I, Britney A Bogard, hereby submit this original work as part of the requirements for the degree of Master of Science in Computer Science.

It is entitled:
Topic modeling: a novel approach to drug repositioning using metadata

Student's name: Britney A Bogard

This work and its defense approved by:

Committee chair: Anca Ralescu, Ph.D.

Committee member: Anil Jegga, D.V.M., M.Res.

Committee member: Fred Annexstein, Ph.D.
Topic modeling: a novel approach to drug repositioning using metadata

A thesis submitted to the Graduate School of the University of Cincinnati in partial fulfillment of the requirements for the degree of Master of Science in the Department of Computer Science of the College of Engineering and Applied Sciences by

Britney Bogard

B.A. University of Cincinnati October 2014

Committee Chair: Dr. Anca Ralescu Ph.D.
Abstract

The expense of researching and developing new drugs is growing and restricting our opportunities to treat people effectively. Drug repositioning offers a less expensive and more effective opportunity for new treatments of various diseases. Current approaches to drug repositioning generally include assessing similarity between drugs or diseases separately. This study presents the novel idea that comparing scientific meta-data between drugs and illnesses to show similarity in the form of shared scientific concepts can be expanded to effectively introduce new drug-disease indication pairs. This is achieved using topic modeling on data extracted from online disease and drug databases and publication abstracts. Drug-disease pairs that share topics highly were derived from topic modeling results to compare to known indications to assess the accuracy of the approach. Due to limitations that could not be mitigated in the methods used, the results were not conclusive but the high potential of this approach is recognized and suggestions for further research improvements are explained.
Acknowledgements

I would like to thank my advisor Dr. Ralescu for guiding me through this process and helping me find Dr. Jegga for this project. Without your guidance I would not have been confident enough to finish this after finishing my bachelor’s degree. I would also like to thank Dr. Jegga for sharing his passion and ideas with me as well as guidance for finding data in the medical field which I am not so versed in. Thanks to Dr. Annexstein as well for participating in my committee and this process.

In addition, a huge thank you to Dr. Daniels for helping me through this college experience. You were able to turn my college requirements into something that made sense.
# Contents

## 1 Introduction

1.1 Motivation .......................................................... 1  
1.1.1 De novo drug R&D is expensive ................................. 1  
1.1.2 Drug repositioning presents another option ................ 2  
1.2 Objective .......................................................... 3  

## 2 Background

2.1 Problems with de novo drug R&D ................................. 5  
2.2 Strategies for drug repositioning ................................. 6  
2.3 Topic modeling approach ........................................... 7  
  2.3.1 Strategy ...................................................... 7  
  2.3.2 Latent Dirichlet Allocation ................................ 7  
  2.3.3 Meta-data .................................................... 8  
2.4 Potential for results .............................................. 9  

## 3 Methods

3.1 Goal indication pairs ............................................. 12  
3.2 Drug and illness meta-data ....................................... 12  
  3.2.1 Illness and drug databases ................................ 12  
  3.2.2 Publication abstracts ....................................... 14  
3.3 MALLET .......................................................... 16
Chapter 1

Introduction

1.1 Motivation

1.1.1 De novo drug R&D is expensive

The traditional de novo approach to drug development includes finding a target to treat a specific illness as well as research to find a molecule to bind to that particular target. The resources and time it requires to get a newly developed drug from research to market is high and increasing [9]. It currently takes upwards of 17 years to get a drug from the beginning of development to the market [5]. The amount of money required is extremely high and growing as well - currently requiring $5 billion dollars per drug for larger pharmaceutical companies and this is doubling every 9 years [7].

This is a considerable amount of resources and there is a critical need to make this process more economical with the goal of bringing cheaper and potentially safer and more effective drugs to patients faster. This pipeline prevents new, better drugs from reaching the market for diseases that already have some drug available, as well as delays drugs for diseases that don’t yet have a treatment. The cost of developing a new drug also motivates pharmaceutical companies to pursue drugs for markets that have the highest return on investment, often leaving patients with rare diseases without treatment [6]. All patients deserve the best treatment possible regardless of cost. All of these items leads one to conclude that a better alternative to drug development needs to be found.
1.1.2 Drug repositioning presents another option

Bypassing or expediting any part of the drug R&D pipeline would be a welcome improvement. A concept that can be of great value here is drug repositioning, an approach to developing drugs that considers already created drugs for use on other disease indications. Drug repositioning can expedite this process in one of two ways: it can either allow a company to use a drug that has already been developed in the lab, bypassing the expense of development, or it can allow a company to bypass all steps in the pipeline up to the FDA efficacy studies, since currently marketed drugs have already undergone these trials [9]. If the drug has already passed the FDA safety requirements and is considered to be effective for some other use, this reduces the time to market by 10-17 years [5].

