A Semantic Cohort Builder for Clinical Trials by Generating Knowledge graphs from an Enterprise EMR Lakehouse

Amna Basharat [*1 2] **Muhammad Uzair** [*1 2] **Fatima Maryam** [*1 2]

Abstract

Selecting a relevant patients' cohort is an expensive prerequisite for conducting evidence based clinical research. To this end, research efforts are focused on using electronic health records (EHRs) based systems. The challenge though is: records are typically stored on distributed servers in a public or private cloud. Moreover, EHRs contain a mix of structured, unstructured, and semistructured data; therefore, searching for eligible patients in this data swamp is significantly inefficient and prone to errors. Consequently, enterprises are adopting lakehouses to store processed information. The major contribution of this paper is a set of novel semantic models to enable semantic search over the processed data by interconnecting different data sources, stored in a lakehouse, by using Knowledge graphs (KGs). We present the design of a novel cohort retrieval system, satisfying inclusion and exclusion criteria of a clinical study, that utilizes a semantics driven dynamic query engine to generate and execute cohort selection queries on heterogeneous data. For evaluation, we employ a case study based approach that utilizes cohort selection queries to assess the ability of our system in searching for patients that meet the eligibility criteria. We demonstrate through experimentation that our KG driven semantic cohort builder performs effectively in terms of eligibility criteria

Keywords: semantic · knowledge graph · cohort selection · EHR · lakehouse · clinical trial

[*]Equal contribution [1]FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan [2]. Correspondence to: Amna Basharat <(amna.basharat,i222819,i181573)@nu.edu.pk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Cohort selection for specialized clinical trials is a cardinal pillar of the evidence based processes in medicine that determine the efficacy and effectiveness of new treatments and interventions. But, it is also considered as one of the most difficult, complex, time consuming, and expensive step in the process (11). Determining the efficacy of a new treatment requires patients' population that must satisfy a predefined inclusion and exclusion criteria for a clinical trial(11). In specialized scenarios, the complex search criteria may even require researchers to do time consuming manual reviews and analyses of electronic health records (EHRs) – stored across different and distributed subsystems – to identify qualified patients. The other challenge is that the data include structured tables, such as laboratory results and medications, and unstructured text such as admission notes, radiology reports, progress notes, and discharge summaries. In order to bring some order in these data swamps, data lake technologies are gaining significant attention in eHealth ecosystems that will help in identifying the most relevant sources for the given cohort selection criteria. State-of-the-art research is advocating the use of knowledge graphs for managing lakehouse data (7). This is possible, as knowledge graphs provide powerful abstraction for enabling semantics driven search, knowledge discovery and retrieval. Various querying approaches over knowledge graphs are reported in the literature such as (13) including the ones that are relevant for data lakes (2).

Recent cohort selection approaches such as those described in (10) match patients to multi-factor inclusion and exclusion criteria. Researchers have utilized a combination of pattern based, knowledge intensive, and feature weighting techniques for the purpose (12). In contrast, others have explored deep learning approaches and concluded that the hybrid models performed better in the realm of using NLP for cohort selection (8). Some have reported that unstructured data contains more relevant information on patients as compared to the structured data, and for searching information in them to select cohorts, database queries are insufficient (4). Ontology-based query answering techniques leverage description logic to achieve this objective (1). The use of

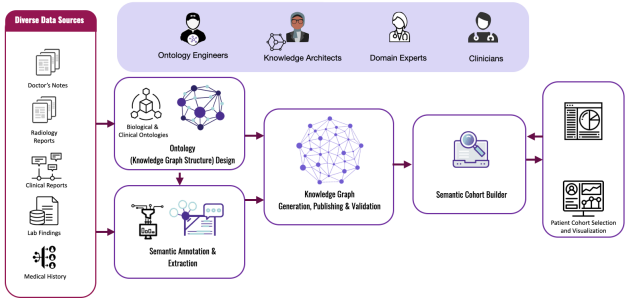

Figure 1. The High level KG Architecture Supporting the Semantic Cohort Builder.

Knowledge graphs in healthcare domain is proving to be promising because of their ability to unify and aggregate disparate data sources (3). Emerging research approaches are automating the process to learn high quality knowledge bases that link diseases and symptoms directly from EHR systems (6). Moreover, their use in personalized health, disease specific knowledge, and for gaining insights from hidden (apparently unrelated and unused) data is also getting attention (9).

