A PICO-based Knowledge Graph for Representing Clinical Evidence*

Yongmei Bai
School of Public Health; National Institute of Health Data Science
Peking University
Beijing, China
baiym2018@126.com

Huage Sun
School of Mathematics and Statistics
The University of Melbourne
Melbourne, Victoria, Australia
huage.sun@student.unimelb.edu.au

Jian Du*
Institute of Health Data Science
Peking University
Beijing, China
corresponding author,
dujian@bjmu.edu.cn

ABSTRACT

Background: Compared to unstructured representation of clinical evidence in bibliographic databases such as PubMed, structured results data available in clinical trial registries may be more timely, complete, and accessible, but these data remain underutilized.

Objective: The clinical trial information is extracted from the semi-structured records on ClinicalTrials.gov to construct a PICO-based knowledge graph for representing clinical evidence. The knowledge graph is expected to give a whole picture on the research protocol and reported results of clinical trials, thus it can be quickly searched, visualized, and exported in batches and on-demand.

Methods: We collected 6279 clinical trials about COVID-19 in ClinicalTrials.gov as the raw data, including 71 pieces of research with results. Information extraction, structured data standardization, visualization and query work were carried out in a semi-automated manner. The knowledge graph was constructed by neo4j 2.25 and Pyhon3.7.

Results: Two knowledge graphs are constructed. The first COVID-19 Trial Knowledge Graph (CTKG) contains 66856 nodes with 10 types and 1217673 relationships with 9 types in total. The second COVID-19 Trial Results Knowledge Graph (CTRKG) contains 1067 nodes with 12 types and 1405 relationships with 13 types in total. The graphs allow for queries, batch exports, and provide data for comprehensive clinical evidence based on PICO.

Conclusions: Our work validated the idea of "computable evidence synthesis" via presenting prespecified PICO data elements results data in trial registries in standardized, structured formats with consistent ontologies. Queries and batch export of information can be acquired in Graph Database built by neo4j through Cypher. It can help researchers obtain the latest data in batches and form a basis for the synthesis of real-world research evidence. Our methodology is also generalizable to other conditions and can incorporate registered clinical trial data from more platforms to achieve field unification of multi-source heterogeneous data.

KEYWORDS

COVID-19; Knowledge Graph; Neo4j; Clinical Trial; Cypher

1 Background

Ravaud, P et al. proposed "The 'one-off' approach of systematic reviews is no longer sustainable; we need to move toward producing 'living' evidence syntheses (i.e., comprehensive, based on rigorous methods, and up-to-date)[1]." This could lead to better health care decision-making. Clinical evidence is usually represented as scientific claims by the PICO format. PICO stands for Population (patients with a condition), Intervention, Comparison and Outcomes. For example, drug A is (Intervention) effective for the relief of B condition in C Population in Comparison with X drug (or placebo) for Y Outcome (symptoms relieved, etc.). For a long time, clinical evidence is predominantly disseminated in unstructured, natural language scientific publications that describe the results of randomized control trials (RCTs). The community from natural language processing (NLP), semantic web and health informatics has developed several approaches to making the clinical evidence structural and computable. It is difficult to quickly integrate the same high-quality RCT research of PICO. Previous studies have shown that the information from the clinical trial registration platform can help solve this problem[2].

The representative developments include building semantic representation for clinical trials and medical guidelines[3, 4], using linked data technologies to improve discovery of knowledge in systematic reviews by using the PICO framework as an ontology to aid in knowledge synthesis [5, 6]. The focus of NLP efforts is to identify the knowledge entities, namely the PICO elements and the general relations (e.g., hasPopulation, hasIntervention, hasOutcome) from RCTs. Nevertheless, they rarely take the effect-relations between interventions and outcomes into account. Most recently, the intervention-outcome-effect (i.e., improved, increased, decreased, no difference, no occurrence, etc.) as an important semantic relationship is introduced to inform augmentation mining and evidence inference [7] [8]. Given a treatment A, a comparator B, and an outcome, one can infer the reported relationship between A and B with respect to a concerned outcome, and provide evidence supporting this from the text.