Many drugs are researched, developed and successfully complete the safety requirements but are not proven to be effective enough for the illness it was intended to treat: the drug’s indication. 95% of drugs do not pass FDA clinical trials but the R&D cost used to develop a drug does not have to go to waste if a new or alternate use can be found [7]. One of the most popular examples of repositioning includes sildenafil, which was originally intended to treat pulmonary arterial hypertension, and is now also indicated for erectile dysfunction, leprosy and multiple myeloma [9]. Due to the needs in this field many approaches to drug repositioning have been found. Unfortunately, most of the current approaches being researched require often hard to obtain or inaccurate data - such as patented molecular make up of drugs [5]. Knowing this, even though drug repositioning is a promising prospect, much more work is needed to find an effective method which does not depend on questionable data.

One such approach proposed here is using widely available meta-data about drugs and diseases for drug repositioning computation. Using metadata combined with text analysis could be a viable and effective option. Meta-data includes any information surrounding drugs and illnesses - including information databases, drug labels and scientific publications. There have been some promising studies using textual analysis to compare side effect meta-data of drugs [3]; however using the information surrounding diseases and drugs as a whole could be a more comprehensive solution.
1.2 Objective

This study is intended to be an expansion of the topic modeling research done in a previous study on applying topic modeling to FDA drug label information for drug repositioning[3]. In this previous study side effect and box warning data for drugs is used with topic modeling to find similar drugs for reuse. This is closer to assessing similarity through pathology or drug targets since drugs that work on the same targets could produce the same side effects [5]. This approach is rather narrow, focusing on one piece of data surrounding only one side of the equation: drugs.

The research done here is meant to expand on that approach and use a more broad idea of similarity in general. Both the meta-data used and the items being compared are expanded. General scientific meta-data (potentially including side effects, box warnings, etc.) will be used to compare both drugs and illnesses together. This introduces a rather new idea to drug repositioning - using the similarities between drugs and diseases. The intention is to demonstrate that if a drug and disease are discussed in the same scientific framework or share scientific concepts they could also be an indication pair. This would bring out indication pairs that potentially would not be found any other way.

Success of this approach would also mean that repositioning computations could forever improve due to a growing body of information surrounding diseases and drugs. Topic modeling algorithms, in general, improve in accuracy with the amount of information available in the collection. As research continues and the number of publications or information available grows, this would add to the body of information for each illness and drug, thus increasing the accuracy of the method in general and allow it to potentially find new connections and pairs that it did not find before the addition of new information.
Chapter 2

Background
2.1 Problems with de novo drug R&D

As mentioned above, there are many drawbacks to continuing to leverage the de novo drug R&D approach including the following:

- It’s getting more expensive and taking longer to get a drug to market. This results in more expensive drugs for consumers and consumers waiting longer than necessary for needed treatments. It also means that consumers may not be getting the best drug possible due to these long pipelines. This could be due to researchers running out of viable drug molecules and/or increased regulations by the FDA.

- Due to this cost, drugs with smaller markets will not be researched and developed due to a low ROI [11]. A low ROI often prevents people with rare or orphan diseases from getting treatment [6]. Pharmaceutical companies simply do not want to invest the resources into a product that will not return their investments.

These downsides, combined with new computational approaches made possible by the internet, are causing pharmaceutical companies to take a different perspective. In 2009, 30% of newly marketed drugs were previously marketed with different indications - drugs that were repositioned in the market [8]. Unfortunately, right now pharmaceutical companies primarily rely on serendipity to find these new drug indications, which includes finding prescriptions that are being written for drugs for unintended uses. Finding drugs that are being used for different indications in clinical settings is a slow and rather ineffective process [9], which will not easily allow for the growth the drug market needs to keep up with demand and patients.

A computational drug repositioning approach presents the potential for much lower costs and time to bring a drug to market. Lower cost means that a pharmaceutical company will be more inclined to serve the smaller markets and will be able to get more drugs to market faster. Newly found drugs could be found for illnesses that are more effective and safe (less side effects) and less expensive than current treatments [9]. Therefore, a systematic, less expensive solution to drug repositioning would benefit currently treated patients and those waiting for a treatment.
2.2 Strategies for drug repositioning

The two main strategies to repositioning *in silico* are disease-based or drug-based. The disease based approach focuses on finding similar diseases that could use the same drug. Thus, through some assessment of a group of illnesses, a few may be found to be extremely close to each other. In this case a drug that is effective in treating one of the similar illnesses may also be effective in treating the other. The drug-based approach focuses on finding similar drugs that could share indications. For example if two drugs were found to be similar through some assessment, perhaps an illness that one of the drugs treats could also be treated by the other. The similarity between two drugs or two illnesses can be assessed through analysis molecularly, chemically or any other data concerning the drugs or illnesses that could be considered significant.