We believe that by using knowledge graphs (KGs), we can improve and enhance the cohort retrieval process by interconnecting different data sources, stored in the gold zone of a lakehouse, and developing semantics based models that enable semantic search over the processed data. As compared to existing prior art, that either focus primarily on structured or unstructured data sources, we employ a semantics driven approach towards unification of this heterogeneous data to support the cohort selection process for clinical trials. While the approach we propose is generic, we have validated it by focusing on oncology trials and evaluating its effectiveness on the real patients' data that is obtained from an EMR based lakehouse. In comparison to similar efforts of semantically integrating data for clinical trials (5) that utilize RDF graphs, our approach utilizes state of the art LPG based knowledge graphs for achieving efficiency at scale.

We present the following major contributions in this paper: (1) the design and development of a high level knowledge graph driven architecture that enables semantics driven cohort selection (Section 2); (2) the design of a novel cohort retrieval system, satisfying the inclusion and exclusion criteria of a clinical study, that utilizes a semantics driven dynamic query engine to generate and execute cohort selection queries on heterogeneous EHR data stored in a lakehouse; (3) a case study based approach that utilizes cohort selection queries to assess the ability of our system in searching for patients that meet the eligibility criteria. We demonstrate through experimentation that our KG driven semantic cohort builder performs effectively in terms of satisfying the eligibility criteria .

Finally, we conclude the paper with an outlook for future research.

2. KG Driven System Architecture for Semantic Cohort Builder for Clinical Trials

Figure 10 shows the overall KG driven system architecture. Semantic annotation and data extraction is done on the unstructured patients' notes to extract useful information that is not available in the structured data. The knowledge graph structure is driven by ontologies and an automated knowledge graph generation pipeline is created to semantically unify the structured data with the annotations obtained from the unstructured data. A semantic cohort builder is then built on top of this knowledge

graph, which interacts with the structured query builder and the cohort visualization module at the user interaction layer. In the subsequent subsections, the approach for knowledge graph construction and the cohort builder is discussed in more detail.

Knowledge Graph Schema and Construction. In this research, two types of knowledge graphs for the field of oncology – data-driven and ontology augmented – are constructed and used. The first type contains only the information of the patients that is stored in the structured tables and extracted from unstructured patients' notes. The second type is ontology

augmented, which contains the NCIt ontology [1], the ICD10 diagnosis [2] , and the GPI [3] , and the relevant ontology mappings.

Data-Driven Knowledge Graphs. The data-driven (DD) graphs for oncology have 13 different types of nodes and 11 types of relationships. The schema of the data-driven graph is shown in Figure 2. They link the patients to their

[1] https://nciterms.nci.nih.gov/
[2] https://icd.who.int/browse1
[3] https://www.findacode.com/drugs/gpi-codes.html

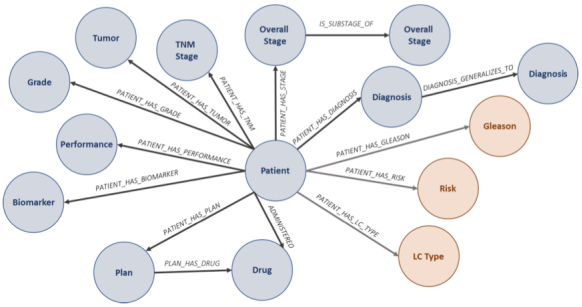

Figure 2. A snapshot of Data Driven Graph.

stage, tumor type, diagnoses, plans, administered drugs, performance measures, biomarkers, Gleason scores, risk groups, lung cancer type, all of which are custom defined cancer information nodes, and are stored with the date as a property of relevant relationships.

Ontology-Augmented Knowledge Graphs. The ontology-augmented (OA) graphs for oncology replaces the custom stage, TNM stage, grade, performance, biomarker, Gleason, lung cancer types, risk group, and diagnoses nodes with the ones that are defined in the NCIt ontology, thus reducing the number of class types from 13 to 7, but increasing the relationship types from 11 to 24 (including relationships within ontologies and those for ontology matching). One such oncology graph is shown in Figure 3.2. It gives us the additional relationship of relating tumors to their location in different anatomical parts of a human body, presenting the additional opportunity to derive additional useful insights.