Efforts aimed at increasing the pace of evidence synthesis have been primarily focused on the use of published articles, but these are a relatively delayed, incomplete, and at times biased

source of study results data. Existing studies results have shown that the same trial results are quite different from those reported in ClinicalTrials.gov and journal articles. Negative results in ClinicalTrials.gov tend to be frequently reported than in journal articles. This may be related to the fact that journal articles always report positive results. Further efforts should focus on decreasing existing published discrepancies to enhance the transparency and accuracy of data reporting in clinical trials [9, 10]. Compared to those in bibliographic databases, structured results data available in trial registries may be more timely, complete, and accessible, but these data remain underutilized [11]. Recently, a linked graph for registered COVID-19 clinical trials was built [12] to facilitate the retrieval of clinical trials. NLP tools were leveraged to extract and normalize the clinical trial information from both their eligibility criteria free texts and structured information from ClinicalTrials.gov. For example, an information extraction tool Critera2Query [13] is used to automatically identify entities from the text of eligibility criteria. Nevertheless, they did not extract the relations between interventions and outcomes, so they can only support the query of clinical trial information instead of clinical decision-making, such as determining which medical treatments work from the reported results in Clinicaltrials.gov.

To solve this problem, we believe that the information in registered clinical trials can be made into a structured knowledge graph. Given the increasing importance of text analysis in biology and medicine, we believe a local graph database of clinical trials will provide helpful computing infrastructure for researchers. To overcome this, automation technology is being explored by many researchers. Our research implements this process in a semi-automated progressing. We have carried out knowledge extraction, structured data visualization and query work. The specific contents include: 1) Automatically obtaining clinical trial data from registration platform. 2) Visualizing the structured data of the clinical trial registration platform through a knowledge graph. 3) Standardize indicators to realize the searchability of medical knowledge.

2 Materials and Methods

2.1 Data Source

Our research flowchart is shown in Figure 1. ClinicalTrials.gov is a database storing privately and publicly funded clinical studies conducted around the world. ClinicalTrials.gov provides storage and download of structured data, such as csv format, but the key information is not It is not completely accurate, such as not distinguishing intervention from comparison, and the combination of drugs involved in Intervention cannot be accurately displayed; in addition, the outcome indicator lacks a detailed description. The U.S. National Library of Medicine (NLM) [14] distributes ClinicalTrials.gov in eXtensible Markup Language (XML) formatted text files for each study. As a computer understandable language, XML can improve the processing speed, interpretability and extensibility of text information. but it is difficult to query information in that format. We used self-compiled software tools to parse the ClinicalTrials.gov data files and load their contents into a graph database to show the entities, their relationships and make information easy to query. Although the task is conceptually straightforward, the huge difference between clinical trial registries and results publications makes the task nontrivial. We collected 6279 COVID-19 clinical trials from ClinicalTrials.gov in XML format on August 3, 2021. Among them 71 clinical trials have results reported. The useful information was extracted from the downloaded XML document. This process was achieved through the open-access package “xml. etree. ElementTree” in Python3.7.

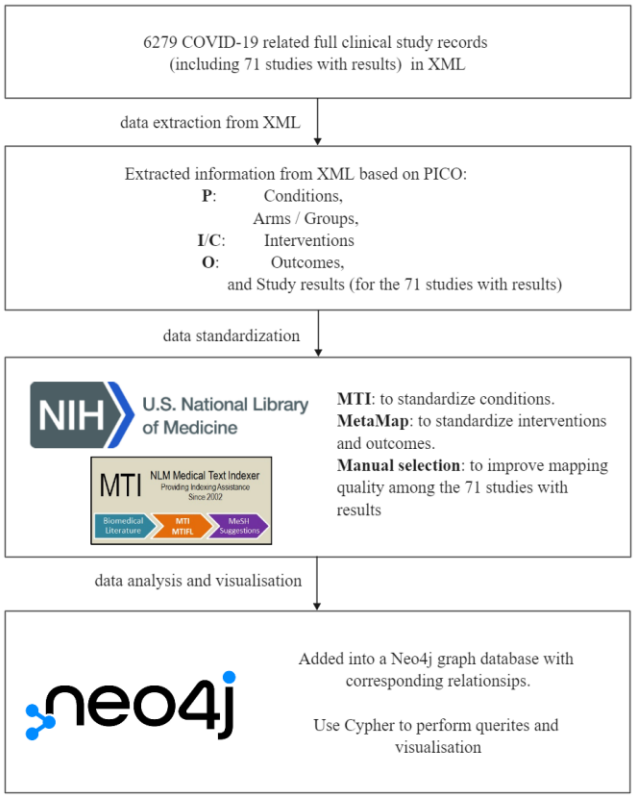

Figure 1: Flowchart of the research

2.2 Data structure

We extracted 14 fields from 6279 clinical trial XML files as following: Nctid, Title, Location, Study design, URL, other IDs, Conditions, Intervention/ Comparison, Outcomes, Funding, Enrollment, Gender, Age and Study type. Seventy one studies with results were added with two fields on this basis,i.e., Start Date and Completion Date.