To assess drug similarity the most simple approach is comparing indications. If two drugs share multiple indications, chances are they operate on similar targets and could share even more indications [5]. This approach does not employ much data nor is very good for finding new indications with drugs that are already rather promiscuous. Chances are higher for finding new indications if more data is involved in the computations.

Another method is using *molecular activity* similarity. This consists of comparing molecular profiles - a measurement of the molecular activity of a compound in the body [5]. These molecular activity profiles can be derived from online databases. This approach can be used to compared drugs to other drugs or illnesses. The biggest problem with this approach is the accuracy of molecular activity profiles. Finding activity profiles is not an easy process and can require many potentially inaccurate assumptions [5]. The chemical properties of drugs are also used to compare drugs but this information is often restricted or patented by companies and unavailable.

There are other strategies such as *molecular docking* that do not use similarity analysis but include finding new targets for existing drugs using data such as 3D molecular models. These center around finding new protein targets that the drug molecule will bind to. This can be done in a lab, using 3D models of the drug and protein molecules or any other way that allows one to determine if they will bind [5].
2.3 Topic modeling approach

Topic modeling is a method for discovering the topics distributed over a collection of documents. In this study each document contained meta data about one illness or drug and the collection contained multiple of these, some drugs being treatments for some illnesses in the group. One of the more recent algorithms topic modeling uses is Latent Dirichlet Allocation (LDA), which was used here. LDA assesses documents, generates the topics that create those documents, then expresses the distribution of the topics among those documents.

2.3.1 Strategy

Through the use of LDA this study uses a combined drug-based and illness-based strategy. Instead of assessing the similarities between different drugs or different illnesses, similarity was primarily assessed between drug disease pairs. The meta-data of each disease or drug was assessed for its topic composition then connections were found where topics were highly similar. This method could be extended to be used as a mainly drug-based or disease-based approach as well if it proved to be useful. This differs from the approaches above in that it’s assessing similarity between two very different types of things, diseases and treatments for those diseases. How these can share similarities will be discussed below.

The goal for this study and benchmark for its success is its ability and reliability in recalling already known indication pairs, or drug illness pairs in which the drug is already known to treat the illness. If the algorithm is able to recall known pairs to some considerable extent it will be considered a potential solution to the problem of drug repositioning. It should be considered that our benchmark list of known pairs does not include future indication pairs that have not been tested and approved yet. Thus, a pair found in this nature may be considered a false negative, but there is currently no way to measure these.

2.3.2 Latent Dirichlet Allocation

*Latent Dirichlet Allocation* is the algorithm that will be used for topic modeling. It is a generative probabilistic model for a collection of data [4]. It is a hierarchical Bayesian model based on the idea
that documents are mixtures over topics where each topic is a distribution over words [4]. Topics contain words and these topics are distributed over the documents in the collection. This algorithm generates topics that documents share, by what percentage and what words those documents consist of. In the figure above representing the LDA statistical model, \( \alpha \) represents the Dirichlet prior for the document topic distribution, \( \beta \) represents the Dirichlet prior for the word distribution through topics. \( \theta \) is the topic distribution over a document, \( z \) is the topic for a specific word in the document and \( w \) is a specific word. Plate M represents the collection of documents and plane N represents the topics, containing words, distributed through the documents [4]. From a higher perspective, if multiple documents contained both words dog and bone, dog and bone would be in a topic together and those documents containing the words would be considered to contain that topic. In this case, where metadata is being used, ideally words used for a specific scientific concept will be grouped together and those documents containing those words will be highly attributable to that topic. For example, a document that contains metadata on asthma would hopefully be highly attributable to a topic that contained words such as lung and breath.

**2.3.3 Meta-data**

Many parts of meta-data could be used surrounding illnesses and drugs. The meta-data includes the FDA study results, side effects, boxed warnings, symptoms and anything else describing the illness or drug in one way or another. For drug repositioning the most commonly used sections here are the drug side effects since they can be reflections of the drug’s targets that we may not be aware of [5]. This side effect analysis approach was used successfully in the previous similar studies using
topic modeling on FDA drug labels [3][2]. In the first of these studies topic modeling was effective in categorizing the drugs by therapeutic application and adverse effects [2]. The second continued this experimentation to effectively pair drugs with common indications, showing potential for a drug repositioning application [3]. These linked drugs together through common side effects but it has not yet been attempted to expand the text analysis of meta-data to using more information than this. If drugs and illnesses could be linked together through shared concepts or trends in research documentation about them this broadens the approach.

