Case Western Reserve University

Department of Electrical
Engineering and Computer Science

Building a Graph of Drug Connections by Mining ClinicalTrials.gov Data

Final project for EECS 435: Data Mining

Authors: Yigit Kucuk (yxk368) - Andrew Morrison (asm115)

April 21, 2016

Table of Contents

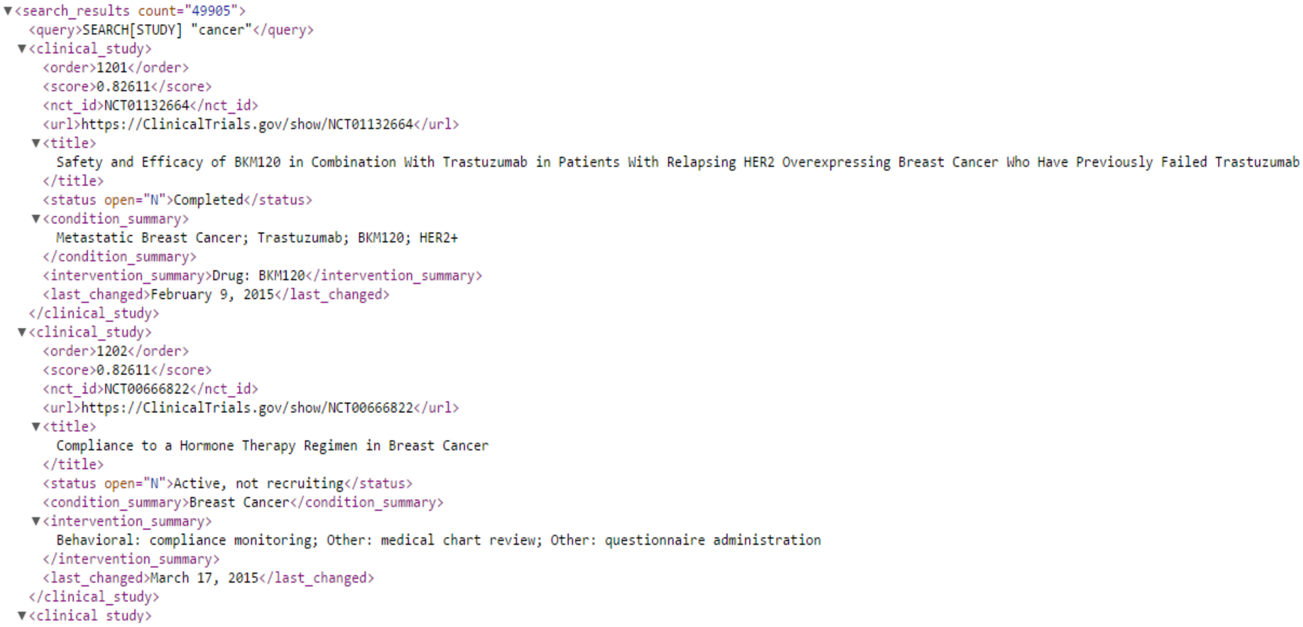
Table of Figures

1. INTRODUCTION

Current medical treatments and drugs are not able to treat all patients equally effectively. The development of drugs is always subject to improvement as drugs become more accessible for people internationally every day. Bioinformatics, one of the most challenging topics of biology and genetics, has created many opportunities and methods for genes, proteins and their interactions to be researched understood. As bioinformatics methods have become well-established, being able to personalize medicine and help people to a greater extent has become a priority in medicine and bioinformatics research. Any vast project such as the Human Genome Project [1] and 23andme.com [2] is, at its core, focused on personalizing medicine and treatments. Despite the considerable progress in fields such as disease gene discovery and protein-protein interactions and networks, personalized medicine is still far from being standard. Drugs are still primarily developed and tested with clinical trials often conducted by medical doctors and representatives from pharmaceutical companies [3]. ClinicalTrials.gov is such a registry of human clinical research studies. It is hosted by the National Library of Medicine (NLM) at the National Institutes of Health (NIH). ClinicalTrials.gov being such a valuable, accurate, and reliable source of drugs, conditions and adverse events, we hypothesize that the many features contained in the clinical trials can be used to create a drug repurposing and repositioning network that can be used for further analysis to help personalize medicine and to validate drug networks created such as DrugBank [5]. Clinical trials are yet to be researched in bulk by many drug companies and researchers. Our methodology involved downloading and analyzing large numbers of trials from ClinicalTrials.gov, becoming familiar with the format of the trials and its shortcomings, and creating a database of the drugs, treated conditions, and adverse events stored within the trials. From the database, we formulated a versatile methodology for drawing connections between drugs that span across multiple studies, thus resulting in a graph or “network” of drugs. We also ran a frequent pattern mining algorithm and compared the results with the proposed graph. Creating the graph and validating it with the FP mining results, we conclude that our graph is more informative than frequent pattern mining alone, and that our methodology allows for building drug graphs that are more versatile and informative than previous attempts to mine connections between drugs using clinical trials data. This idea and project is significant in that it is among the most unique and general networks derived from ClinicalTrials.gov to the best of our knowledge.