3. Methodology

3.1. Engine Builder

The dynamic query engine is generated using engine builder module that takes in the schema and other details of a given graph to generate a query engine that can create queries with any n-number of parameters. These details are fetched using exploratory cypher queries which gets the details such as number of nodes in the graph, properties of a node and how a nodes are connected. It creates sub query functions for each node. For numerical data types the engine builder creates additional variables to use as upper and lower limit while working with the ranges. At this stage it also keeps in store the relations which have some data in it. As we will need to know the relations which have data when we query for the information. The engine builder only keeps the node that are connected to the target node to keep the computational processing in check. Only keeping the connected nodes reduces the processing that it is required.

3.2. Criteria Encoding

The criteria is encoded in a JSON format where every node has its own list of properties which can be set in the JSON file. Each property has

isPropertyDisjunction and *isValueDisjunction* which are boolean flags which tells whether or not the property will be in disjunction or not (ORing) or if the multiple values that were given will be in disjunction or conjunction (AND-ing). Both inclusion and exclusion criteria files option to set any property on any node which is connected to the target node.

3.3. Dynamic query engine

The dynamic query engine is the one that creates query with given nnumber of parameters. This is achieved by creating sub queries for each node. If we have a node n1 all the properties which are part of n1 will be handled in the sub query of n1. The module takes in the JSON file and parses through it invoking only creating subquery for those nodes that have been filled in the JSON. After the sub queries are created they are aggregated along with the target node. Target node is the one that will be returned after the criteria is met. The return patients are entire data which is present in the graph is queried and saved. Algorithm 1 provides an implementation for a dynamic Query engine. For disjunction with multiple values for the same property the query pattern for the cypher query changes, the query engine caters to the different query patterns. Along with this the query engine also works with date query pattern and different comparison operators

3.4. Criteria Diversity, Complexity and Coverage

Other than improved algorithms, work was done to ensure that the cohort builder can provide coverage for multiple criteria types ranging in their

complexity and diversity. For this purpose, an extensive literature review was done using ten relevant and recent research papers.

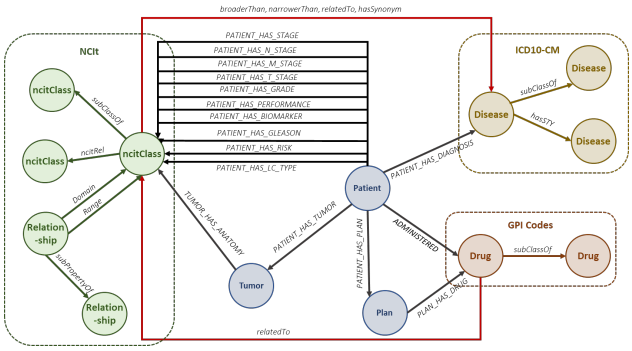

Figure 3. The Schema for Ontology Augmented Graph.

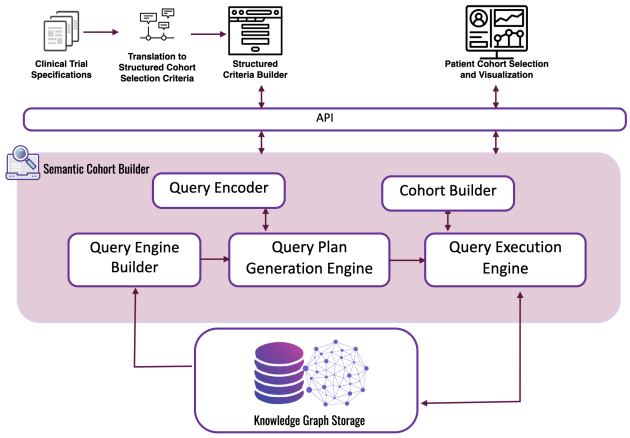

Figure 4. Semantic cohort builder.