Generalizing the data used for analysis also presents more opportunities for growth and improvement since this meta-data could be considered limitless while the body of medical knowledge continues to grow. Also, concerning using publication data with topic modeling analysis, linking documents of this nature together for the purpose of research has been proven effective. Alterations and improvements can be made to the LDA algorithm to capitalize on this type of data but in general LDA has been proven useful for categorizing scientific research [1]. This study was meant to be a proof of concept so no alterations were made to the LDA algorithm for this purpose but results could be improved if, for example, more depth was added to the underlying Bayesian network such as adding authors, universities or side effects.

2.4 Potential for results

The conclusions that could be drawn from results of this study are much more broad than the previous studies discussed. The previous studies can link drug pairs and potentially illness pairs but would not be able to create a network or pairs linking both drugs and illnesses together. The data used is rather specific to drugs or illnesses but not both. These previous approaches are also limited by the availability of a specific type of data (side effects and box warnings). Being able to generalize the data to something that shares similarities between drugs and illnesses in addition to illness-illness and drug-drug connections creates the potential for a comprehensive network. A network of this nature presents drug-illness pairs for potential drug repositioning, the goal of this study, but it also allows graph analysis concepts to come into play. Being able to study drugs or illnesses that demonstrate core graph concepts, such as hubs, centrality and bridges between
clusters would show where items fit in the larger scheme of medical research [9].

On a higher level than leveraging drug-illness similarities, showing that drugs and illnesses that share topics also share indications could present a considerable amount of potential in the field of finding drug indication pairs.
Chapter 3

Methods
Table 3.1: Subset of known indication pairs

<table>
<thead>
<tr>
<th>Disease/Drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immune thrombocytopenia</td>
</tr>
<tr>
<td>Type I diabetes mellitus</td>
</tr>
<tr>
<td>Non-small cell lung cancer</td>
</tr>
<tr>
<td>Nocardiosis</td>
</tr>
<tr>
<td>Glanders</td>
</tr>
<tr>
<td>Cholangiocarcinoma</td>
</tr>
<tr>
<td>Gastric cancer</td>
</tr>
<tr>
<td>Pertussis Tetrazycline</td>
</tr>
<tr>
<td>Measles</td>
</tr>
<tr>
<td>Type II diabetes mellitus</td>
</tr>
<tr>
<td>Rituximab</td>
</tr>
<tr>
<td>Alrestatin sodium</td>
</tr>
<tr>
<td>Gemcitabine hydrochloride</td>
</tr>
<tr>
<td>Imipenem</td>
</tr>
<tr>
<td>Imipenem</td>
</tr>
<tr>
<td>Fluorouracil sodium salt</td>
</tr>
<tr>
<td>Cytarabine</td>
</tr>
<tr>
<td>hydrochloride</td>
</tr>
<tr>
<td>Human normal immunoglobulin</td>
</tr>
<tr>
<td>Insulin lispro</td>
</tr>
</tbody>
</table>

Figure 3.1: Subset of known indication pairs

3.1 Goal indication pairs

This study centered on a list of known indication pairs (i.e., disease-drug pairs) from Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/). This list was the target list for the topic modeling results - getting these pairs from the results with a high amount of certainty would mean the topic modeling was a success. It was attempted to create a more comprehensive list of known indications from the DailyMed drug database but, as discussed in the previous topic modeling study [3], the indications data in the drug entries are not standardized enough to mine this information accurately. The indications list from KEGG was determined to be long enough and sufficiently accurate to be a goal of the current study. Various subsets of these indication pairs were used for different experiments, using databases the entire list was used and for the abstracts a random subset or a subset from a pathway was used.

3.2 Drug and illness meta-data

3.2.1 Illness and drug databases

The first phase of this study was done with drug data derived from DailyMed (http://dailymed.nlm.nih.gov/) and illness data from KEGG and Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov). DailyMed is a database of the most up to date FDA documentation for currently marketed drugs. For each one of the drugs in the indications list that were on DailyMed, the description section was mined and exported to a document (one per drug). From KEGG the illness descriptions were
also used and exported to documents. OMIM was slightly different from KEGG in that it only had information regarding diseases with a genetic component but the KEGG entries had OMIM IDs that corresponded to illnesses in the indications list assisting in the matching of data. In the OMIM entries the clinical features section was used rather than the description sections (this was a similar type of data but with a different title). All of this data was mined from exports from the websites, varying amounts of effort was needed to extract the needed data but since the resulting data was rather small in size, all of the illnesses and drugs from the known indications list could be used in the same topic modeling computation.