According to the clinical trial registration information, valid fields were extracted and divided into 5 categories: 1) The metadata of Clinical trials: including 10 fields study information. 2) Conditions: since this study uses COVID-19 as an example, all clinical studies included the same condition, which is COVID-19. And we also extracted all other diseases and/or conditions in these fields. The complications of the disease are not distinguished. 3) Population Feature: three demographic characteristics of the population, i.e., age, gender, and enrollment were extracted. 4)

Intervention/comparison: According to the registration information of the existing clinical trial registrants, the intervention/comparison measure types are divided into 7 categories. This classification has the subjectivity of the research designer. Since the CT.gov web information only displays the intervention field, most current studies do not distinguish between intervention and comparison. We have obtained the measure type grouping situation from the XML documents, experimental and comparator group. As long as there exists evident cue word "experimental" in the type, the measure type is classified as intervention, and the null value or other value is classified as comparison. However, the type of many observational clinical trials is empty, and its grouping cannot be represented correctly. 5) Outcomes: We extracted all the outcomes descriptions from the XML data. As many clinical trials have not yet produced results or are not completed, many outcomes only stay at the trial design stage without corresponding reported result details. Fields contained in all categories, information in nodes are shown as Figure 2.

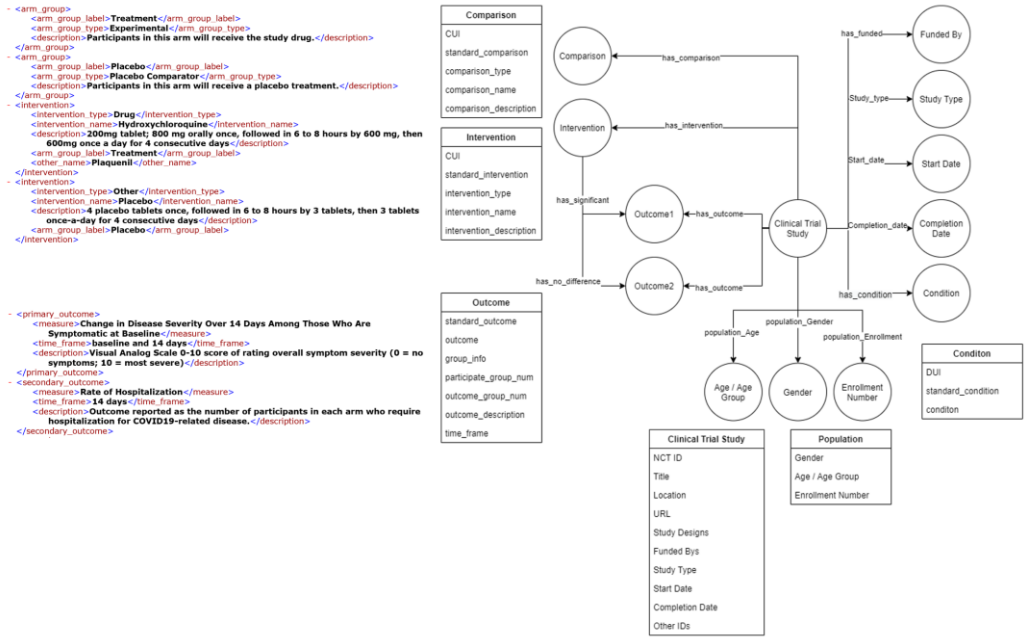

Figure 2: The entities and relationships represented in our knowledge graph

2.3 Data terms standardization

"The lack of a standardized outcome classification system results in inconsistencies due to ambiguity and variation in how outcomes are described across different studies."[15] In order to make better use of the visualized results and search for the research integrated into the knowledge graph, , we give a standardized terms on fields such as conditions, intervention/comparison, and outcome.

The Medical Text Indexer (MTI) produced by the NLM was adopted to extract standardized medical subject heading [12] terms in 6279 clinical trials conditions. And applied the results to 71 clinical trials that had results. It has been mapped to the MeSH Unique ID (DUI) number. DUI number is MeSH Unique ID finds Descriptor, Qualifier, and Supplemental Concept Records by their Record Unique Identifier https://www.nlm.nih.gov/mesh/mbinfo.html.