Open-Source Tools CTKB. .some open-source projects and tools were explored such as CKTB. CKTB is a relational database of frequent criteria used in clinical trials and it was thus used to extract the top ten most frequent criteria parameters such as disease condition, etc. However, in the case of CTKB it failed to provide detailed information with regards to numeric ranges for different measurements such as those within medical lab tests. Moreover, there was a lack of temporal data to specify how much time needs to be elapsed for a patient to qualify for a cohort in terms of their past procedures done or past medications taken. Therefore, in the end CTKB was helpful however not that significant to our goal of ensuring criteria coverage.

Open-Source Tools C2Q. Another tool explored was C2Q that takes clinical trials eligibility criteria from ClinicalTrials.gov in the form of free text and uses different NLP techniques to annotate and categorize recognized entities with labels such as measurement, procedure, condition, etc. C2Q was able to thus provide additional details with regards to measurements and temporal data that CTKB was not. Moreover, APIs of ClinicalTrials.gov were also explored to aid in further research work. The relevant key people related to C2Q codebase were also emailed however no response was received. Additionally, work was done to reconstruct the open-source code of C2Q so its output of structured JSON criteria from free text could be made a part of the knowledge engineering pipeline in order to make the cohort builder more user-friendly. This was done to ensure that the end-to-end experience of the user would be NLP driven. The end user would need to simply write the criteria in free text and it would get relevant results from the cohort builder. Currently, the reconstruction is still a work in progress.

Schematic Coverage of the Graph by the Cohort Builder Some competency questions/scenarios that were used to test the cohort builder were not based on real clinical trials criteria however they helped indicate the schematic level coverage of our knowledge graph. These CQs were constructed to ensure that the cohort builder is able to capture data from all present graph schematics such as various Node Types, Relationship Types, Node Properties and Relationship Properties.

Schematic Coverage of the Graph Evaluation. These CQs were also varied in terms of the conjunctions, disjunctions, numeric ranges and number of graph schema aspects

Algorithm 1 Algorithm for dynamic Query engine

Given a set of nodes N, a set of properties P for each node, and a subquery function S, a combined query Q is formed by:

for all $n \in N$ **do**
  Call S for n
  if P is not $None$ **then**
    Return S of n with given parameters
    Append S to L
  else
    Append an empty string to L
  end if
end for

for all $e \in L$ **do**
  Concatenate e together
  Concatenate RETURN statement to Q
end for

Execute Q

for all $r \in R$ **do**
  Append r to T
end for

covered. Work was done to verify whether multiple data types such as string, numeric, boolean, and temporal can be handled.The cohort builder and its underlying dynamic query builder was constantly evolving and therefore, these CQs were executed against all versions from 1.1 to 1.4 inclusive to ensure that there were no breaking changes across versions and addition of new changes across versions was not affecting the previous features and implementation.

Schematic Coverage Queries Evaluation. Doing so also allowed for measuring the accuracy of the cohort builder as the query generated automatically by the cohort builder's query builder was matched and evaluated against the manual queries written to extract relevant patients in Neo4j browser. This also allowed for comparison of the syntactic and semantic similarities between the manual queries and the automatically generated queries and it was observed that in most case both the manual and generated queries were exactly the same and in some cases there was syntactic variation between the two but no semantic variation and therefore, all relevant patients were extracted correctly.

Schematic Coverage Queries Issue Diagnosis. The use of such CQs also helped in diagnosis of some cases that the builder was unable to handle such as empty strings given in string properties, issues in setting values for some properties related to Node Types such as that of Provider, issues with numeric ranges being unable to handle precise less than and greater than cases and always generating queries against less than equal to/greater than equal to. Lastly, it was observed that null values appeared for some relationship properties such as 'tumor size' and 'tumor size unit' within the output files of cohort builder.

Schematic Coverage Queries Issue Documentation and Fixes. . A documented test log was generated to keep track of the dates when each CQ/test case was executed, on which version of the cohort builder, issues/errors reported and the final results/pass status against each test case execution. These issues were addressed such as that of numeric ranges which were fixed so that precise '¿', '¡', '¿=' ,'¡=' and '=' could be set for numeric values especially against lab test values. An additional case of multiple different and same relationships properties between two nodes was also handled. The exclusion criteria handling was also further looked at to ensure relevant possibilities are being covered.

