北京大学 健康医疗大数据国家研究院
NATIONAL INSTITUTE OF HEALTH DATA SCIENCE AT PEKING UNIVERSITY

A PICO-based Knowledge Graph for Representing Clinical Evidence

Yongmei Bai: baiym@bjmu.edu.com

Jian Du: dujian@bjmu.edu.com

September 30, 2021

School of Public Health, Peking University, No. 38 Xueyuan Road, Beijing (China);
National Institute of Health Data Science, Peking University, No. 38 Xueyuan Road, Beijing (China)

Outline

1. **Introduction**
2. **Materials and Methods**
3. **Results**
4. **Discussions & Conclusions**
5. **Future work**

Introduction

Data sources

- We collected 6279 clinical trials about COVID-19 in ClinicalTrials.gov as the raw data, including 71 pieces of research with results. We completed the information extraction, structured data standardization, visualization and query work were carried out in a semi-automated manner.

Introduction

The specific content

- 1) Automatically obtaining clinical trial data from registration platform.
- 2) Visualizing the structured data of the clinical trial registration platform by using a knowledge graph.
- 3) Standardize indicators to realize the searchability of medical knowledge.

Introduction

The function of graph database

Our graph database enables researchers to query and obtain clinical registration trial information related to the ‘PICO framework’ designed by themselves in batches. And display it in a visual form.

Materials and Methods

ClinicalTrials.gov

For a long time, clinical evidence is predominantly disseminated in unstructured, natural language scientific publications that describe the results of randomized control trials (RCTs).

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 (ie, comprehensive, based on rigorous methods, and up-to-date).

Materials and Methods

ClinicalTrials.gov

It is difficult to quickly integrate the same high-quality RCT research of PICO. 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.

Previous studies have shown that the information from the **clinical trial registration platform** can help solve this problem.

Materials and Methods

PICO

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.).

Data Collection and Processing

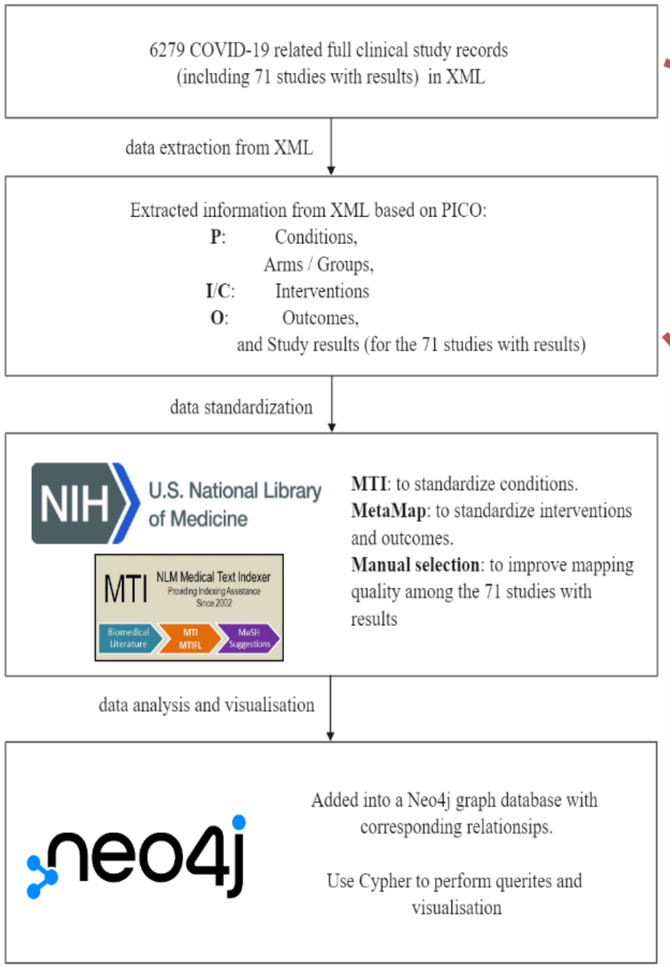
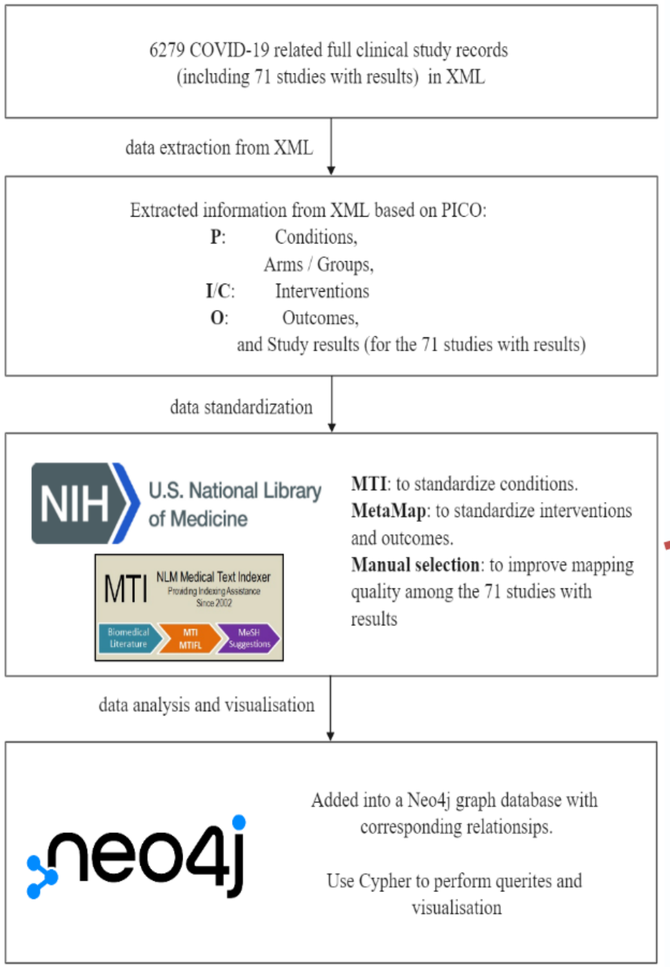


