MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220152

NLP-Assisted Pipeline for COVID-19 Core Outcome Set Identification Using ClinicalTrials.gov

Fatemeh Shah-Mohammadi, Irena Parvanova, Joseph Finkelstein

Center for Biomedical and Population Health Informatics, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

Core outcome sets (COS) are necessary to ensure the systematic collection, metadata analysis and sharing the information across studies. However, development of an area-specific clinical research is costly and time consuming. ClinicalTrials.gov, as a public repository, provides access to a vast collection of clinical trials and their characteristics such as primary outcomes. With the growing number of COVID-19 clinical trials, identifying COSs from outcomes of such trials is crucial. This paper introduces a semi-automatic pipeline that can efficiently identify, aggregate and rank the COS from the primary outcomes of COVID-19 clinical trials. Using Natural language processing (NLP) techniques, our proposed pipeline successfully downloads and processes 5090 trials from all over the world and identifies COVID-19-specific outcomes that appeared in more than 1% of the trials. The top-of-the-list outcomes identified by the pipeline are mortality due to COVID-19, COVID-19 infection rate and COVID-19 symptoms.

Keywords:

Core outcome set; Natural language processing (NLP), COVID-19.

Introduction

Design of clinical trials lacks consensus on which common outcomes should be measured. Significance of development of core outcome sets (COS) for clinical trials has been emphasized by researchers in various fields [1-3]. Identification of COS can realize cross-study aggregations and comparisons, speed up meta-data analysis and enhance efficiency and reproducibility [1-2]. Therefore, development of an automated and semi-automated approach that identifies and ranks COS has been an important objective for research informatics.

Novel Coronavirus Disease 2019 (COVID-19) caused a worldwide pandemic outbreak, which as of May 2021 has resulted in more than 153 million infections and over 3 million deaths worldwide [3-4]. Due to the agonizing impact of the pandemic on the world population, large portion of the scientific community worldwide has focused on understanding and battling the virus. As a result, an enormous number of studies have been conducted and published over the last year, with increasingly granular data sets. For this reason, the importance of categorizing outcome sets using relations and classes has increased in COVID-19 research. Currently, multiple studies, both in the USA and internationally, have focused on establishing core outcome sets for trials in COVID-19 patients in attempt to establish COVID-19 standards and controlled terminologies. Variety of approaches have been taken in this direction. For instance, the Clinical Characterisation and Management Working Group of the WHO Research and Development Blueprint programme, the International Forum for Acute Care Trialists, and the International Severe Acute Respiratory and Emerging Infections Consortium have developed a minimum set of common outcomes, which includes three elements: measures of viral infection, measures of patients' survival, and patients' progression through the healthcare system [5]. Similarly, other studies also used international experts by implementing online surveys, such as the Delphi method in multiple languages, to establish main clinical outcomes in clinical trials for COVID-19 patients [6-7]. Additional methods for establishing core outcomes sets for clinical trials included organizing international workshops of experts of over 100 countries [8]. Other studies used publicly available data repositories, such as ClinicalTrials.gov to identify common outcome sets in COVID-19 clinical trials [9]. In addition, multiple COVID-19 databases have been established, such as the German Corona Consensus Dataset (GECCO), where core dataset consisting of 81 data elements. These elements included information about demography, medical history, symptoms, therapy, medications or laboratory values of COVID-19 patients [10].

In this study, we aimed at introducing a semi-automatic pipeline to identify, collect, and rank the COSs from the primary outcomes of COVID-19 clinical trials using Natural language processing (NLP) techniques. The proposed pipeline allowed for the download of 5090 international trials and the identification of core COVID-19-specific outcomes on a larger scale than in previous attempts.

Methods

The main data source used in this paper is ClinicalTrials.gov (CTG). CTG serves as a mandatory repository for clinical trials and is maintained by the U.S. National Library of Medicine [11]. In addition to user interface, CTG offers a RESTful application programming interface (API) that facilitates the automation of submission of search query from a computer program, and returns the results in different formats such as XML and CSV for further processing. We have implemented our CTG query pipeline using Python 3.7 [12] and the URLLIB.request, Pandas, and Xml.etree libraries. In the following we describe components of our developed pipeline from the input query to the final output. Figure 1 also illustrates the steps involved in the pipeline.