2. BACKGROUND AND RELATED WORK

ClinicalTrials.gov consisted, as of 2/16 (The date the project assigned), of 208,568 studies with results available to public from over 170 countries. [3] Clinical trials may have up to 32 fields with a variety of information that has mostly non-unified English explanations such as how many patients were involved, how the treatment was conducted, which drugs and dosages were involved, adverse events and treated conditions, ages and nationalities of the participants etc. Although it is mandated by the federal law [4] to report clinical trials to this source, the language used in the trials is often informal and usually more of a combination of reports than a database. Even though significant effort [6] has been shown to unify the language and medical terms used in the trials, it is still an ongoing process. For example, often times we see terms like "Aspirin mg-100" that corresponds to a certain dosage of Aspirin used for that particular stage of the study, leading to complications in making connections with trials who simply listed the drug as "Aspirin" without dosage information. Although certain articles are very explanatory of how the study is conducted, such information is irrelevant and usually a setback for the purposes of our project. (Figure 1) Clinical trials can be either 'completed' or 'in-complete' status. In addition, they either have results or they do not have results, denoted by 'has results' indicator for each status and in their corresponding xml elements. For the purposes of our study, we choose to only use interventional studies with 'has results' and 'completed' to build our graph and run the mining algorithms. These restrictions amounted to about 19,000 trials for us to mine.



```
▼<search_results count="49905">
   <query>SEARCH[STUDY] "cancer"</query>
  ▼<clinical_study>
     <order>1201</order>
     <score>0.82611</score>
     <nct_id>NCT01132664</nct_id>
     <url>https://ClinicalTrials.gov/show/NCT01132664</url>
    ▼<title>
       Safety and Efficacy of BKM120 in Combination With Trastuzumab in Patients With Relapsing HER2 Overexpressing Breast Cancer Who Have Previously Failed Trastuzumab
     </title>
     <status open="N">Completed</status>
    ▼<condition_summary>
       Metastatic Breast Cancer; Trastuzumab; BKM120; HER2+
     </condition_summary>
     <intervention_summary>Drug: BKM120</intervention_summary>
     <last_changed>February 9, 2015</last_changed>
   </clinical_study>
  ▼<clinical_study>
     <order>1202</order>
     <score>0.82611</score>
     <nct_id>NCT00666822</nct_id>
     <url>https://ClinicalTrials.gov/show/NCT00666822</url>
    ▼<title>
       Compliance to a Hormone Therapy Regimen in Breast Cancer
     </title>
     <status open="N">Active, not recruiting</status>
     <condition_summary>Breast Cancer</condition_summary>
    ▼<intervention_summary>
       Behavioral: compliance monitoring; Other: medical chart review; Other: questionnaire administration
     </intervention_summary>
     <last_changed>March 17, 2015</last_changed>
   </clinical_study>
  ▼<clinical study>
```

Figure 1: Overview of Clinical Trial Results XML View

Our project was only concerned with the fields that contain information on treated conditions, adverse events, interventions (often drugs), and phase information. Phase can have a value between 1 and 4 inclusive, and denotes the confidence level of the study that has passed inspections and proven to be effective in an ever growing number of patients. [7] The partial information about how the drugs are launched and the steps (Phases) of the development is given in Fig. 2.