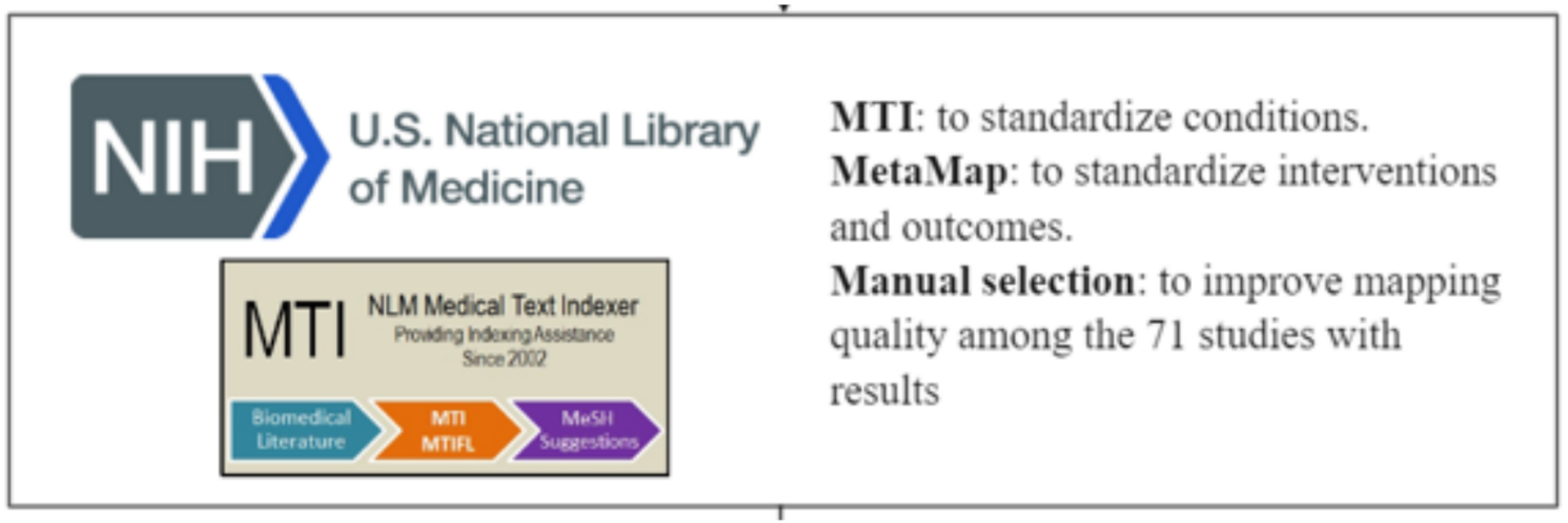
Fig.1 Flowchart of the research

6279 COVID-19 related full clinical study records
(including 71 studies with results) in XML
data extraction from XML
Extracted information from XML based on PICO:
P: Conditions, Arms / Groups,
I/C: Interventions
O: Outcomes,
and Study results (for the 71 studies with results)

Our study 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. 71 studies with results added two fields on this basis, Start Date and Completion Date.