MetaMap is a highly configurable program developed by NLM to map biomedical text to the Unified Medical Language System (UMLS) Metathesaurus or, equivalently, to discover referred Metathesaurus concepts in the text. We standardized the intervention/comparison of 6279 clinical trials by using MetaMap tool. Among them, 5706 clinical trials have intervention/arm/group. No specific intervention was written in the 573 clinical trials. It may be because some clinical trials have not yet started, or because some clinical trials are observational studies without arm/intervention information. Finally, we analyzed a total of 5311 clinical trials. Since the clinical trials of has results announced specific plans for the combination of interventions, such as interventions including HCQ and AZT, the interventions used the two drugs as a result of the specific implementation. Therefore, we checked the measure groupings in 71 clinical trials with results manually. For standardized interventions/comparisons, we have performed Concept Unique Identifiers [16] [16] number mapping, and those that cannot be mapped are indicated by their original description. A key goal of Metathesaurus construction is to understand the intended meaning of each name in each source vocabulary and to link all the names from all of the source vocabularies that mean the same thing (the synonyms). Each CUI contains the letter C followed by seven-digit numbers https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html.

Since clinical trials that have not yet found results are only at the design stage, we only mapped the outcomes of 71 studies through the MetaMap tool and standardized them with manual

work. In order to improve the visualization, all the unified standardized names first to be select to show is the abbreviation, and the detailed description in the field is contained in the nodes.

2.4 Data visualization

Graph database query language is indispensable for medical information queries. The Clinical Quality Language (CQL) [19] is a useful tool for defining search requests for data stores containing FHIR data. There are only few execution engines that are able to evaluate CQL queries as FHIR data represents a graph structure.

We built the graph database in Neo4j 4.25. The Neo4j graph database and its query language, Cypher, provide efficient access to the complex Reactome data model, facilitating easy traversal and knowledge discovery. Cypher is a graph database query language that is easy to understand and does not require any deep programming knowledge [19]. Therefore, based on previous research experience, we used information from ClinicalTrials.gov for text mining, information extraction and the establishment of graph database in neo4j by Cypher.

At present, many researchers make a great effort in the establishment of graph database in Neo4j, such as the developed applications to link Cytoscape and Neo4j by Summer et al. and providing a high-performance pathway data resource to the community through adopting graph database technology by Fabregat et al. [10, 18].

3 Results

3.1 The Clinical Trials Results reported in registered platform vs. in bibliographic database

We collected clinical trial studies related to COVID-19 at https://clinicaltrials.gov/ on August 3, 2021. A total of 6279 studies were collected, of which 71 studies have results. As we all know, the publication of bibliographic databases is delayed compared with clinical trials. In order to find the time difference between COVID-19-related clinical trials and publications, we searched 71 studies of registered clinical trial numbers on the Dimensions platform. 59 publications (containing 4 not open access studies, 8 not report the results) related to 35 studies in 71 registered clinical trial were retrieved. The completion time of clinical trials is on average (79.81 ± 125.73) days earlier than the bibliographic publications.

3.2 Knowledge Graph Information

Totally, there are 6279 registered clinical trials related to COVID-19 included in our research. Among them, 6208 (98.87%) studies have not yet registered results. So we generated two knowledge graphs. One contains all 6279 registration studies related to COVID-19, which is used to summarize the latest clinical trial studies to help researchers understand the trial progress and design plan. The other one includes 71 clinical trials that have registered results. Based on the framework of the first graph, we added fields for each intervention/comparison's group name, baseline data, and post-test data to the outcome node in the second graph. The grouping of each intervention is clearer than the enrollment node. This data helps researchers to generate evidence synthesis based on clinical trial data from studies screened by the same conditions.

The first COVID-19 Trial Knowledge Graph, which contains 66856 nodes with 10 types and 1217673 relationships with 9 types in total. The second is COVID-19 Registered Results Knowledge Graph contains 1067 nodes with 12 types and 1405 relationships with 13 types in total. The statistics of nodes and relationships can be seen in Tables 1 and 2, respectively. The graph data files and codes are public available at: https://github.com/baiym13/COVID-19-Trial-Knowledge-Graph/find/main.