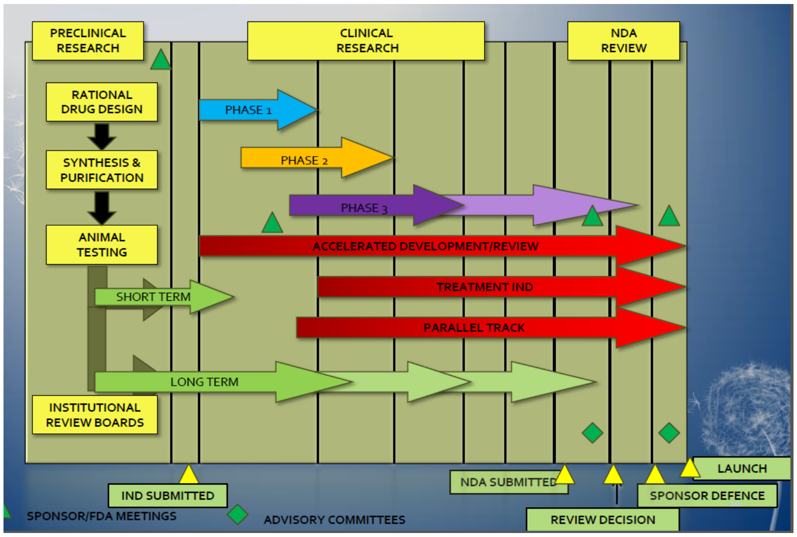


Figure 2: Simple Drug Launch Process and Phases

Many attempts have been made to create a database of clinical trials and adverse events and conditions [9-12] many being significant and sufficient for many of the research or supplementary purposes. Despite the existence of such efforts, none of them were primarily concerned with the drugs, especially and specifically combinational drugs which means more than one drug per intervention field. Although [12] is a database of adverse events, it is still insufficient for the drug-centric analysis that we propose. Xu et.al [9] created an adverse mined database with significant data mining effort, but their study was still insufficient for building a graph from the interventions (particularly drugs), as suggested in the future work part of the study. Our study is unique in the manner of creating a graph solely based on the mined data from the drugs used in combination of others. Our results show that this approach is a promising new approach.

3. DESIGN AND METHODOLOGY

Creating our graph of drugs consists of three main stages. The first stage is to create a database containing relevant information extracted from clinical trials XML files. A row in the database is formatted as follows:

Drug name | Effect name | Effect type | Phase

In the above tuple, “Effect name” denotes either a name of a condition the drug was used to treat, or the name of an adverse event (ie. a side effect) caused by the drug. Whether the effect name denotes a treated condition or an adverse event is specified in the “Effect type” part of the tuple. Finally, each tuple in the database contains the phase of the trial it was extracted from. A single trial can yield many such tuples.

The minimum amount of information that should be stored at this stage is the drug name, effect name, and effect type. Any information beyond that is simply for use when weighting connections between drugs. We note that there is a great deal more information that could be stored at this stage, particularly prevalence of adverse events. However, due to the time and expert knowledge required to construct a more complex weighting function, we restricted our database to the above four fields.

The second stage is to construct a preliminary graph from the database. We will refer to this preliminary graph as the “indirect graph” because it will not contain any direct connections between drugs, but rather, will contain only connections between drugs and effects. To construct the preliminary graph, rows are read from the database and parsed into a node representing the drug, a node representing the effect, and an edge between those nodes, which is weighted either positively or negatively and scaled by the phase. This is illustrated in the following example, where one drug treats nausea and has a positive connection, and another drug causes nausea and has a correspondingly negative connection:

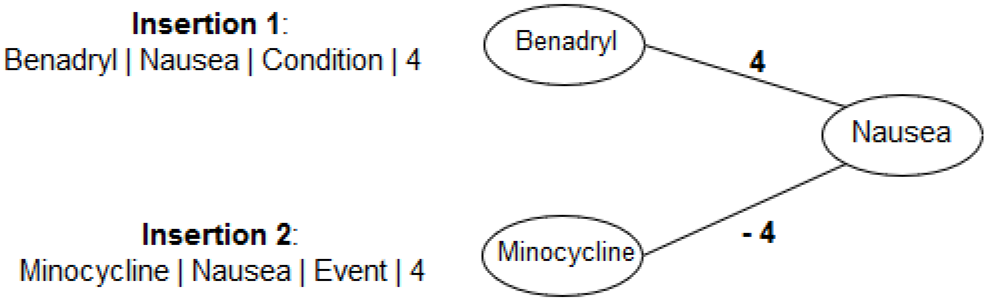

Insertion 1:
Benadryl | Nausea | Condition | 4
Insertion 2:
Minocycline | Nausea | Event | 4
Benadryl
4
Nausea
- 4
Minocycline

Figure 3: Weighting based on treated condition vs. adverse event

A given drug/effect combination may be encountered many times from different trials. However, duplicate nodes are not allowed in the graph. Consider the example of encountering a phase 4 study showing that ibuprofen treats pain, and then later encountering another study that also uses ibuprofen to treat pain. There are two options for weighting the ibuprofen/pain connection:

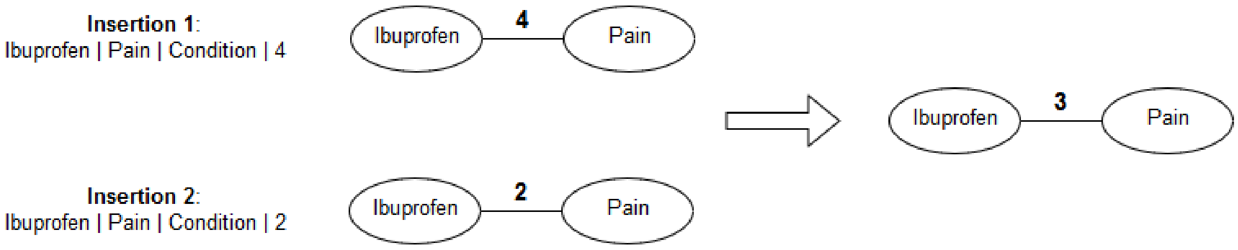


Figure 4: Weighting by taking the average

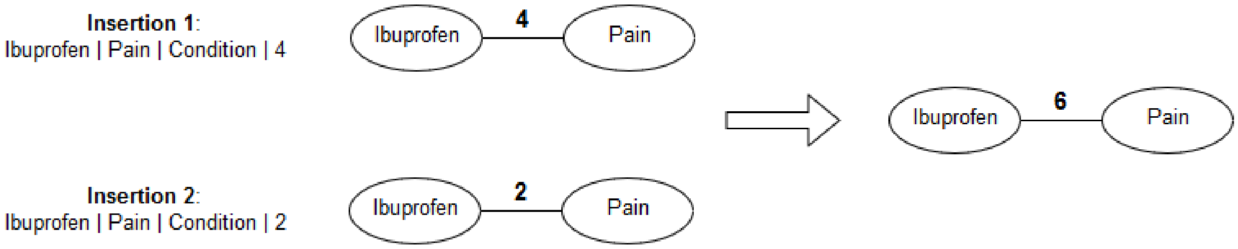


Figure 5: Adding the weights together

The choice of whether to take the average or to add the weights together largely determines the meaning of the graph, as will be explained in the following paragraphs. We chose to add the weights together, as in Figure 5.

Once the entire indirect graph has been built, the final stage is to prune effect nodes from it, leaving only drug nodes. When an effect node is pruned, all drugs connected to it become interconnected with each other. This is illustrated in the following figure, where the effect node to be pruned is shown in red:

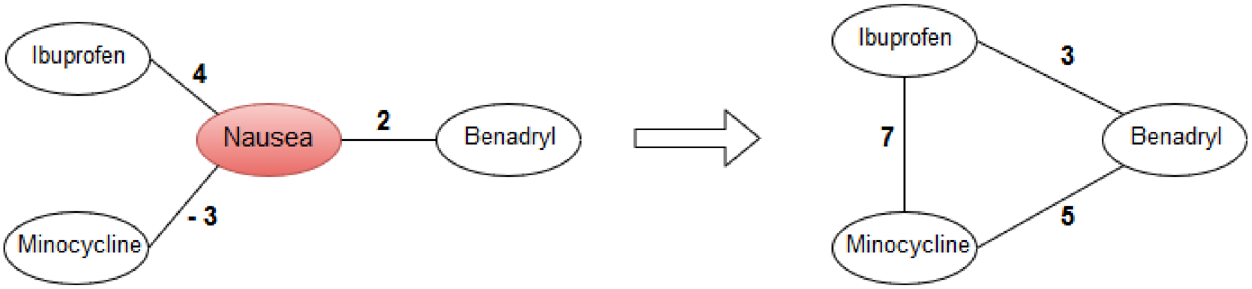


Figure 6: Pruning an effect node and weighting connections between drugs

Given two drugs connected to the same effect node, a connection will be made between those drugs when the effect node is pruned, and it will be weighted as follows:

- If one connection is positive and the other is negative, the connection between the drugs will be the sum of the absolute values of the two connections.
- Otherwise, the connection between the drugs will be the average value of the two connections.

According to this weighting scheme, the strongest connections are made when one drug treats the side effects of another drug, because there is high potential for the two drugs to be used together to provide a better treatment experience for a patient. Conversely, two drugs that cause the same adverse event will have a correspondingly negative connection to each other, and two drugs that treat the same condition will have a positive connection drawn between them, although with no added strength since there is not much point in using them both at the same time.

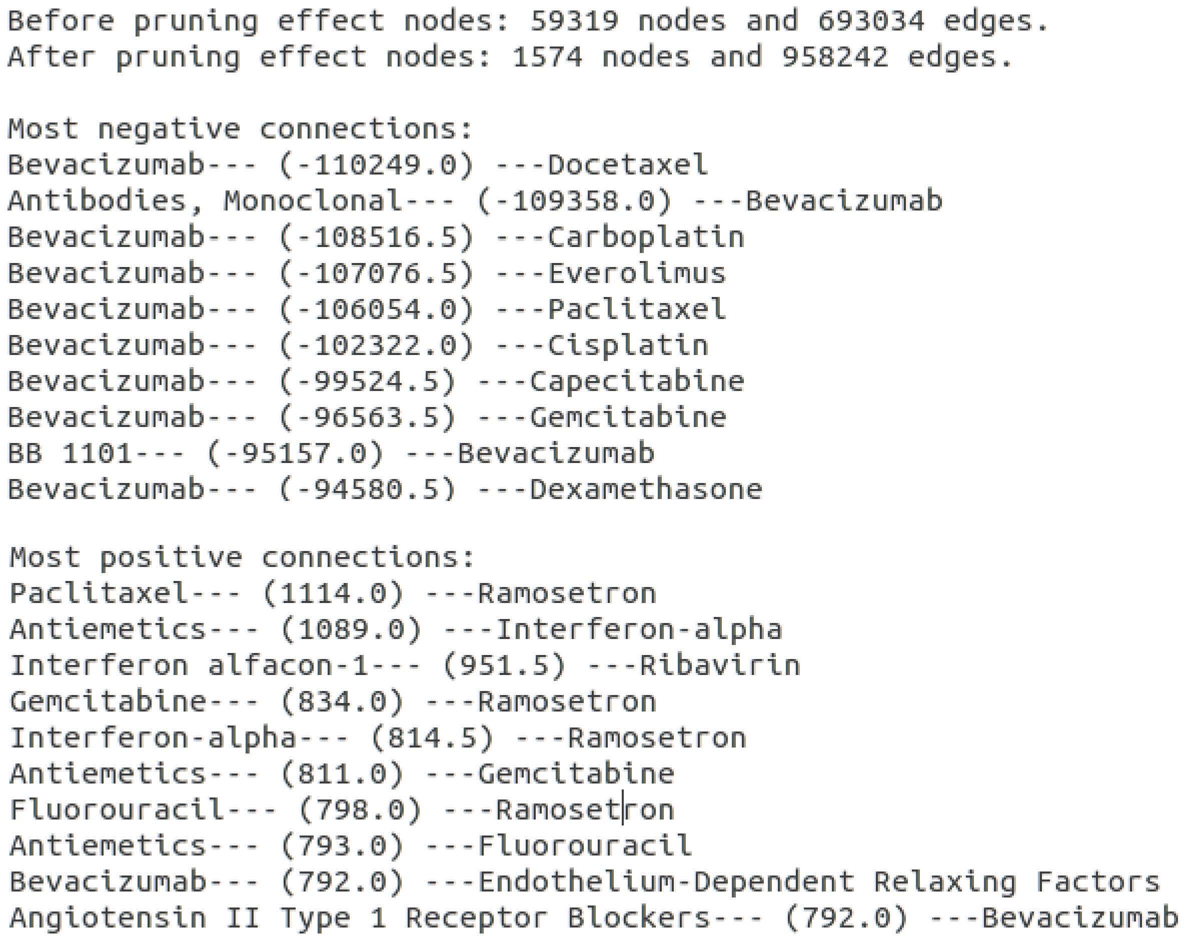
Referring back to the indirect graph, recall the choice of averaging weights versus adding weights. If we choose to average the weights in the indirect graph, as in Figure 4, the max and min possible weight values will be 4 and -4. From there, it follows that all weights in the final graph will be in the range of [-4, 8]. Thus, choosing to take the average as in Figure 4 is useful if you want connection weights in the final graph to all be across the same fixed scale, allowing for easy comparison of drug connections.