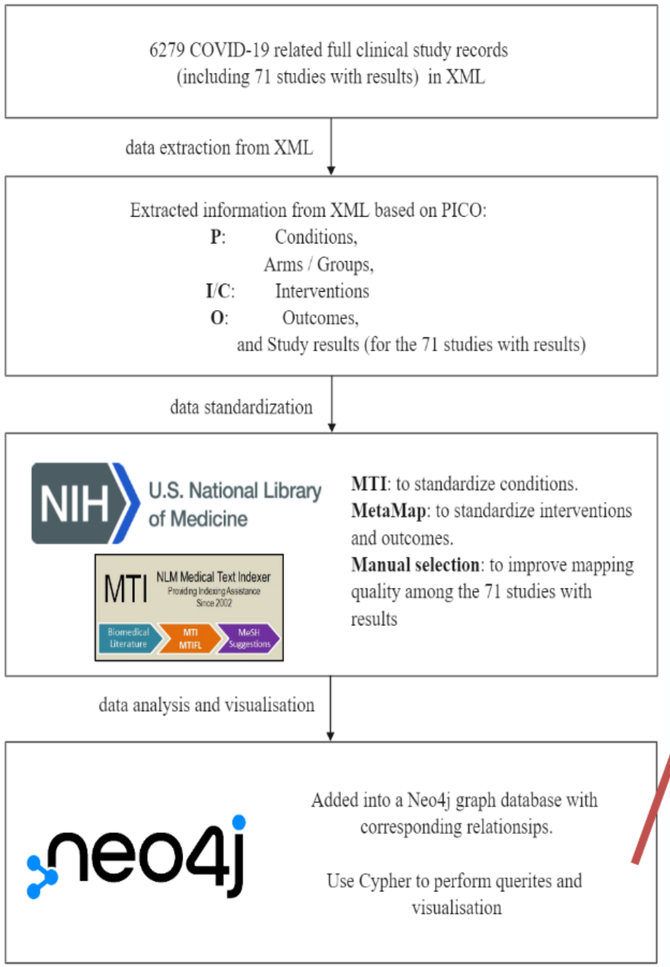
Data Collection and Processing

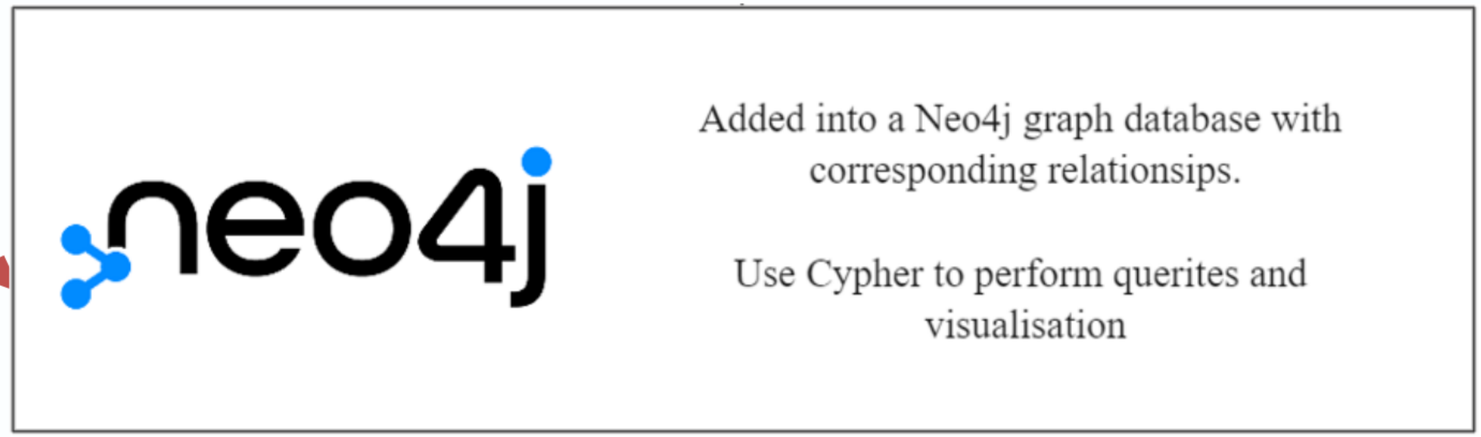




- **The Medical Text Indexer (MTI) produced by the NLM was adopted to extract standardized medical subject heading terms in 6279 clinical trials conditions.**
- **We standardized the intervention/comparison of 6279 clinical trials by using MetaMap.**
- **Mapped the outcomes of 71 studies through the MetaMap and standardized them with manual work.**

Data Collection and Processing





Graph database query language is indispensable for medical information queries.