Table 1: Node types and the number of unique entities in Knowledge Graph

	KG1	KG2
clinical trial	6279	71
condition	1395	104
intervention	5730	98
comparison	7168	50
outcome	44784	537
funding	13	3
enrollment	1049	58
gender	3	3
age	426	22
study tyoe	9	2
start date	—	58
completion date	—	63

Table 2: Relationship types and the number of unique relationships in Knowledge Graph

	KG1	KG2
has condition	10900	104
has intervention	67184	98
has comparison	1059270	50
has outcome	44784	537
funded by	6279	71
enrollment	6285	71
gender	6267	71
age	10424	71
studytype	6279	71
start date	—	71
completion date	—	71
has significant	—	69
no difference	—	50

Conditions: There are 1395 nodes of conditions in 6279 studies, of which only 1278 have DUIs. We use the original description to supplement the conditions that are not mapped by the MTI tool. We used 0 to supplement the missing values of standard interventions and DUI in our knowledge graph. Some clinical trials regard the

occurrence of COVID-19 as the clinical outcome of certain high-risk groups. When the high-risk situation is refers to a disease, COVID-19 cannot be used as one of the complications. Therefore, in the list of conditions, the complications of the disease are not distinguished. There are 127 descriptions of conditions in 71 studies (the same concept is not de-duplicated), all of which have DUIs. After deduplicating 23 identical concepts in the same clinical trial, a total of 104 nodes are left about conditions in the knowledge graph. The standardization method is the same as that of 6279 clinical trials. A total of 26 standard condition-concepts were obtained after deduplication.

Intervention/Comparison: We use MetaMap to standardize intervention/comparison terms. The standardized results were mapped to CUI. The results that cannot be mapped to the CUI used the original description. The problem of combining measures has been solved by manual verification, such the same drugs Favipiravir, Avigan and Aavilavir were merged to one. The standardized results preferentially use the abbreviation from MetaMap mapping through manual screening.

Outcomes: The text of the original outcome description cannot be mapped to a single field in MetaMap, and there may be multiple entity types in the same field. Therefore, in this study, we only extracted the standardized results of the outcome of 71 clinical trials with registered results, taking into account the complexity of the mapping results. Field in outcome: the empty value in the field, Time frame, is filled by 0. Other fields involve numerical values, so NA is used to fill in missing values.

Other nodes: funding, enrollment, gender, age, study type, start date, completion date are extracted directly from the download XML files.

In addition to the relationship directly displayed in the structured data of clinical trials, we have established a new relationship between the node intervention/comparison and the outcome, filtered based on the P-value extracted from the XML file. The established relationships are “has significant” or “no difference”. This is used to help researchers understand the effects of current measures on specific outcomes more intuitively.

3.3 Research Visualization

The knowledge graph below shows the basic information (age groups, genders, study types, start dates, completion dates, conditions, interventions, outcomes) of the 71 clinical trials with registered results. For visualization purposes, this graph only shows the top 300 nodes. The nodes at the center of the graph have more relationships with other nodes. Several obvious clusters appear on the graph, which provides an overview of the 71 clinical trials results. The main age group of enrollments is 18 and older, with 11 clinical trials including children in the study population. Most of the studied populations are on both male and female, while 4 studies focus on the population of male and one study only focus on the population of female. The number of participants lies between 1 and 20460. The majority (63 out of 71) of studies are interventional, and the rest are observational. More than half of the clinical trials started in the period from March to May 2020 and were finished in the period from May 2020 to May 2021. Among all the studies with results, National Institutes of Health (NIH) has funded 5 studies, related industries have funded 23 studies, and the rest of the studies are funded by others, including individuals, universities and organizations). Most clinical trials focus solely on COVID-19, while 5 studies respectively investigated the relationship between COVID-19 and androgenetic alopecia, pregnancy, neoplasm, influenza and adolescent depression. As for interventions, among the 300 nodes, two drug interventions in three different trials were found to have significant difference on the experimental groups, which are Avigan and HCQ (Figure 3).