Schematic Coverage Extended Work. Effort was made to incorporate as much information as possible from available data sources either that in 10g or otherwise such as previously the graph had lab test values such as liver bilirubin, red blood cells hemoglobin, and serum creatinine without their units. Later, units were incorporated by making each specific lab a Node and storing the date, value and unit of that lab test between Patient and Lab nodes. Moreover, queries making use of NCITClass nodes were explored to aid in semantics enhancement although it is still a work in progress along with the functionality of being able to extract the most recent lab values, procedures done, drugs taken, etc

Criterias taken from ClinicalTrials.gov. The main focus was on five main cancer types: Breast Cancer, Multiple Myeloma, Colon Cancer, Non-Small Cell Lung Cancer and Non-Small Cell Lung Cancer. A specific comprehensive clinical trials eligibility criteria was taken for each of these cancer types from ClinicalTrials.gov and encoded and fed to the cohort builder. Criterias for these cancer types were selected to demonstrate various levels of complexity and diversity within the criterias and this was achieved by taking criterias from the past ten to fifteen years and of clinical studies that have already been completed. Moreover, diversity in selected criterias was ensured by choosing criterias of clinical trials conducted by different universities, research groups, big pharmaceutical companies, US federal agencies and non-profit organizations. Additionally, diversity within criterias selected for clinical studies was ensured through selecting criterias for such clinical studies that were taking place in the Americas, Europe and Asia. The details of these criteria is presented in 1 and the actual criteria points corresponding to these Trials are presented in 1. An example query automatically generated by cohort builder's dynamic query engine for breast cancer cohort building is illustrated in Figure 3.4. The remaining queries generated can be found

Table 1. Details of the five Competency Scenarios based on Oncology Clinical Trials Criteria taken from ClinicalTrials.gov

CQR-No	Cancer Type	Trial Conductors	Location Countries	Intervention	Trial Enrollment	Trial Completion	Patients Returned by Cohort Builder
CQR-1	Breast Cancer	Acetylon Pharmaceuticals Incorporated, Columbia University, National Cancer Institute	USA	ACY-1215 + Nab-Paclitaxel	17	September, 2020	45
CQR-2	Multiple Myeloma	Acerta Pharma	UK	Acalabrutinib + Dexamethasone	27	April, 2019	80
CQR-3	Colon Cancer	National Cancer Institute (NCI), Eastern Cooperative Oncology Group	Canada, Puerto Rico, United States	Cetuximab	3397	November, 2012	23
CQR-4	Non-Small Cell Lung Cancer	Hoffmann-La Roche	Singapore, Australia, Belgium, Canada, Spain, Turkey, Japan, Germany and 11 more counries	Atezolizumab	667	January, 2019	838
CQR-5	Non-Small Cell Lung Cancer	AGC Biologics S.p.A.e	Italy	Doxorubicin+ NGR-hTNF	28	May, 2011	9

in the Appendix section.

```
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*breast.*' OR DISEASE.prefLabel =~'(?i).*neopla
sm.*')with PATIENT, COLLECT(DISEASE.prefLabel) as conditions WHERE ANY (condition IN conditions WHERE condition =~'(?i).*breast.*') AND
 ANY (condition IN conditions WHERE condition =~'(?i).*neoplasm.*')
MATCH (PATIENT:Patient)-[rel_ncitclass]-(NCITCLASS:ncitClass) WHERE (NCITCLASS.prefLabel =~'(?i).*HER2.*')
MATCH (PATIENT:Patient) WHERE (PATIENT.gender =~'(?i).*M.*')  AND  (PATIENT.age >= 45)
MATCH (PATIENT:Patient)-[rel_Labs]-(LABS:Labs) WHERE (LABS.liver_bilirubin <= 1.5)  AND  (LABS.liver_ast <= 137.5)  AND  (LABS.liver_alt
 <= 100)  AND  (LABS.blood_red_blood_cells_hgb >= 9)  AND  (LABS.kidney_creatinine <= 1.95)
RETURN DISTINCT PATIENT
Execution time for
 0.4204599857330322 seconds
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*HIV.*' OR DISEASE.prefLabel =~'(?i).*arrhythmi
a.*' OR DISEASE.prefLabel =~'(?i).*psychiatric.*' OR DISEASE.prefLabel =~'(?i).*heart failure.*')
RETURN DISTINCT PATIENT
Execution time for
 0.07843518257141113 seconds
Total Number of patients: 45
```

Figure 5. Semantic cohort builder.

