Aggregated Searchable Interface of Clinical Trials Data

Department of Computer and Information Science, The University of Pennsylvania, Philadelphia PA

Elizabeth Conrad
The University of Alabama
ecconrad1@crimson.ua.edu

ABSTRACT

Though there is a vast amount of medical literature available to the public, it can sometimes be difficult to find and even more difficult to consume. Patients should have easy access to the expanse of information that is out there in an easily digestible format. The interface detailed in this paper offers patients with an easy and streamlined way to access an aggregation of important data in one place. Patients can search by condition and quickly view the most popular interventions, outcomes, and associated publications.

INTRODUCTION

Being an informed patient is both vitally important and unfortunately complicated. Existing methods of gathering information concerning particular conditions are not as comprehensive as they could be. Our project, BrowsingHealth, aims to bring the vast expanse of available medical data into one easily navegable location. In addition to providing interested patients with a place to access information concerning the populations, interventions, conditions, and outcomes associated with published clinical trials and literature, it also may serve to assist in the formulation of and expedite in the answering of queries posed by clinicians. The PICO (problem/population, intervention, condition, outcome) framework is used by physicians who practice evidence-based medicine (EBM) [1]. Adhering to the PICO framework can often produce more robust queries than natural language would alone [1]. Due to the relevance of the framework and its intuitive nature for users not involved in the medical field, we organized the data and interface around its structure. Even with minimal data cleaning and aggregation, the interface works well to display relevant information.

IMPLEMENTATION

The repository used for the system was downloaded from ClinicalTrials.gov. The website provides XML files of every study registered on the site which may be downloaded in one compressed folder. Once the data was downloaded, BeautifulSoup was used to parse the XML data and it was loaded into a Mongo database. The database included one document per condition, which were aggregated by string matching. Each condition document had a frequency count of how many studies listed that condition, associated interventions and frequencies, and associated outcomes and frequencies. With the data aggregated, searching for the most commonly used interventions and outcomes both for individual conditions and also overall was made simpler, but an

interface was needed to more elegantly streamline the process of viewing and analyzing the database.

The interface was built using a MERN stack, which includes Mongo, Express, React, and Node. The homepage to the website includes a search bar that shows suggestions as text is entered for conditions that contain the given string, sorted by how many studies exist for the respective condition. Beneath the search bar are Quick Access links.

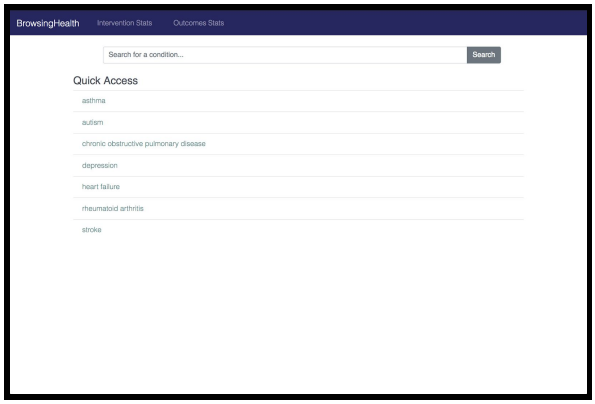

Figure 1. Homepage

The Quick Access links are associated with conditions for which additional data cleaning was done. Using SNOMED IDs, different names for conditions that are considered to be the same were aggregated. For example, the page for "depression" will also include PICO information from studies associated with "depressive disorder," "depressive episode," "depressive illness," and "depressed."

The condition page includes tables of the interventions, outcomes, publications, and matching conditions for the given condition.

The interventions and outcomes tables are sorted by the number of registered trials by default, but this can be changed by clicking another column.