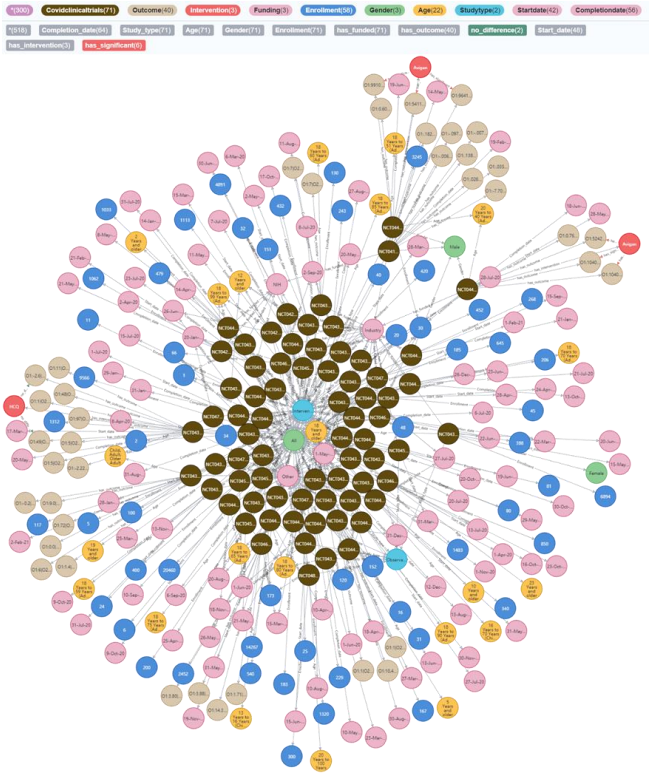

Figure 3: Knowledge graph related to COVID-19

3.4 Case Query

3.4.1 Nodes information of Clinical trial knowledge graph An example of a single clinical trial query is shown in Figure4. By using Cypher, the study record of 'NCT04452435' is extracted. The general information of the study and the population features are connected directly to the clinical trial node. In this clinical trial, the intervention is C21 (an experimental drug) and the comparison is placebo, therefore, the relationship between the clinical trial and intervention is captured by a directed connection from the trial node to the intervention node, named as has_intervention. Similarly, the comparison node is connected to the clinical trial node with a directed relationship, named as has_comparison. In this clinical trial, the intervention measure was found to take effect on the experimental groups, resulted in statistical significance on several outcomes. Therefore, on the graph, the three red relationships, named as has_significant, indicate that the intervention, C21, led to a significant difference on the three outcomes compared to the

comparator group. Similarly, by examining the p-value, the intervention was also found to have no significant difference on the other 8 outcomes. In these case, the green lines, named as no_difference, are used to formulate this relationship. The last outcome, oxygen (supply), has no connections with the intervention, which means no statistical information can be extracted to deduct their relationships and no connection is drawn (Figure 4).

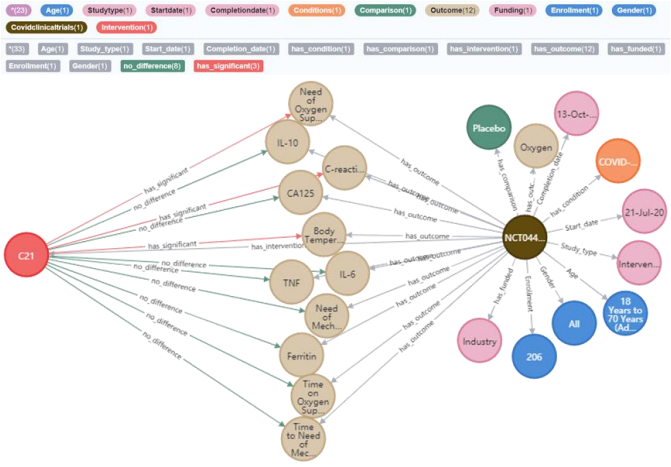

Figure 4: Node information of Clinical trial knowledge graph

3.4.2 Case query for outcome data Taking “c-reactive protein” as an example, we use the Cypher language to query all the outcome nodes with the standardized name "CRP". The result shows a total of 8 CRP nodes. Through the expansion of the node-related relationship, it can be seen that there are to 6 clinical trials, 3 interventions and 1 comparison. And it contains 4 relationships “has significant” (Figure 5).

Cypher query command:

MATCH (n{standard_outcome:'CRP'}) RETURN n;