Figure. 1– Pipeline workflow

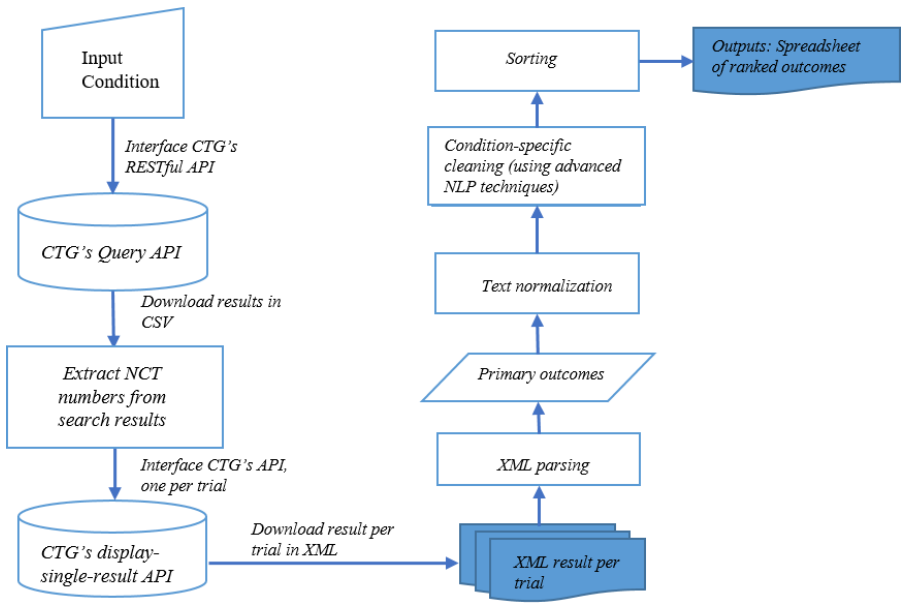

Table. 1– overview of the output of the pipeline at step 3

#	Primary outcome	Description	n	NCT number
1	Mortality	Cumulative incidence	81	NCT04360824, NCT04450017, […]
2	Death	Death from any cause	35	NCT04644198, NCT04374071, […]
3	All-cause mortality	Death from trial start date to 30 days recorded in the electronic health record.	24	NCT04659876, NCT04800770, […]
4	Time to clinical improvement	Time to clinical improvement (TTCI) was defined as time (days) from randomization to a decline of 2 categories on the seven-category ordinal scale of clinical status or live discharge from the hospital, whichever came first	23	NCT04669925, NCT04584580, […]
5	Clinical improvement	In the current study the clinical improvement will be determined by: Temperature Heart rate (pulse) Respiratory rate Oxygen saturation Need for oxygen Hospital stay time. CT lung involvement at day 0 and day 14.	21	NCT04354519, NCT04501458, […]
6	In hospital mortality	In-hospital mortality after administration of ABO compatible convalescent plasma or indication (but not plasmapheresis for absence of compatible convalescent plasma) for comparison group	20	NCT04493268, NCT04436484, […]

Step1: Interfacing with the CTG API search endpoint

The only input term for the search query was selected to be the word "COVID-19". This input term is embedded into a URL to interface CTG RESTFul API at https://clinicaltrials.gov/ct2/results/download_fields?cond=COVID-19. Other parameters of the API call include as follow: *down_count* which specifies the number of records to download, *down_chunk* which represents which set of records to include in the downloaded file relative to the option selected for the *down_count*, and *down_fmt* specifies the format of the results table that can be XML, CSV, PDF, TSV, and PLAIN.

This call is equivalent to using the term "COVID-19" for the *"condition or disease"* field when one decides to use the CTG user interface. In our implementation, we set the number of *down_count* to 10,000 and *down_chunk* to be 1. Each call returns a table of results in CSV format. Using the term "COVID-19", our interface with the search endpoint outputted one CSV formatted table with 5090 rows at the time of our inquiry. We used Python's Pandas library to parse the table into a DataFrame object. This DataFrame stores the details about the trials that meet the input condition. The details include the National Clinical Trial (NCT) number of each trial and the location in which the study has been conducted. The NCT number is a unique number which is assigned to a registered study and can be used to download the full record of that study. It should be noted that the result of our search covers the studies from all over the world.