Cohort Builder User Interface's Interactive Prototype. Lastly, the following user interface depicted in Fig1 was designed based on the wiki- data's query builder. It was designed as a lofi prototype and mockup to envision how the graphical interface would be for the end user of the cohort builder. An initial paper prototype along with research was constructed to ensure that major user interaction flow paths and ease of structured criteria input through a form-based interface can be understood. This paper prototype was later converted to a figma interactive prototype for clinical trial inclusion/exclusion criteria along with its relevant cohort generation. A screenshot of this Figma Prototype can be seen in Figure 6.

Interactive Prototype Design Choices. The outer cards help select the entities and the inner cards help set properties within those entities. In order to handle multiple data types, different form elements were used such as large text boxes for descriptive node properties such as the name for disease conditions that should be matched for a patient to be selected as part of a cohort. Additionally, radio buttons and checkboxes are used to enable which option within the numeric ranges for properties such as Lab Test Values, Patient Age, etc can be opted. Furthermore, the first toggle button was added to the prototype to ensure that the user is able to specify that values for a specific property such as patient gender should be and-ed or or-ed (conjunction or disjunction). A second toggle button was also added to give control to the user so that they can choose whether any specific property should be and-ed or or-ed with other properties. An example of this would be that lab values for two properties bilirubin and hemoglobin needed to be and-ed or or-ed as in does the user require patients which have a specific value x for bilirubin AND and a value y for hemoglobin or they require patients which have a specific value x for bilirubin OR and a value y for hemoglobin. Further details of this prototype can be found in Appendix.

Table 2. Detailed Criteria points of the five Competency Scenarios based on Oncology Clinical Trials taken from ClinicalTrials.gov

CQR-No	Cancer Type	Criteria
CQR-1	Breast Cancer	Disease: Breast Cancer AND Biomarker: HER2 AND Age: Greater than or Equal to 45 AND Bilirubin: Less than or Equal to 1.5 AND AST: Less than or Equal to 137.5 AND ALT: Less than or Equal to 100 AND HGB: Greater than or Equal to 9 AND Creatinine Less than or Equal to 1.95 AND NOT Disease: HIV, Arrhythmia, Psychiatric condition, Heart failure
CQR-2	Multiple Myeloma	Disease: Multiple Myeloma AND Performance: ECOG Performance Status 2 AND NOT Disease: History of heart disease, Systolic Heart failure, Diastolic Heart failure, Heart failure, Myocardial Infractio Cardiovascular disease, Arrhythmia
CQR-3	Colon Cancer	Disease: Malignant Neoplasm of Colon AND Performance: ECOG Performance Status 2 AND Age: Between 18 and 99 inclusive Bilirubin: Less than or Equal to 1.5 AND HGB: Greater than or Equal to 9 AND Creatinine Less than or Equal to 1.95 AND NOT Disease: HIV, Malignant Neoplasn of Rectum, Pulmonary Fibrosis, Systolic Heart failure, Diastolic Heart failure, Heart failure Unstable Angina, Allergy to Drugs and Medication
CQR-4	Non-Small Cell Lung Cancer	Disease: Lung Cancer AND Biomarker: PD-L1 AND NOT Disease: HIV, Hepatitis B, Hepatitis C, Idopathic Pulmonary Fibrosis, Autoimmune related conditions, Central Nervous System related conditions
CQR-5	Non-Small Cell Lung Cancer	Disease: Malignant Neoplasm of Lung AND Performance: ECOG Performance Status 2 AND Age: Greater than or Equal to 18 AND Bilirubin: Less than 1.5 AND AST: Less than 100 AND ALT: Less than 112.5 AND Creatinine: Less than 1.95 AND Tumor Size: Greater than or Equal to 20 AND Tumor Size Unit: mm AND NOT Disease: Arrhythmia, Systolic Heart failure, Diastolic Heart failure, Heart failure, Unstable Angina, Peripheral Vascular disease, Allergy to Drugs and Medication, Central Nervous System related conditions, Hypertension

4. Results

The results for the cohort builder were consolidated based primarily on the execution time that it took for the criteria to be matched and relevant patients retrieved. Furthermore, the number of patients returned against each criteria was noted along with the complexity of the criteria as elaborated by the number of node types, node/relationship properties, conjunction, disjunction and numeric ranges within that specific criteria. The results of these metrics for the five cancer types mentioned above are presented in Table 3.