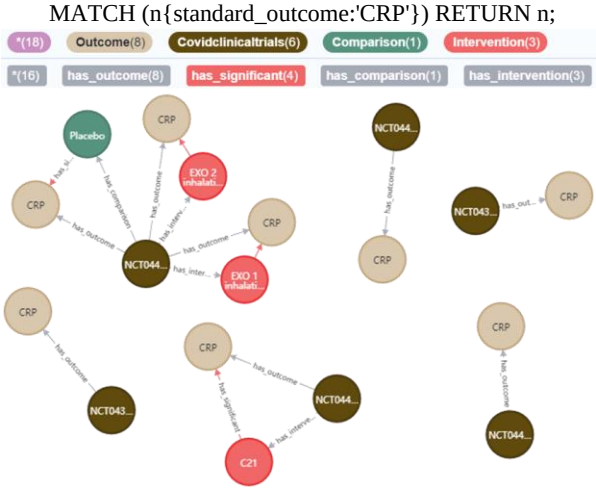

Figure 5: Result of the query for specific outcomes

All the node information obtained by the query can be exported as a csv file or a Json file in neo4j. This study takes all pieces of information in Json format in the CRP query result as an example (Figure 6). The intervention/comparison measures are numbered “O+number”in the XML files. The number of repeated occurrences of the fields “group info”, “participate group num” and “outcome group num” corresponds to the “time frame” in order.We parse the Json format to export operable data through self-compiled code.

Figure 6: Information of outcome nodes example

3.4.3 Relationship Query One feature of our knowledge graph that is different from the past is the establishment of a relationship between intervention/comparison and outcome. The purpose of this is to help researchers find all the measures that can or cannot improve the outcome of the disease in a particular trial. According to the automatically obtained P-value, the relationship “has_significant” or “no_difference” are defined according to the 95% test level. Taking the "has_significant" query as an example, the result is shown in Figure.7. Through standardizing intervention/comparison, the number of various nodes can be counted and the measures that affect COVID-19 most can be found. The research purpose is well practiced (Figure 7).

Cypher query command:

MATCH p=()-[r:has_significant]->() RETURN p LIMIT 300

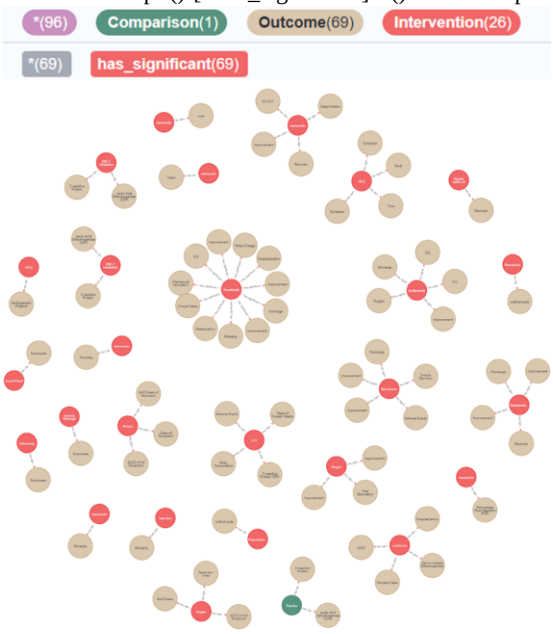

Figure 7: Result of the query for relationship

4 Discussion and Conclusion

Meta-analyses are increasingly used to address the evidence synthesis problem. But they filter out a lot of information during the data processing.[20]. Evidence-based medicine is a labor-intensive task, and its operation process determines that the results of evidence synthesis are delayed. The latest bibliographic publications are difficult to be included in the latest evidence synthesis results.

Instead of using unstructured claims in scientific publication, our work validated the idea of "computable evidence synthesis" via presenting prespecified PICO data elements results data in trial registries in standardized, structured formats with controlled vocabularies. We used COVID-19 as a case in our research. By parsing the XML file, more detailed information can be obtained than the csv or txt format downloaded from Clinicaltrials.gov. In addition to the necessary elements contained in PICO, we also extracted data from the clinical trials results. To help form the basis of computable medical evidence. Query and batch export information in Graph Database built by neo4j through Cypher language. It can help researchers obtain the latest data in batches and form a basis for the synthesis of real-world research evidence. Compared with publications in bibliographic database, these data include negative and positive outcomes. More comprehensive and objective. Our methodology is also generalizable to other conditions, such as cancer clinical trials. we will incorporate registered clinical trial data from more platforms to achieve field unification of multi-source heterogeneous data. Developing more visualized and knowledge graphs of disciplines or diseases in our future research.

ACKNOWLEDGMENTS

This work was funded by Chinese Scientific and Technical Innovation Project 2030（2018AAA0102100), the National Natural Science Foundation of China (71603280, 72074006), Peking University Health Science Center and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2017QNRC001).

REFERENCES

[1] Ravaud, P., Crequit, P., Williams, H. C., Meerpohl, J., Craig, J. C. and Boutron, I. Future of evidence ecosystem series: 3. From an evidence synthesis ecosystem to an evidence ecosystem. *Journal of Clinical Epidemiology*, 123 (Jul 2020), 153-161.