There were two experiments done with this data - one using the illness data from KEGG and one using the illness data from OMIM. For both, the drug data used was from DailyMed. If drugs or illnesses from the indications list could not be found they were simply discarded from this experiment. 500 drugs from the original 1066 in the list were found and used from DailyMed, 179 illnesses were found and used from OMIM and 352 from the KEGG database.

ENTRY H00079 Disease

NAME Asthma

DESCRIPTION Asthma is a complex syndrome with many clinical phenotypes in both adults and children. Its major characteristics include a variable degree of airflow obstruction, bronchial hyperresponsiveness, and airway inflammation. Inhaled allergens encounter antigen presenting cells (APC) that line the airway. Upon recognition of the antigen and activation by APC, naive T cells differentiate into TH2 cells. Activated TH2 stimulate the formation of IgE by B cells. IgE molecules bind to IgE receptors located on mast cells. The crosslinking of mast-cell-bound IgE by allergens leads to the release of biologically active mediators (histamine, leukotrienes) by means of degranulation and, so, to the immediate symptoms of allergy. Mast cells also release chemotactic factors that contribute to the recruitment of inflammatory cells, particularly eosinophils, whose proliferation and differentiation from bone marrow progenitors is promoted by IL-5. The activation of eosinophils leads to release of toxic granules and oxygen free radicals that lead to tissue damage and promote the development of chronic inflammation.
Example of a KEGG illness description

It is significant to note that these data sources tended to have rather small documents per drug or illness; each of the items may have had one or two paragraphs. Most of the illness files from KEGG were around 1 KB. The OMIM resource was relatively more extensive with document size ranging from 5-25 KB each. The drug documents were comparable in size to the OMIM descriptions with files mostly between 5 and 20 KB. Consideration should also be brought to the nature of this metadata - the DailyMed drug data is geared toward providing information to patients and doctors about how the drug works while the illness documents tended to be a broad overview of the illness. This is in contrast to the publication abstracts used in the next section of the study.

3.2.2 Publication abstracts

In the research done using publication abstracts only a subset of the KEGG indications list was used due to data size and time restrictions. If the entire indications list was used it would have taken an extensive amount of time to collect all of the abstracts needed - each illness or drug requiring a call to a web server. These subsets of the indications were first selected various times in a random manner only selecting drugs that treat multiple illnesses and illnesses that have multiple drugs. This was done because it was proven in a previous study that grouping similar drugs for reuse using topic modeling was more effective when the drugs were narrowed to only drugs with multiple indicated uses [3]. Since a random grouping of drugs and illnesses was not the most effective way to group subsets (some samplings could include more similar items, some not) an average was taken over many samplings of the pairs. To find the average the entire method was ran 20-30 times, each time with a different random subset of items.

Another test was done using select drug and illness pairs that were included together in one pathway, as determined by the KEGG database. It was thought that this would group similar items together more effectively than a random selection. Pathways were selected from KEGG that had a high amount of items in them, including both drugs and illnesses. Known indication pairs that had one item (either the drug or the illness) in the pathway were selected and used in that collection. It was reasoned that known disease-drug pairs that related to a same pathway would
have a higher likelihood of connecting via shared topics or concepts. In other words, if a disease and drug impact a common pathway the chances of that drug and disease relating could be high. This would be manifested either as a side-effect of the drug or as a potential drug repositioning candidate or could be simply a previously known or approved indication.

To get the publication abstracts, PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) was used in combination with Entrez - a tool allowing users to query the PubMed database. A Python package called Bio (http://biopython.org/) was used which automated the process of using Entrez. Through Bio and Entrez there is the ability to query publications from PubMed and get any sections you wish - in this case the illness or drug name was used as the query search string and ten abstracts of papers in the search results were used. The ten abstracts were then compiled into one document for each illness or drug. This was done for each item in the selected subset of known indications. Initially a list of only twenty indication pairs were used for the topic modeling and that was expanded to thirty and fifty to determine if this affected the results. In general the results did not largely depend on the number of indication pairs used, but rather using the appropriate number of topics for the topic modeling algorithm, various numbers of topics were used for each try to find the optimal number for this as well.

For each illness-drug indication pair for which abstracts were gathered, abstracts that were shared between the two items were not used in the data. There were many publications gathered for the known indication pairs in which the research included both the drug and illness names. These names inside of the documents would link the two items together automatically in the topic modeling algorithm. It was discovered during research that a publication may not be returned in both queries but both the illness and drug terms were used in the abstract. Further measures had to be taken to remove these abstracts from the data as to not skew results. This resulted in retrieving a higher amount of abstracts (12-15) then pruning those that had drug or illness names in them.

