. 4, pp. 523–529, Apr. 2015.

[112] N. Novac, "Challenges and opportunities of drug repositioning," *Trends Pharmacol. Sci.*, vol. 34, no. 5, pp. 267–272, May 2013.

[113] P. Sanseau *et al.*, "Use of genome-wide association studies for drug repositioning," *Nat. Biotechnol.*, vol. 30, no. 4, pp. 317–320, Apr. 2012.

[114] R. M. Plenge, E. M. Scolnick, and D. Altshuler, "Validating therapeutic targets through human genetics," *Nat. Rev. Drug Discov.*, vol. 12, no. 8, pp. 581–594, Aug. 2013.

[115] Y. Okada *et al.*, "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature*, vol. 506, no. 7488, pp. 376–381, Feb. 2014.

[116] A. Power, A. C. Berger, and G. S. Ginsburg, "Genomics-enabled drug repositioning and repurposing: insights from an IOM Roundtable activity," *JAMA*, vol. 311, no. 20, pp. 2063–2064, May 2014.

[117] M. C. Cobanoglu, Z. N. Oltvai, D. L. Taylor, and I. Bahar, "BalestraWeb: efficient online evaluation of drug–target interactions," *Bioinformatics*, vol. 31, no. 1, pp. 131–133, Jan. 2015.

[118] M. X. LaBute, X. Zhang, J. Lenderman, B. J. Bennion, S. E. Wong, and F. C. Lightstone, "Adverse Drug Reaction Prediction Using Scores Produced by Large-Scale Drug-Protein Target Docking on High-Performance Computing Machines," *PLOS ONE*, vol. 9, no. 9, p. e106298, Sep. 2014.

[119] X. Liu *et al.*, "In Silico target fishing: addressing a 'Big Data' problem by ligand-based similarity rankings with data fusion," *J. Cheminformatics*, vol. 6, p. 33, Jun. 2014.

[120] "Catalog of Published Genome-Wide Association Studies," *National Human Genome Research Institute (NHGRI)*. [Online]. Available: https://www.genome.gov/26525384/catalog-of-published-genomewide-association-studies/. [Accessed: 18-Nov-2018].

[121] X. Sun, F. Han, J. Yi, N. Hou, and Z. Cao, "The effect of telomerase activity on vascular smooth muscle cell proliferation in type 2 diabetes in vivo and in vitro," *Mol. Med. Rep.*, vol. 7, no. 5, pp. 1636–1640, May 2013.

[122] Z.-Y. Wang and H.-Y. Zhang, “Rational drug repositioning by medical genetics,” *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1080–1082, Dec. 2013.

[123] M. Griffith *et al.*, “DGIdb: mining the druggable genome,” *Nat. Methods*, vol. 10, no. 12, pp. 1209–1210, Dec. 2013.

CURRICULUM VITAE

Majid Rastegar-Mojarad

Place of birth: Bushehr, Iran

Education:

B.Sc., Shahid Beheshti University, Tehran, Iran, August 2005
Major: Computer Engineering

M.Sc., Iran University of Science and Technology, Tehran, Iran (February 2008)
Major: Computer Engineering
Thesis title: New clustering algorithm in relational database and implementation into SQL

M.Sc., University of Wisconsin-Milwaukee (December 2013)
Major: Biomedical Engineering (Informatics track)
Thesis title: Classifying Drug-Drug Interactions with Two-Stage classifier and Rule-based Post-processing