Conversely, choosing to sum the weights when building the indirect graph results in unbounded connection weights in the final graph. For example, if there are thousands of clinical trials connecting "Ibuprofen" to "Pain," then the connection weight between Ibuprofen and Pain will be increased repeatedly, up to a very large value. Thus, summing the weights essentially integrates frequent pattern mining into the graph building process. As a result, a highly weighted edge in the final graph indicates that one or both of the associated drugs is highly prevalent in clinical trials. The sign of the connection (either positive or negative) still indicates whether the drugs are harmonious or not, but the magnitude of the harmony is more difficult to infer in this case.

The procedure described here does not adhere strictly to any previous graph algorithms to our knowledge. It is an algorithm that takes advantage of nearby connections to draw connections between nodes that were previously unconnected. However, the primary issue addressed here is how to achieve efficiency in building a very large, high-connectivity graph of this type, given the inherent need for list data structures that need to be frequently accessed and modified. To that end, we made clever use of hash tables in our implementation, and achieved linear run time in as many aspects of the graph building process as possible, leading to very good performance.

4. RESULTS

To build our graph, we downloaded all completed interventional studies with results available on clinicaltrials.gov, which amounted to about 19,000 studies and over 1,600,000 rows in the database described in Section 3. The final graph contained 1574 drugs and over 950,000 edges. Some of the strongest connections are shown below.



```
Before pruning effect nodes: 59319 nodes and 693034 edges.
After pruning effect nodes: 1574 nodes and 958242 edges.

Most negative connections:
Bevacizumab--- (-110249.0) ---Docetaxel
Antibodies, Monoclonal--- (-109358.0) ---Bevacizumab
Bevacizumab--- (-108516.5) ---Carboplatin
Bevacizumab--- (-107076.5) ---Everolimus
Bevacizumab--- (-106054.0) ---Paclitaxel
Bevacizumab--- (-102322.0) ---Cisplatin
Bevacizumab--- (-99524.5) ---Capecitabine
Bevacizumab--- (-96563.5) ---Gemcitabine
BB 1101--- (-95157.0) ---Bevacizumab
Bevacizumab--- (-94580.5) ---Dexamethasone

Most positive connections:
Paclitaxel--- (1114.0) ---Ramosetron
Antiemetics--- (1089.0) ---Interferon-alpha
Interferon alfacon-1--- (951.5) ---Ribavirin
Gemcitabine--- (834.0) ---Ramosetron
Interferon-alpha--- (814.5) ---Ramosetron
Antiemetics--- (811.0) ---Gemcitabine
Fluorouracil--- (798.0) ---Ramosetron
Antiemetics--- (793.0) ---Fluorouracil
Bevacizumab--- (792.0) ---Endothelium-Dependent Relaxing Factors
Angiotensin II Type 1 Receptor Blockers--- (792.0) ---Bevacizumab
```