Step 2: Interfacing with the CTG to download the trial records using NCT number

As mentioned the resultant DataFrame at the previous step contains the NCT number for each trial. We used this number to interface with CTG again to download the full records for each trial. CTG offers an API to display a single record in XML by calling https://clinicaltrials.gov/ct2/show/NCT_Number?displayxml=true. For each NCT number, an API call is made and the records of that study are saved in XML format for further parsing in the next step.

Step 3: Parsing the XML file for primary outcome

After downloading the records for each trial in XML format, for every trial our developed pipeline parses the primary outcomes and the description associated with each. We used Python's built-in *Xml.etree* module to parse the XML data into a tree with different nodes. We then defined a function iterating over every node. Each node has further children nodes. Our pipeline only extracts the children nodes associated with the node tagged as "*primary_outcome*". This node's children nodes are tagged as *measure* (which records the name of the outcome), *description* (which provides the description for each outcome), and *time_frame* (which records the time frame for the study). Our developed pipeline parses the texts associated to *measure* and *description* tags and stores them in a DataFrame along with the NCT number.

Step 4: Normalizing the outcome texts

In this step, the pipeline takes the resultant DataFrame in the previous step, and normalizes the texts in the columns storing the primary outcomes and their description. Given a string, the pipeline uses the Natural Language Processing (NLP) techniques to apply the following changes:

- Change the string to all lowercase and strip the spaces from both ends.
- Replace all punctuations with space.
- Replace all occurrences of two or more space marks with one space mark.
- Deduplication

Then the pipeline ranks the outcomes based on their appearance frequency in the entire 5090 trials. Table. 1 shows the results of this step (due to the limitation in space, we only listed the first 6 rows). It should be noted that 9243 unique outcomes were identified at this stage. Considering the pipeline output at this stage, we noticed that the top outcomes happen tens of times across different trials. While, there were thousands of outcomes that occurred only once. After manual inspection, we found that some of them semantically are identical to the top outcomes but they appeared unique as they were written in a specific way. Then, we decided to add an additional component to the pipeline to aggregate such outcomes together. We also found that the top outcomes could even be aggregated. For instance, "Death", "Mortality" and "All-cause mortality", the most common outcomes according to Table. 1, are semantically the same and can be aggregated.

Step 5: Grouping the similar outcomes

As mentioned, due to the identification of a large number of overlapping outcomes from the output at the previous step, we added another step into our developed pipeline where we use more advanced NLP techniques and human judgement to group the overlapping outcomes and rank the resultant groups by the number of outcomes under each. For instance, the outcomes "hospitalization", "need for hospitalization", "duration of hospitalization in days", "length of hospitalization", and "length of hospital stay (days)", all convey the similar message and the pipeline has aggregated them under outcome "hospitalization".

Table. 2– General statistics

Variable	
Number of trials downloaded using the query term "COVID-19"	5090
Percentage of trials listing outcome	100%
Number of unique primary outcomes parsed (after normalization and deleting duplicates, before conducting condition specific cleaning)	9243
Number of unique primary outcomes after condition specific cleaning	2349
Time required by the pipeline to download and parse the trials and group the overlapping outcomes	16 mins

Results

We examined the outcomes for the 5090 clinical trials, and the results are shown in tables 1 to 3. Table. 1 provides an overview of the pipeline output at step 4. While the remaining tables are the outputs of the pipeline at the final stage.

Table. 2 provides general statistics related to our application of the pipeline using input query "COVID-19". As listed in this table, the total number of trials downloaded by the pipeline is 5090. After normalization and deduplication, 9243 unique outcomes were identified. Condition-specific cleaning (clustering related outcomes) reduced the number of unique outcomes to 2349. Last row in this table shows the time required by the pipeline to output the final core outcomes and ranked them by their frequency. It can be seen that the pipeline is able to facilitate the process of identifying the outcomes of thousands of trials in such a short time (16 minutes), and thus aiding the design of future clinical trials.