Compiling the abstracts for these search terms resulted in much larger files in the collection, most being 20-30 KB in size. In general, this was an increase in the information amount for both illnesses and drugs. It was expected for these to have better results using this method due to
document size and a lower number of drugs and illnesses in the collection. This metadata took a different tone than the data used from the databases since it was describing research done in and around the topic. It was expected to have a much higher amount and variety of medical vernacular that may improve the results. The amount of data could also be easily increased to potentially improve the algorithm if a higher number of abstracts were used per document.

3.3 MALLET

MALLET, an open source topic modeling tool [10], was used for this study. It was created by Andrew McCallum at the University of Massachusetts Amherst and is used by many for finding trends in text data. It uses Latent Dirichlet Allocation to find these topic trends in the data. MALLET outputs the keywords for the topics found as well as what topics are contained in each document and to what percentage. Input for the package consists of any additional stopwords (if necessary), the number of topics to find and the collection of documents to be used for the LDA.

3.3.1 MALLET input

As expected, the main input for mallet is the collection of documents explained above. In addition to this MALLET requires a stopword list and a number indicating the number of topics to generate. The stopword list indicates to the algorithm which words it should skip during computations. These words are essentially filler that would not add meaning to a topic if they were selected. MALLET already includes a list of stopwords (words like 'the' and 'and') but more were needed for this context of data. General medical terms that were common in these documents were added to the list such as patient, blood and diagnosis. Initially outputted topics from MALLET were used to create these additional stopwords lists. The KEGG, OMIM and DailyMed datasets were used with MALLET and from the topics generated all the words with insignificant meanings were added to the additional stopword list for future use of the algorithm. This was done again with the publication abstract dataset since the vernacular used in the metadata was different and words regarding publication details needed to be removed (cities, university names, etc).

The other input needed for MALLET is the number of topics the algorithm would generate.
This number was experimented with throughout the duration of this study. The previous study on FDA labels and repositioning determined fifty two topics to be the idea number to represent all of the medical topics needed to represent major categories of drugs [3]. This number was not used as an accepted amount of topics in this study because different amounts of drugs and illnesses were used, not necessarily including the same subset of types that the previous study used. The previous study used eight hundred and seventy drugs but the trials done here were anywhere from 30 drugs and illnesses to 852. With such a varying amount of items the number of concepts expressed in them were expected to change as well. For this reason experimentation was done to determine the best values for the topic variable for the number of items we used.

### 3.3.2 MALLET output

The output from MALLET used in this study primarily was the composition file. This file contained one line for every document and each of those lines lists number pairs pertaining to the topic numbers and how much of that document consists of that specific topic.

Thus if the algorithm generated 30, after the file name there would be 30 pairs of numbers. For
each pair, the topic number then some decimal. In the example above the first pair of numbers states that the document is approximately 47% created from topic 6. This data was used to find all of the indication pairs after using mallet - connections were made between items when they both showed a high level of composition in the same topic.

The other piece of output used from MALLET was the export of the topic keys. This showed the top words the composed the topics generated from MALLET. The topic keys were only used for a frame of reference for the items being linked together and for the creation of the stop word documents. The topic words outputted from MALLET seemed to be random in relevance. Some topics created made sense conceptually and others contained words that were not related. This observation did not change based on the relative success of recalling the known indications or on the data sources being used - publications or databases. This file was often useful in determining if results of a trial were ineffective - for example if there was a drug or illness name that was high on the keys list. Often a drug or illness name in the key meant that the name was the main reason for the linking between items.

### 3.4 Deriving Results

The composition file explained above was used to derive results from MALLET. Drug-illness pairs were created when both the drug and the illness were highly composed of the same topic. Thus if a drug and an illness both had a high decimal for topic number 6, a connection would be made between them. The threshold for this was arbitrary at the beginning then research was done to determine the most effective combination between the threshold and the number of topics. The effectiveness of a threshold value was directly related to the number of topics. If there was a lower
number of topics selected, the connection values tended to be higher in general and a low threshold value would retrieve too many topics where they could not all possibly be new potential indications. Otherwise if the number of topics was higher too high of a threshold value would result in no pairs at all. Optimal values were ultimately found for both to be used in the results of this study and will be explained.

After the top pairs were derived from the composition file these pairs were compared to the known indications list. For each group of pairs found from the MALLET output an accuracy percentage was found - the number of correct indication pairs found divided by the total number of indication pairs found. This percentage represents the true positive rate, the pairs retrieved from the results that are also known to be correct. The inverse of this percentage is the false positive amount, the pairs found that are not known pairs. It should be considered that the false positive percentage is only reflected in our current understanding of indications, some of the found pairs may prove to be correct in the future.
Chapter 4

Results
4.1 Variables

There was an extensive amount of variables to consider when using topic modeling for drug repositioning. The opportunities for how to do it and the data to use are endless. In this study an educated guess was made on how to most effectively use the metadata to find new repositioning opportunities. Experiments were done to find the most effective way to narrow the threshold and topic numbers specifically, which helped with the resulting averages derived.