The results for some other criterias that were not based on real clinical trials criteria/ schematic level graph coverage criterias are presented in Table 4.

5. Discussion

The cohort builder can thus be used to tackle the challenge of building patient cohorts with improved patient selection, reduced costs and time, and improved outcomes for patients. This will be of benefit to clinicians, nurses, doctors, medical professionals, medical researchers, and other stakeholders in the cancer clinical trials ecosystem. However, it is also important to note that the current cohort builder has much room for improvement in terms of the complexity of the criterias it can handle especially with regards to elapsed time based criterias related to procedures or previous drugs taken by the patients. Moreover, it's important to note that sometimes the exclusion criteria for a specific real cancer clinical trial in present within the knowledge graph however as long as no patients are connected to those exclusion criteria nodes; the exclusion cannot be applied as the cohort builder looks at the relationship of a specific patient with that of other entities such as drugs, procedures, conditions etc. To elaborate, for the breast cancer criteria we selected, we did have histone inhibitors, valproic acid, calcium channel blockers and macrolides present as nodes with the Drug NodeType label however since none of these drugs wee connected to any patient in the graph with the relationship of ADMINISTERED, thus these could not prove to be helpful in the exclusion criteria.

6. Conclusion

In conclusion, the cohort builder described in this text provides a solution to the challenge of building patient cohorts with improved patient selection, reduced costs and time, and improved outcomes for patients in the oncology clinical

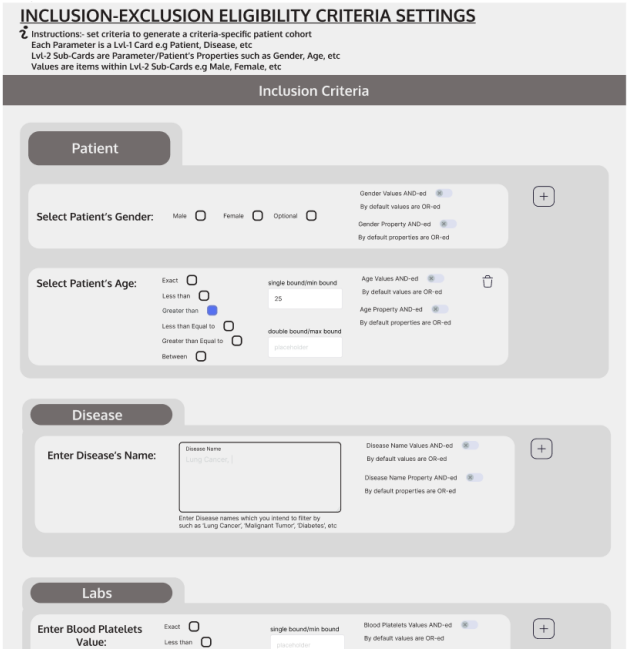

Figure 6. Cohort Builder's UI Interactive Prototype

Table 3. Results obtained against five Competency Scenarios based on Oncology Clinical Trials Criteria taken from ClinicalTrials.gov, where the abbreviations used in the table header are as follows: CQR-No is the identifier of competency question based on criteria taken from ClinicalTrials.gov, CT refers to Cancer Type, PR refers to number of Patients Returned in the cohort generated by the cohort builder, MET is the Mean Execution Time needed to match the criteria and extract the relevant results, NoD is the Number of Disjunction used in the criteria, NoC is the Number of Conjunctions used, NoNR is Number of Numeric Ranges used, NoMNT is the Number of Match clauses or NodeTypes covered in the criteria; and lastly, NoP is the Number of Properties (node and relationship) covered in the criteria.