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.*

```
- <arm_group>
    <arm_group_label>Treatment</arm_group_label>
    <arm_group_type>Experimental</arm_group_type>
    <description>Participants in this arm will receive the study drug.</description>
  </arm_group>
- <arm_group>
    <arm_group_label>Placebo</arm_group_label>
    <arm_group_type>Placebo Comparator</arm_group_type>
    <description>Participants in this arm will receive a placebo treatment.</description>
  </arm_group>
- <intervention>
    <intervention_type>Drug</intervention_type>
    <intervention_name>Hydroxychloroquine</intervention_name>
    <description>200mg tablet; 800 mg orally once, followed in 6 to 8 hours by 600 mg, then
      600mg once a day for 4 consecutive days</description>
    <arm_group_label>Treatment</arm_group_label>
    <other_name>Plaquenil</other_name>
  </intervention>
- <intervention>
    <intervention_type>Other</intervention_type>
    <intervention_name>Placebo</intervention_name>
    <description>4 placebo tablets once, followed in 6 to 8 hours by 3 tablets, then 3 tablets
      once-a-day for 4 consecutive days</description>
    <arm_group_label>Placebo</arm_group_label>
  </intervention>
```

```
- <primary_outcome>
    <measure>Change in Disease Severity Over 14 Days Among Those Who Are
      Symptomatic at Baseline</measure>
    <time_frame>baseline and 14 days</time_frame>
    <description>Visual Analog Scale 0-10 score of rating overall symptom severity (0 = no
      symptoms; 10 = most severe)</description>
  </primary_outcome>
- <secondary_outcome>
    <measure>Rate of Hospitalization</measure>
    <time_frame>14 days</time_frame>
    <description>Outcome reported as the number of participants in each arm who require
      hospitalization for COVID19-related disease.</description>
  </secondary_outcome>
```

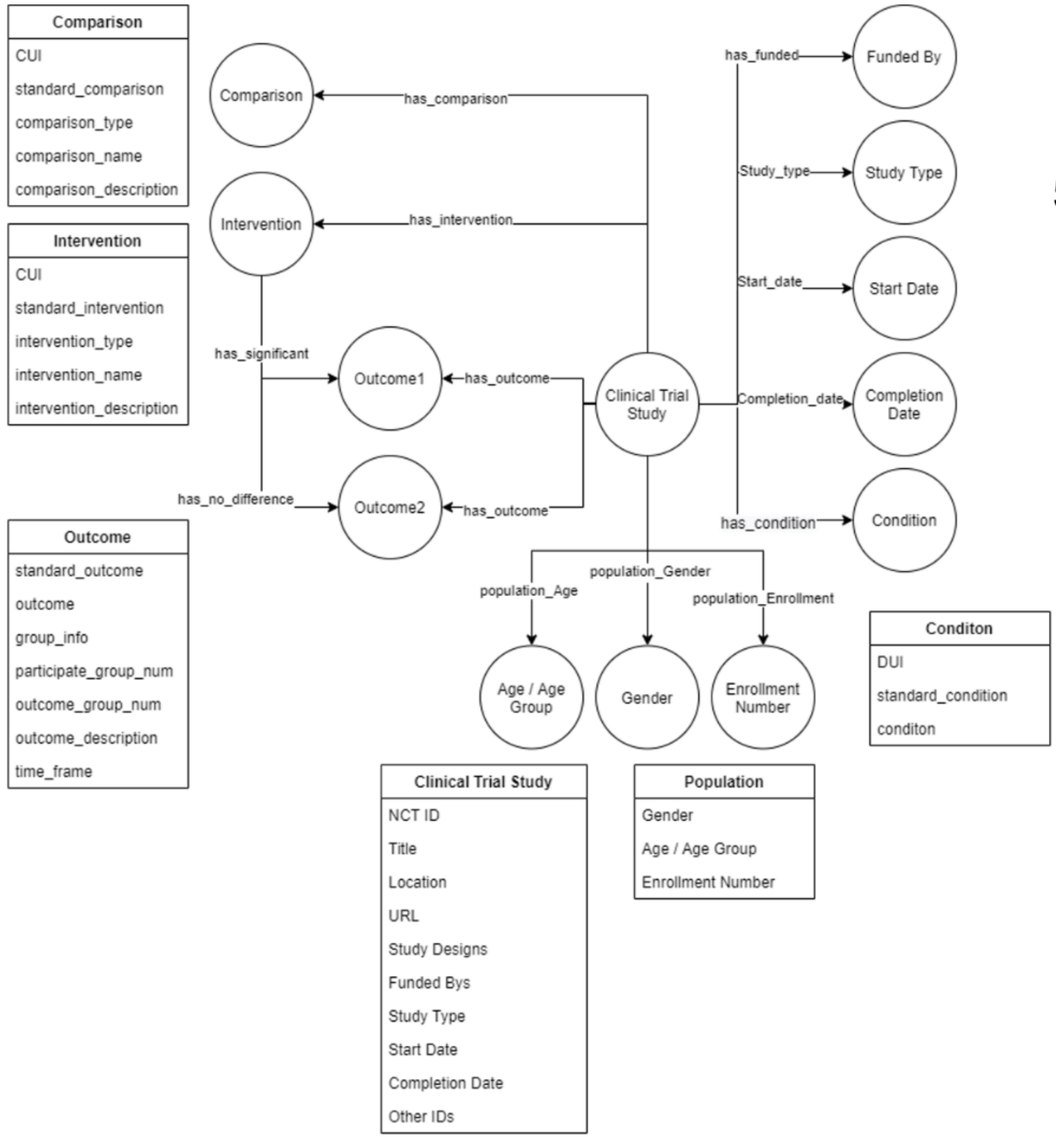


Figure 2: The entities and relationships represented in our knowledge graph

5 Categories:

1) Clinical trials study information: including 10 fields study information.
2) Conditions.
3) Population Feature.
4) Intervention/ comparison.
5) Outcomes: We extracted all the outcomes from