4.1.1 Connection threshold

![Figure 4.1: Accuracy of algorithm as connection threshold varies](image)

To find optimal threshold the number of pairs was held constant at 30 and the accuracy of the algorithm was graphed at multiple threshold values. While the best threshold value varied with the number of topics, the optimal number of topics for 30 random indication pairs tended around 30 or 35, so 30 topics were used to calculate the generally best threshold value.

4.1.2 Number of Topics

To find the optimal number of topics to use the accuracy of the algorithm was graphed at multiple topic values. This also needed to take into consideration the threshold value so the optimal threshold value of .17 was used for this computation. The same 30 random indication pairs were used as described earlier.
4.2 Using KEGG and OMIM

When using the KEGG illness descriptions combined with the DailyMed drug clinical pharmacology sections the resulting pairs were close to random with 193 out of the 2248 pairs found (using 100 as the number of topics) being in the original indications list, a 8.5% accuracy rate. These computations were not done with the threshold and topic values listed above. Since the number of items used did not vary, the algorithm was simply attempted multiple times with different threshold and topic values until the best accuracy rate was found. This was not a promising result. This may have happened due to the incompleteness of the KEGG illness descriptions - they are short and less medically complicated. This was improved upon in the next trial by using the OMIM database instead of KEGG and continuing to use the DailyMed drug meta-data. These were longer more comprehensive documents and it was thought that these would have better results, but a success rate of around 7.6% was found (varied by 1-2% depending on threshold used), this was the highest recall rate found with this data using 50 topics for the MALLET input as suggested by the previous topic modeling study using a similar number of items. Since it is known that more data increases the accuracy of topic modeling, larger data sources were tried next.

4.3 Using PubMed abstracts

At first the results for using the PubMed abstracts seemed extremely promising - 50%+ positive pair retrieval rate, but it was discovered that this was due to research abstracts having other illness
or drug names in them. Once this error was discovered and corrected, the accuracy of the method dropped. The first attempts at finding an average recall rate were around 20%, using various threshold values and numbers of topics. This value was generally true for experiments using 20 indication pairs in the data pool, for various thresholds between .08 and .17 and topic amount between 20 and 35.

Once the research surrounding the threshold and topic values was complete better results were found using the 30 indication pairs and 35 topics used in those trials. The threshold value found above (.17) wasn’t as useful since it resulted in less than desired numbers of found indications (1-5) so a threshold of 11 was found to be optimal to find around 30 pairs per run. Once these variables were used a recall rate of 35% was found among the new indications with a standard deviation of 21%. This can be contrasted with a random recall rate of 7.5%. While some of the results in this average were highly accurate with the number of found pairs close to 30 and correct pairs as high as 50%, other runs only retrieved 1-5 pairs and even others retrieved over 70 indication pairs with an accuracy as low as 8%. One could deduce that some distributions of the random pairs selected for these runs were better or more similar to each other while others were extremely unsimilar only being able to create 4 pairs, or there is some aspect of the data that is skewing the results, as explored below.

4.3.1 Example of results

In examining specific results of an iteration of the algorithm some trends were found that seemed to be skewing the data. These were analyzed on a case by case basis and addressed when possible (such as drug and illness names in other item’s abstracts) but some did not seem to have solutions. One example is the indication pairs above - rabies and the rabies vaccine are obviously matched together due to shared terms, the drug Docetaxel seems to show up for every item and only three correct pairings were found.

Upon further investigation into why shortcomings like this happen the topics that linked these items together were investigated. While Docetaxel seemed to be related to everything it was composed most highly by topic 28 as shown in Figure 4.4 - which doesn’t seem necessarily out of
Amoebiasis Doxycycline, Amoebiasis Etoposide, Diphtheria Docetaxel, Diphtheria Etoposide, Diphtheria Rabies vaccine, Febrile seizures Docetaxel, Febrile seizures Etoposide, Febrile seizures Rabies vaccine, Propionic acidemia Azathioprine sodium, Propionic acidemia Carglumic acid, Propionic acidemia Doxycycline, Propionic acidemia Ifosfamide, Propionic acidemia Rabies vaccine, Prostate cancer Docetaxel, Rabies Docetaxel, Rabies Rabies vaccine, Renal cell carcinoma Docetaxel, Rheumatoid arthritis Docetaxel, Rheumatoid arthritis Etoposide, Rheumatoid arthritis Rabies vaccine, Type II diabetes mellitus Docetaxel, Type II diabetes mellitus Etoposide, Type II diabetes mellitus Rabies vaccine