[2] Atal, I., Zeitoun, J. D., Neveol, A., Ravaud, P., Porcher, R. and Trinquart, L. Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries. *BMC Bioinformatics*, 17 (Sep 2016), 14.

[3] Huang, Z., ten Teije, A. and van Harmelen, F. *SemanticCT: A Semantically-Enabled System for Clinical Trials*. Springer International Publishing, City, 2013.

[4] Hu, Q., Huang, Z. and Gu, J. Semantic representation of evidence-based medical guidelines and its use cases. *Wuhan University Journal of Natural Sciences*, 20, 5 (2015/10/01 2015), 397-404.

[5] Mavergames, C., Oliver, S. and Becker, L. *Systematic Reviews as an Interface to the Web of (Trial) Data: using PICO as an Ontology for Knowledge Synthesis in Evidence-based Healthcare Research*. City, 2013.

[6] Mavergames, C., Beecher, D., Becker, L. A. and Ali, A. Cochrane's Linked Data Project: How it Can Advance our Understanding of Surrogate Endpoints. *journal of law medicine & ethics*, 47, 3 (9/27/2019 2019), 374-380.

[7] Marshall, I. J., Nye, B., Kuiper, J., Noel-Storr, A., Marshall, R., Maclean, R., Soboczenski, F., Nenkova, A., Thomas, J. and Wallace, B. C. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association : JAMIA*, 27, 12 (Dec 9 2020), 1903-1912.

[8] Mayer, T., Marro, S., Cabrio, E. and Villata, S. Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials. *artificial intelligence in medicine* (5/7/2021 2021).

[9] Pradhan, R. and Singh, S. Comparison of Data on Serious Adverse Events and Mortality in ClinicalTrials.gov, Corresponding Journal Articles, and FDA Medical Reviews: Cross-Sectional Analysis. *Drug Saf.*, 41, 9 (2018/09/01 2018), 849-857.

[10] Summer, G., Kelder, T., Ono, K., Radonjic, M., Heymans, S. and Demchak, B. cyNeo4j: connecting Neo4j and Cytoscape. *Bioinformatics*, 31, 23 (Dec 2015), 3868-3869.

[11] Dunn, A. G. and Bourgeois, F. T. Is it time for computable evidence synthesis? *Journal of the American Medical Informatics Association : JAMIA*, 27, 6 (Jun 1 2020), 972-975.

[12] Du, J., Wang, Q., Wang, J., Ramesh, P., Xiang, Y., Jiang, X. and Tao, C. COVID-19 Trial Graph: A Linked Graph for COVID-19 Clinical Trials. *Journal of the American Medical Informatics Association : JAMIA* (Apr 24 2021).

[13] Yuan, C., Ryan, P. B., Ta, C., Guo, Y., Li, Z., Hardin, J., Makadia, R., Jin, P., Shang, N., Kang, T. and Weng, C. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association : JAMIA*, 26, 4 (Apr 1 2019), 294-305.

[14] Wang, H. D., Abbas, K. M., Abbasifard, M. and al., e. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950-2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. *Lancet*, 396, 10258 (Oct 2020), 1160-1203.

[15] Dodd, S., Clarke, M., Becker, L., Mavergames, C., Fish, R. and Williamson, P. R. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of Clinical Epidemiology*, 96 (Apr 2018), 84-92.

[16] Pan, X. L., Yan, E. J., Cui, M. and Hua, W. N. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. *J. Informetr.*, 12, 2 (May 2018), 481-493.

[17] Henkel, R., Wolkenhauer, O. and Waltemath, D. Combining computational models, semantic annotations and simulation experiments in a graph database. *Database : the journal of biological databases and curation*, 2015 (2015 2015).

[18] Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P. P., Wu, G. M., Stein, L., D'Eustachio, P. and Hermjakob, H. Reactome graph database: Efficient access to complex pathway data. *PLoS Comput. Biol.*, 14, 1 (Jan 2018), 13.

[19] Fette, G., Kaspar, M., Liman, L., Ertl, M., Krebs, J., Stork, S. and Puppe, F. *Implementation of a HL7-CQL Engine Using the Graph Database Neo4J*. City, 2019.

[20] Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T.

A., Thacker, S. B. and Grp, M. Meta-analysis of observational studies in epidemiology - A proposal for reporting. *JAMA-J. Am. Med. Assoc.*, 283, 15 (Apr 2000), 2008-2012.