Knowledge Graph Information

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

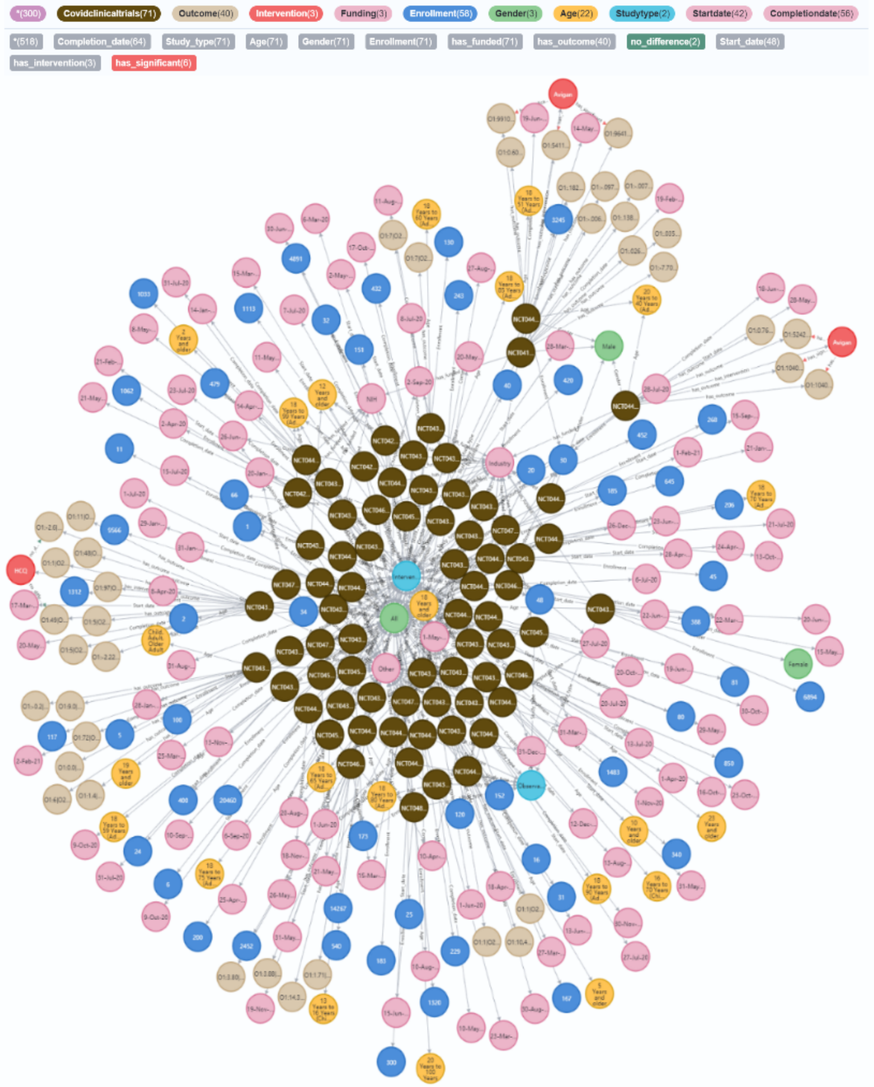
	KG1	KG2
clinical trial	6279	71
condition	1395	104
intervention	5730	98
comparison	7168	50
outcome	44784	537
funding	13	3
enrollment	1049	58
gender	3	3
age	426	22
study tyoe	9	2
start date	–	58
completion date	–	63

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

KG1 is the Knowledge Graph including 6279 studies. KG2 is the Knowledge Graph including 71 studies with results.