Figure 7: Partial summary of final weighted drug network

Manually searching the internet for some of the connections shown above yields interesting results. Many of the highly-weighted connections correspond to drug pairs that, intuitively, make sense. For example, the number 1 connection above involved paclitaxel and ramosetron. Paclitaxel is a chemotherapy medication, and ramosetron is a medication primarily used for treating nausea. The connection between those two drugs is intuitive because nausea is a common side effect of chemotherapy. However, the connection is made even more interesting when one looks up paclitaxel on drugbank.ca. Paclitaxel has an extensive list of drug interactions, but ramosetron is not listed among them, suggesting that ramosetron might be completely safe for use with paclitaxel. Furthermore, a Google search of “paclitaxel and ramosetron” yields few results relating the two medications to each other.

Such a connection is exactly the type that we hoped to discover in our project. The connection represents the hypothesis that paclitaxel and ramosetron might be useful in a joint treatment, and that further study should be done on their joint usage.

Furthermore, we conducted a simple apriori frequent pattern mining effort, in which we found frequent 2-patterns of drugs used together in clinical trials. The results of the frequent pattern mining include many of the highest-ranked drugs in our graph, confirming that our graph construction methodology is more general and informative than frequent pattern mining alone.

5. DISCUSSION

The need for personalized medicine is apparent and immediate as much ongoing research suggests. In this study we have proved that a graph solely built from the mined results data from ClinicalTrials.gov can be used to analyze, validate and even predict new drug-drug connections for the purposes of drug repositioning and repurposing that can contribute to personalized medicine research.

Perhaps the most distinct feature of our graph building process is that drugs do not have to come from the same study to have a connection drawn between them. Any two drugs from any two studies can be connected, based solely on the fact that they have an effect in common. This act of making inter-study connections is highly novel to the best of our knowledge. Furthermore, our process is very versatile. Combined with extraction of different information from the clinical trials, and different weighting functions, our methods can easily be repurposed to potentially predict other types of connections between drugs.

Even though we had challenges like converting and filtering informal English words for medicine practices and drug names and dosages, weighting our edges, and implementing a data mining algorithm for frequent pattern mining, we were able to validate our methods and our results are promising to be valuable to any drug information network research. We believe our work is unique in ways of creating a drug-drug interaction graph that considered conditions, phase information and adverse events while weighting the edges and cross-validating with another data mining approach, namely frequent pattern mining. This work contributes to our understanding of data mining algorithms and approaches that were made clear in our classes.

6. FUTURE WORK

Future directions suggest that our work needs to be validated in a variety of ways and enhanced with the exact dosage information of drugs. The procedure used for assembling the graph is extremely versatile, and the way the edges are weighted during effect node pruning could be adjusted to predict something besides likelihood of useful joint treatment. With more sophisticated methods such as regression or even a neural network, and greater usage of things like demographic information stored in clinical trials, the edges could be weighted to predict other things, such as suitability for a specific patient's personalized treatment.

7. REFERENCES

[1] https://www.genome.gov/10001772/all-about-the--human-genome-project-hgp/

[2] https://www.23andme.com/howitworks/

[3] https://clinicaltrials.gov/ct2/about-studies/learn

[4] http://www.fda.gov/RegulatoryInformation/Legislation/Federal

[5] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.

[6] https://clinicaltrials.gov/ct2/about-studies/glossary

[7] https://clinicaltrials.gov/ct2/help/how-read-study

[8] Luo Z, Zhang G, Xu R. Mining Patterns of Adverse Events Using Aggregated Clinical Trial Results. AMIA Jt Summits Transl Sci Proc. 2013; 112(6): 112-116.

[9] Southworth H, O'Connell M. Data Mining and Statistically Guided Clinical Review of Adverse Event Data in Clinical Trials. Journal of Biopharmaceutical Statistics. 2009; 19: 803-817.

[10] Korkontzelos I, Mu T, Ananiadou S. ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials. BMC Medical Informatics and Decision Making. 2012; 12(3): 1-12.

[11] Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, et al. The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty. PLoS ONE. 2012; 7(3): 1-12