Table. 3 lists the main output of the pipeline that is the top outcomes appeared in more than 1% of the trials. Columns three and four in this table list the occurrence frequency and the percentage of the appearance of each outcome in all 5090 trials, respectively. The last column shows the NCT number for the trials in which the corresponding outcome has listed as primary outcome. By listing the NCT numbers, this column facilitates access to the trials based on search for outcome of interest.

Table. 3– Top 18 outcomes listed in more than 1% of the COVID-19 clinical trials

#	Outcomes	n	%	NCT numbers
1	Mortality due to COVID-19	743	14.6%	NCT04371562, NCT04504734, [...]
2	COVID-19 infection rate	591	11.61%	NCT04567979, NCT04384926, [...]
3	COVID-19 symptoms	426	8.37%	NCT04646109, NCT04463420, [...]
4	Mental health impact of COVID-19	355	6.97%	NCT04470609, NCT04747756, [...]
5	Incidence of adverse events and reactions	342	6.72%	NCT04551547, NCT04382040, [...]
6	Intubation, ventilation, and oxygenation	323	6.35%	NCT04261270, NCT04646109, [...]
7	Clinical improvement and time to clinical improvement	295	5.8%	NCT04542694, NCT04359615, [...]
8	Treatment and response to treatment	232	4.56%	NCT04374695, NCT04535869, [...]
9	COVID-19 antibodies and anti- COVID-19 antibodies	182	3.58%	NCT04591717, NCT04591717, [...]
10	Hospitalization	173	3.4%	NCT04552951, NCT04334382, [...]
11	Assessment of vaccination	128	2.51%	NCT04659941, 'NCT04750343, [...]
12	Change in physical activity	117	2.3%	NCT04768257, NCT04768257, [...]
13	Clinical outcome	113	2.22%	NCT04480593, NCT04444401, [...]
14	Prevalence rate	89	1.75%	NCT04352764, NCT04472078, [...]
15	Health related quality of life	85	1.67%	NCT04751721, NCT04447222, [...]
16	ICU (admission rate and length of stay)	80	1.57%	NCT04374565, NCT04397614, [...]
17	viral load, viral infection and clearance	79	1.55%	NCT04438850, NCT04802408, [...]
18	lung and chest (injury and involvement)	76	1.49%	NCT04794985, NCT04409275, [...]

The three most common COVID-19 clinical outcomes with percentage of the appearance in more than 1% of the trials are: Mortality due to COVID-19 (14.6%), COVID-19 infection rate (11.61%), and COVID-19 symptoms (8.37%).

Discussion

COSs are vital for ensuring comparability of clinical trial data and enabling meta-analyses and interstudy comparison. In response to the need for developing COSs for the rapidly evolving COVID-19 outbreak, multiple studies have focused on establishing main outcomes for trials in COVID-19 patients in attempt to establish COVID-19 standards and controlled terminologies [3-8]. To decide which outcomes should be recommended as COS, they relied on assembling and surveying panels of subject-matter experts that is usually a time-consuming and laborious process. While our developed pipeline can automatically identify COVID-19 specific COSs by finding, downloading and analyzing data of the COVID-19 clinical trials registered in CTG. Based on the results shown in Table. 3, the most frequent clinical outcomes appeared in 743 trials out of 5090 trials is “Mortality due to COVID-19”.

The second and third most frequent outcomes in the trials are “COVID-19 infection rate” and “COVID-19 symptoms” that appeared in 11.61% and 8.37% of the trials, respectively.

Authors in [13] have also implemented a pipeline that identifies common outcome set in stem cell clinical trials. Our work is in concordance with [13] in effective identification of evidence-base disease specific core outcomes.

To analyze the quality of the pipeline’s output, we looked at the differences in results between what pipeline generated and the outcomes listed in references [5-8]. On comparing the outcomes identified automatically by the pipeline to the outcomes identified by references [5-8], we found matches for 15 outcomes. Table. 4 lists the outcomes that not only appeared in at least one reviews but also have been identified by our developed pipeline.