Results



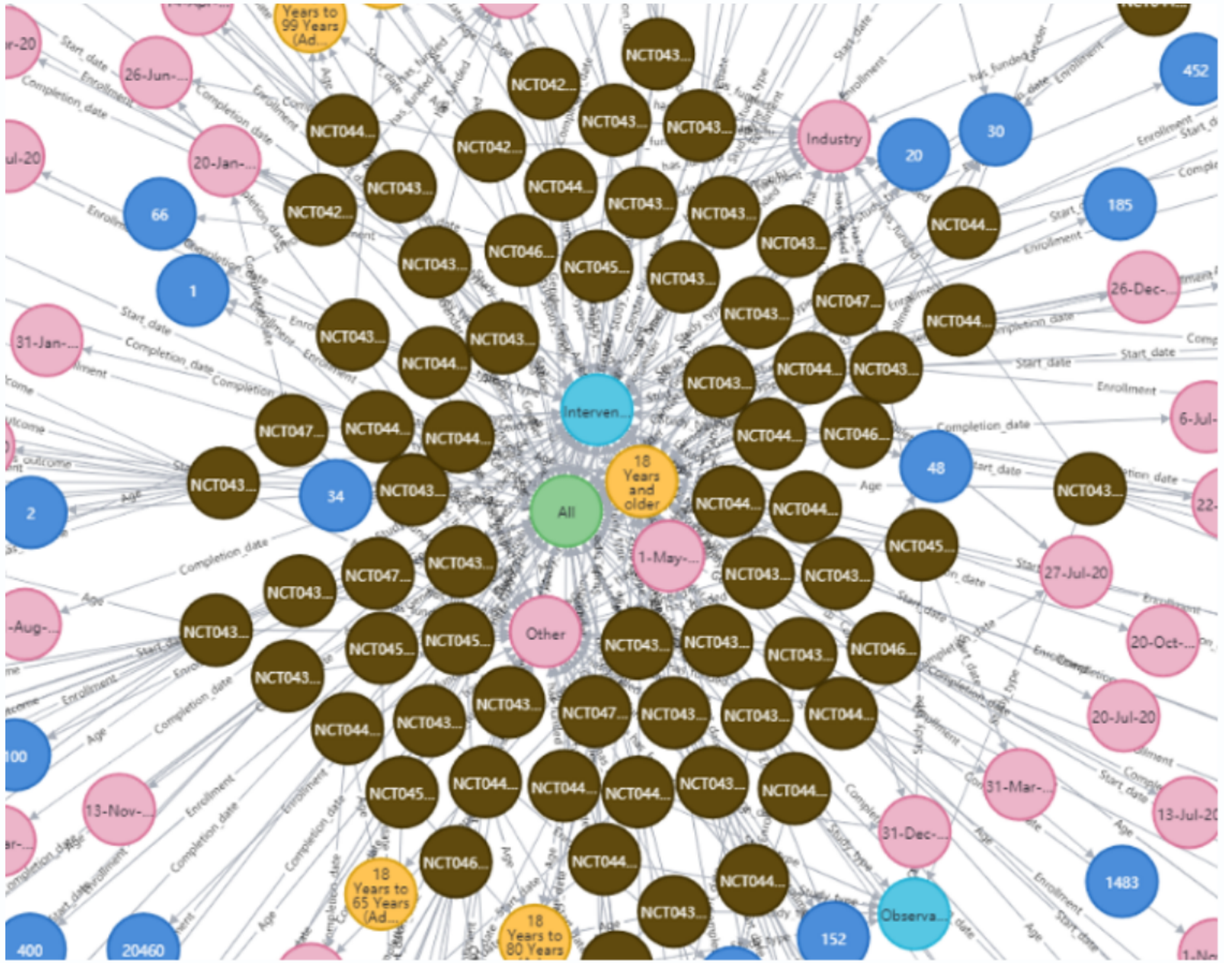


Figure 3: Knowledge graph related to COVID-19

Results

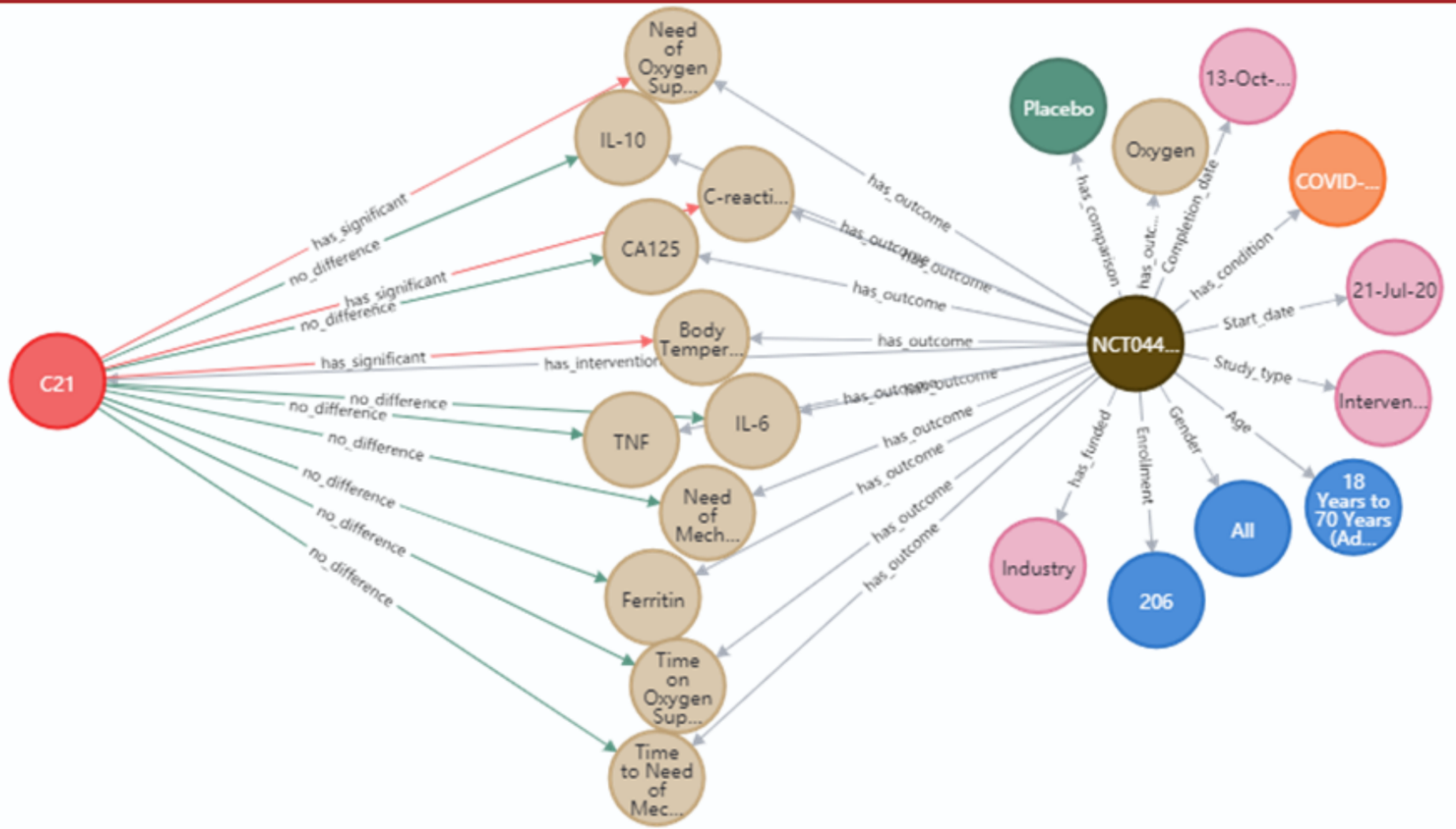


Figure 4: Node information of Clinical trial knowledge graph

Results

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

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. There are 6 clinical trials, 3 interventions and 1 comparison. And it contains 4 relationships “has significant”