Figure 4.3: A set of found indication pairs for one iteration of the algorithm

23 1.42857 rabies vaccine dog vaccination virus baits veterinary vaccines rvg owners rig sag rpep exposure antibodies drit cattle nuclease attitude 27 1.42857 cancer prostate docetaxel ci giai hungary onkol pfs budapest int tumors enzalutamide abiraterone oncology radiation lung cabazitaxel resistant zet 28 1.42857 sep electronic elsevier received serum res induction receptor status follow mediated retrospective male assay ac adult isolated pediatrics improve 29 1.42857 col gene mutations mutation collagen osteoarthritis cartilage mild alpha mutated transgenic genes fibrils cys skeletal chondrodysplasia mutant substitution oa

Figure 4.4: Some topics of an iteration of the algorithm - each being a set of keywords
the ordinary though made up of rather general terms. Thus without being able to pinpoint the reasoning for inconsistencies like this many were unable to be addressed.

4.4 Narrowing by Pathway

Narrowing the items by pathway did not prove to help the randomness and deviation presented in calculating the averages above. Pathways hsa00140 and hsa04514 were selected from the KEGG database since they had some of the highest numbers of drugs and illnesses. Unfortunately, even though these contained the highest numbers, they were still rather low containing about 23 and 9 drugs and illnesses per pathway. These drugs, illnesses and their indications were used in the data sets to compute average accuracy here. One average was computing for each pathway. Results with these were at a 0% accuracy rate, even worse than the random subsets.
Chapter 5

Conclusions

5.1 Conclusion

This study presents a novel approach to the problem of drug repositioning based on information extracted from online databases and publication abstracts using topic modeling techniques to assess similarity.

The principal idea pursued here is to enable the use of drugs (already approved) for diseases for which they were not initially indicated.

Departing from previous studies, where drug-to-drug similarity or disease-to-disease similarity was used, here the drug-to-illness similarity is considered. Chemical and/or molecular similarity was exploited in previous studies. By contrast, drug-to-illness similarity is evaluated here based on shared topics in scientific publications. This meta-data based approach, which would convey the fact that scientifically, an illness and drug can be discussed in the same framework, may capture similarities otherwise not present on the basis of chemical or other types of analysis.

The preliminary results reported here suggest that a more comprehensive, standardized pool of data and a more specific topic modeling approach may present more promising results. Multiple problems arose that could not be mitigated using the methods in this study. One major one of these is unintended shared terms skewing data; for example, an illness name appearing in a drug name or the drug and illness sharing some specific term that is specifically indicative of each other.
These instances caused a close similarity to be deduced erroneously: not through shared scientific concepts but already developed names. There was also the problem that topic modeling only finds significance in individual words while much of the vernacular in the medical field requires multiple words for a specific meaning.

A 35% recall rate with the considerably high standard deviation of 21% is not compelling enough to say that this approach works with the methods used in this study. It does however imply that these methods can be improved upon significantly to potentially prove the success of this approach in future research.

There still stands the potential implication of this study that suggests sharing scientific concepts implies sharing protein targets for drugs, but a word by word general topic modeling approach may not be the way to demonstrate this effectively. An algorithm that is able to find significance in multiple word concepts and effectively deal with erroneous term connections skewing data would show more confidence in this approach. The methods that were used here were not conclusive to prove or disprove the implication but show the necessity to investigate this potential further with a more targeted way to show shared concepts.

If a more effective technique was found to improve upon this and find a way to connect protein targets to scientific concepts, there is the potential for an extremely economical and efficient way of finding new drug-illness indications to offer cheaper and safer alternatives to patients (or a new treatment in general). It is more worthwhile in the long run as our body of knowledge in the medical field continues to grow surrounding current drugs and illnesses as well as new ones. The algorithm would become more effective without extra effort as new information becomes available.

Improvements of the current results may be obtained from improvements to the topic modeling approach. Some suggested directions:

- Either improvement in finding an effective distribution of drugs and diseases for the execution of this algorithm or the ability to include all drugs and diseases in one execution. Choosing subsets randomly as was done in this study has the potential to retrieve and use a group of drugs and illnesses that are potentially highly dissimilar. Alternatively, considering all of the drugs and all of the illnesses could provide a more meaningful result.
• Implementation of a topic modeling algorithm that is able to take multiple word concepts into consideration. In these scientific documents often individual words do not hold a significance comparable to multiple word phrases. Being able to link together multiple word terms between papers would better model the sharing of scientific concepts rather than the sharing of specific words. Potentially a solution to this problem would also be able to address the erroneous terms that were skewing results.
Bibliography