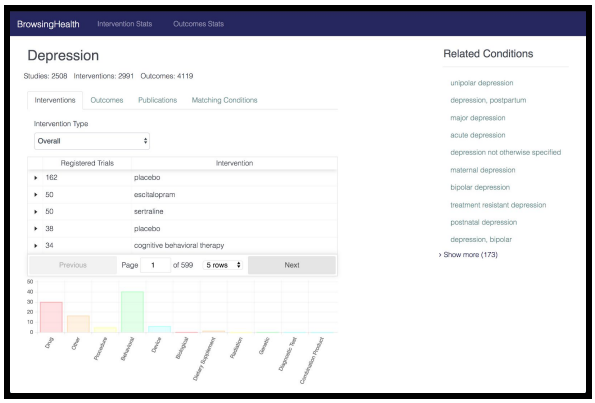

Figure 2. Condition page, interventions tab

Each row of the interventions table includes the name of the intervention and the number of registered trials for that intervention. Each row may be expanded and links are shown for all the associated trials so users may view them.

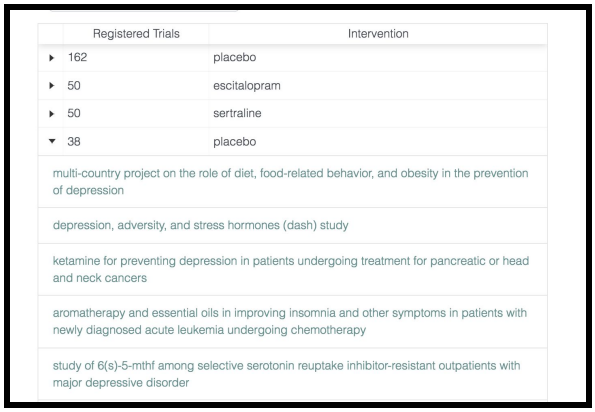

Figure 3. Expanded interventions row

Since there are multiple types of interventions (e.g. drug, behavioral, dietary supplement, etc.), there is a drop down menu provided which allows the interventions to

be filtered by any of the given types. The default filter is none, thus all interventions are shown.

In addition to the interactive interventions table, a graph is shown which displays, in percentages, the distribution of the types of interventions administered for the given condition.

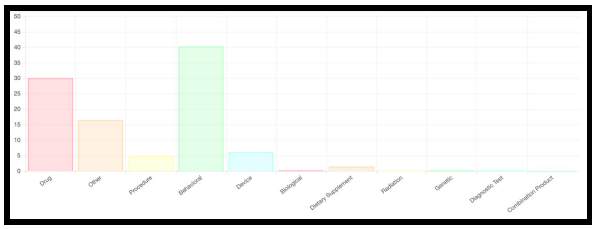

Figure 4. Intervention distribution graph

The outcomes tab contains a similar expandable table with the name of the outcome and the number of registered trials.

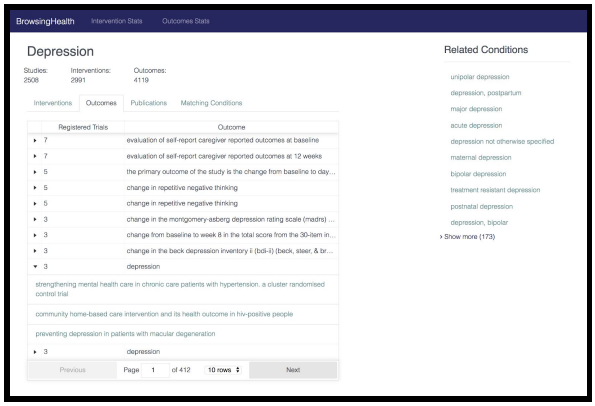

Figure 5. Outcomes tab

The publications tab contains all the information provided with the XML data provided by ClinicalTrials.gov for a given trial regarding associated publications. This information is the PubMed ID and the citation.

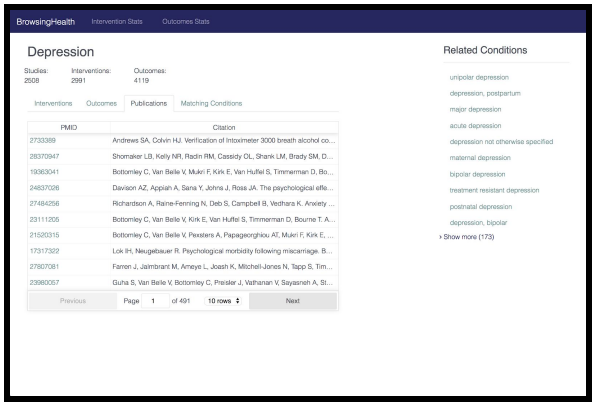

Figure 6. Publications tab

The matching conditions tab shows all the conditions that have been aggregated together for the displayed information and the counts of the trials under each condition name.