Table. 4– Outcomes identified by published reviews

#	Outcome	Source
1	Mortality due to COVID-19	[5-8]
2	COVID-19 infection rate	[6]
3	COVID-19 symptoms	[4,6,7]
4	Mental health impact of COVID-19	[5,6,8]
5	Incidence of adverse events and reactions	[7]
6	Intubation, ventilation, and oxygenation	[5,6,7]
7	Clinical improvement and time to clinical improvement	[5]
8	COVID-19 antibodies and anti- COVID-19 antibodies	[7,8]
9	Hospitalization	[5-8]
10	Change in physical activity	[5,8]
11	Clinical outcome	[5]
12	Health related quality of life	[5,7]
13	ICU (admission rate and length of stay)	[7]
14	viral load, viral infection and clearance	[5,6,8]
15	lung and chest (injury and involvement)	[5-8]

Conclusions

Due to the lack of standardization in the current state of COVID-19 data collection, identification of COVID-19 common data elements could greatly facilitate data harmonization for the future COVID-19 research and clinical trial design. To address this necessity, in this work we have introduced a semi-automated COVID-19-based generation of clinical outcomes pipeline which interfaces with CTG application programming interface. Our developed COVID-19 specific outcome pipeline successfully downloaded and processed 5090 clinical trials. The primary outcomes of those trials were grouped based on text similarity and ranked based on frequency. The quality of the pipeline automatic output has been analyzed by comparing its output and the outcomes identified in comprehensive reviews. This pipeline can be added as a new option to the CTG search engine to obtain an evidence-based COVID-19 specific results at once.

References

[1] J. Sheehan, S. Hirschfeld, E. Foster, U. Ghitza, K. Goetz, J. Karpinski, et al., Improving the value of clinical research through the use of Common Data Elements, *Clin Trials* **13** (2016), 671-676.

[2] N.S. Redeker, R. Anderson, S. Bakken, E. Corwin, S. Docherty, S.G. Dorsey, et al., Advancing Symptom Science Through Use of Common Data Elements, *JNurs Scholarsh* **47** (2015), 379-388.

[3] WHO-China Joint Mission, Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), 2020.

[4] C. Mattiuzzi, G. Lippi, Which lessons shall we learn from the 2019 novel coronavirus outbreak?, *Ann. Transl. Med.* **8 (**2020).

[5] J.C. Marshall et al., A minimal common outcome measure set for COVID-19 clinical research, *Lancet Infect Dis* **20 (**2020).

[6] N. Evangelidis et al., COVID-19-Core Outcomes Set (COS) Survey Investigators. International Survey to Establish Prioritized Outcomes for Trials in People With Coronavirus Disease 2019, *Crit Care Med* **48** (2020), 1612-1621.

[7] X. Jin et al., Core Outcome Set for Clinical Trials on Coronavirus Disease 2019 (COS-COVID), *Engineering* **6 (**2020), 1147-1152.

[8] A. Tong et al., COVID-19-Core Outcomes Set (COS) Workshop Investigators. Core Outcomes Set for Trials in People With Coronavirus Disease 2019, *Crit Care Med* **48** (2020), 1622-1635.

[9] I. Parvanova, J. Finkelstein J. Identifying Core Outcome Sets in COVID-19 Clinical Trials Using ClinicalTrials.gov, *Stud Health Technol Inform*, 2021.

[10] J. Sass et al., The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond, *BMC Med Inform Decis Mak* **20** (2020), 341.

[11] D. Zarin et al., The ClinicalTrials.gov results database--update and key issues, *N Engl J Med* **364** (2011), 852-860.

[12] B. Ekmekci, C.E. McAnany, C. Mura, An Introduction to Programming for Bioscientists: A Python-Based Primer. *PLoS Comput Biol* **12** (2016), e1004867. https://doi.org/10.1371/journal.pcbi.100486

[13] A. Elghafari, J. Finkelstein, Introducing an Ontology-Driven Pipeline for the Identification of Common Data Elements, *Stud Health Technol Inform.* **272** (2020), 379-382.

Address for correspondence

Fatemeh Shah-Mohammadi, Center for Biomedical and Population Health Informatics, Icahn School of Medicine, New York, NY, Email: Fatemeh.shah-mohammadi@mountsinai.org