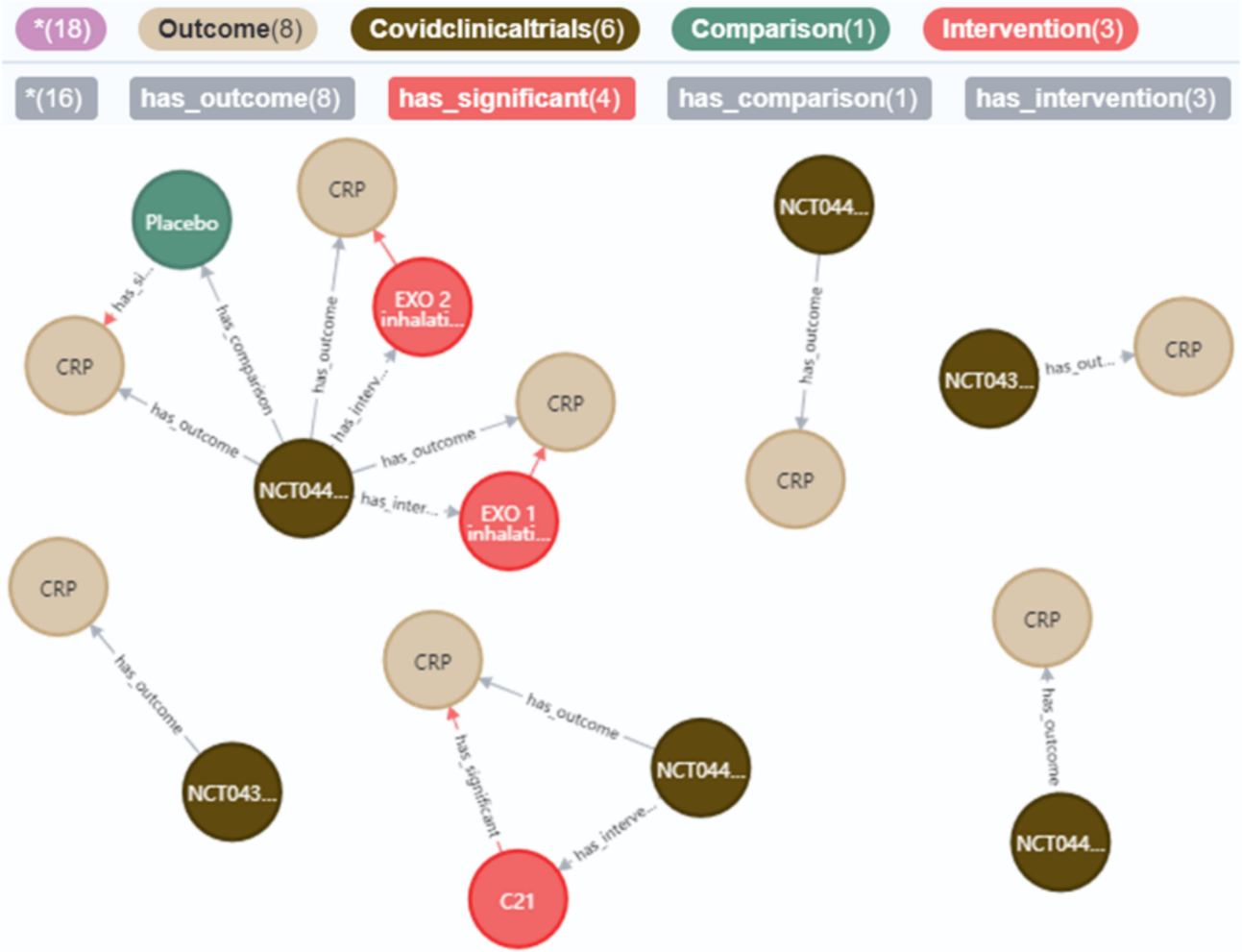


Figure 5: Result of the query for specific outcomes

Results

OUTCOME_ID	NCTID	TIME_FRAME	OUTCOME_DESCRIPTION	GROUP_INFO_LIST	PARTICIPATE_GROUP_NUM	OUTCOME_GROUP_NUM
OUTCOME_1	NCT04452435	Time Frame: Day 1 through Day 29	···	C21 Treatment|Placebo Treatment	O1:45|O2:46	O1:0.19|O2:0.22
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:446|O2:455	
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:446|O2:455	O1:-23.035|O2:-18.671
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:368|O2:367	O1:-58.935|O2:-30.908
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:241|O2:250	O1:-78.411|O2:-62.038
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:159|O2:181	O1:-103.789|O2:-88.881
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:221|O2:240	O1:-122.339|O2:-112.588
OUTCOME_113	NCT04401579	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Remdesivir Plus Baricitinib|Remdesivir Plus Placebo	O1:217|O2:219	O1:-131.333|O2:-122.342
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:88|O2:90	
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:77|O2:79	O1:56.0|O2:62.8
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:80|O2:74	O1:36.1|O2:43.7
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:78|O2:76	O1:21.7|O2:26.1
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:74|O2:71	O1:13.9|O2:26.3
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:69|O2:73	O1:15.8|O2:38.3
OUTCOME_181	NCT04335136	NA	···	Group A (Active) APN01|Group B (Placebo Control)	O1:36|O2:37	O1:4.9|O2:8.5
OUTCOME_343	NCT04491240	Time Frame: 14 days	···	EXO-1|EXO-2|Placebo	O1:10|O2:10|O3:10	O1:78.3|O2:75.4|O3:61.5
OUTCOME_344	NCT04491240	Time Frame: 14 days	···	NA	NA	NA
OUTCOME_345	NCT04491240	Time Frame: 14 days	···	NA	NA	NA
OUTCOME_377	NCT04475588	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Arm A - Itolizumab + BSC|Arm B - Best Supportive Care (BSC)	O1:20|O2:10	
OUTCOME_377	NCT04475588	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Arm A - Itolizumab + BSC|Arm B - Best Supportive Care (BSC)	O1:18|O2:8	O1:-61.69|O2:-103.6
OUTCOME_377	NCT04475588	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Arm A - Itolizumab + BSC|Arm B - Best Supportive Care (BSC)	O1:16|O2:5	O1:-81.65|O2:-107.2
OUTCOME_377	NCT04475588	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Arm A - Itolizumab + BSC|Arm B - Best Supportive Care (BSC)	O1:11|O2:3	O1:-90.99|O2:-127.5
OUTCOME_377	NCT04475588	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Arm A - Itolizumab + BSC|Arm B - Best Supportive Care (BSC)	O1:3|O2:2	O1:-103.2|O2:-127.6
OUTCOME_407	NCT04323592	Time Frame: Days 1, 3, 5, 8, 11, 15 and 29	···	Exposed to Methylprednisolone|Non-exposed to Methylprednisolone	O1:83|O2:90	O1:-82.08|O2:-34.34

Figure 6: Information of outcome nodes example

Discussions & Conclusions

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.

Discussions & Conclusions

- 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 CT.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

Future work

- Incorporating 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 the future research.
- Results of clinical trial evidence synthesis vs. Bibliographic database evidence synthesis——Grey data

Acknowledgement

Thank for

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)

Thanks!

National Institute of Health Data Science, Peking University, Beijing, China

Yongmei Bai: baiym@bjmu.edu.com
Jian Du: dujian@bjmu.edu.com

北京大学 健康医疗大数据国家研究院
NATIONAL INSTITUTE OF HEALTH DATA SCIENCE AT PEKING UNIVERSITY