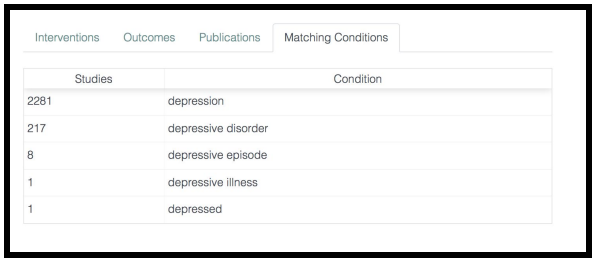

Figure 7. Matching conditions tab

The last important piece of the condition page is the related conditions column to the far right. It currently uses simple string matching to display any conditions that contain the full string of the condition of the current page. It shows the top ten related conditions, but allows the user to expand and show all the available related conditions in the database.

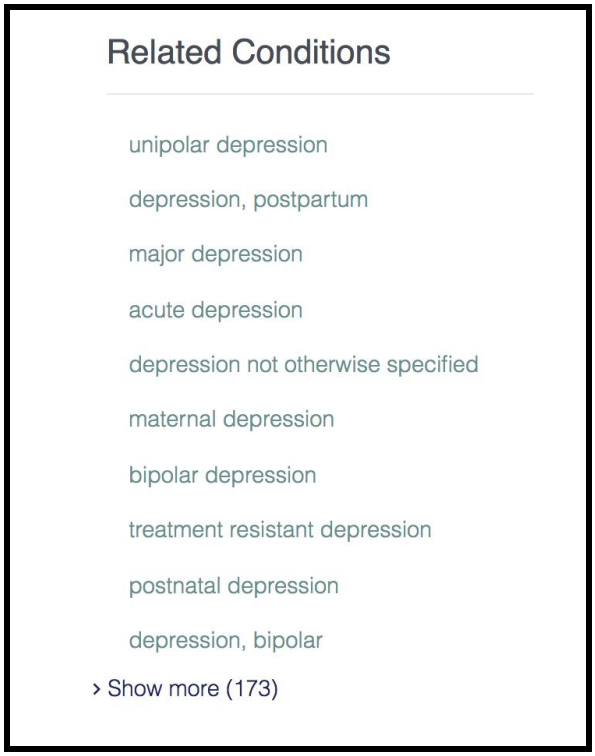

Figure 8. Related conditions panel

FUTURE WORK

While the interface shows a lot of useful data, there are numerous aspirations to improve both the underlying data and the interface used to interact with it. Though some very basic cleaning measures have been taken, more thorough work needs to be done to ensure that each condition page contains the information for all available relevant trials. Additionally, the interventions and outcomes have not been cleaned at all (beyond aggregating exact string matches), and some items that are worded slightly differently should be further aggregated. Extraneous interventions and outcomes such as "placebo" need to be removed or at least hidden by default as well.

Beyond the cleaning of the existing database, the ultimate goal will be to include less structured sources of medical literature, which would create a need to start using NLP techniques to automatically read and identify PICO elements.

Regarding the interface, as with any interface, the number of potential additional features is nearly limitless, but the most pressing feature is the ability to on-demand select a "bucket" of conditions (e.g. depression, anxiety, OCD, etc.) and have the page display the aggregation for all of these conditions in real time, rather than being limited to the information for one condition at a time. This will allow patients with comorbidities and physicians working with such patients to view relevant literature for those comorbidities simultaneously very easily.

CONCLUSION

BrowsingHealth is a promising tool for anyone interested in easily accessing medical literature and associated statistics, whether those users are in the medical field or simply patients with a desire to be informed. With the eventual goal of serving as an all-inclusive hub for medical literature, it aims to bring clarity to the medical world.

REFERENCES

[1] Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of PICO as a Knowledge Representation for Clinical Questions. *AMIA Annual Symposium Proceedings*, *2006*, 359–363.