CQR-No	CT	PR	MET	NoD	Noc	NoNR	NoMNT	Nop
CQR-1	C-50	45	0.497	4	9	6	5	10
CQR-2	C-90	80	0.416	7	2	nil	3	3
CQR-3	C-18	23	7.456	8	7	5	5	8
CQR-4	C-34-NSC	838	0.402	6	3	nil	3	3
CQR-5	C-34-SC	9	0.777	9	9	6	6	10

Table 4. Results obtained against Competency Scenarios used to evaluate the Schematic Level Coverage of the Graph by the Cohort Builder

CQR-No	PR	MET	NoD	Noc	NoNR	NoMNT	Nop
CQ-1	0.144	256	nil	2	nil	3	3
CQ-2	0.114	8	1	3	nil	3	5
CQ-3	0.012	60557	nil	1	2	3	4
CQ-4	0.144	503	nil	2	3	3	4
CQ-5	1.333	199	nil	5	1	6	7
CQ-6	0.061	1866	nil	2	1	3	3
CQ-7	0.042	26	1	2	1	3	3
CQ-8	0.123	34	nil	4	1	5	5
CQ-9	0.099	9	nil	4	nil	5	5
CQ-10	0.082	219	nil	1	nil	2	2
CQ-11	3.658	391	nil	5	4	4	4
CQ-12	0.06	1204	nil	1	nil	2	2
CQ-13	0.314	2315	nil	4	2	2	3

trials ecosystem. The dynamic query engine allows for the creation of queries with any number of parameters, and the criteria coverage has been improved through the exploration of tools such as CKTB. The results of the cohort builder, as

presented in Table 1, demonstrate its effectiveness in retrieving relevant patients based on complex and diverse criteria for five different cancer types.

7. Appendix

```
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*neoplasm.*' OR DISEASE.prefLabel =~'(?i).*lung
.*')with PATIENT, COLLECT(DISEASE.prefLabel) as conditions WHERE ANY (condition IN conditions WHERE condition =~'(?i).*neoplasm.*') AND
 ANY (condition IN conditions WHERE condition =~'(?i).*lung.*')
MATCH (PATIENT:Patient)-[rel_ncitClass]-(NCITCLASS:ncitClass) WHERE (NCITCLASS.prefLabel =~'(?i).*PD-L1.*')
RETURN DISTINCT PATIENT
Execution time for
 0.2993690967559814S seconds
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*central nervous system.*' OR DISEASE.prefLabel
 =~'(?i).*HIV.*' OR DISEASE.prefLabel =~'(?i).*hepatitis B.*' OR DISEASE.prefLabel =~'(?i).*hepatitis C.*' OR DISEASE.prefLabel =~'(?i).
*idopathic pulmonary fibrosis.*' OR DISEASE.prefLabel =~'(?i).*autoimmune.*')
RETURN DISTINCT PATIENT
Execution time for
 0.18882095152216797 seconds
Total Number of patients: 838
```

Figure 7. Small cell lung cancer

```
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*malignant neoplasm.*' OR DISEASE.prefLabel =~'
(?i).*lung.*')with PATIENT, COLLECT(DISEASE.prefLabel) as conditions WHERE ANY (condition IN conditions WHERE condition =~'(?i).*maligna
nt neoplasm.*') AND  ANY (condition IN conditions WHERE condition =~'(?i).*lung.*')
MATCH (PATIENT:Patient)-[rel_ncitClass]-(NCITCLASS:ncitClass) WHERE (NCITCLASS.prefLabel =~'(?i).*ECOG Performance Status 2.*')
MATCH (PATIENT:Patient) WHERE (PATIENT.age >= 18)
MATCH (PATIENT:Patient)-[rel_Labs]-(LABS:Labs) WHERE (LABS.liver_bilirubin < 1.5)  AND  (LABS.liver_ast < 100)  AND  (LABS.liver_alt < 1
12.5)  AND  (LABS.kidney_creatinine < 1.95)
MATCH (PATIENT:Patient)-[rel_Tumor]-(TUMOR:Tumor) WHERE (rel_Tumor.size >= 20)  AND  (rel_Tumor.sizeUnit =~'(?i).*mm.*')
RETURN DISTINCT PATIENT
Execution time for
 0.6928448677062988 seconds
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*central nervous system.*' OR DISEASE.prefLabel
 =~'(?i).*systolic.*' OR DISEASE.prefLabel =~'(?i).*diastolic.*' OR DISEASE.prefLabel =~'(?i).*arrhythmia.*' OR DISEASE.prefLabel =~'(?i
).*heart failure.*' OR DISEASE.prefLabel =~'(?i).*peripheral vascular disease.*' OR DISEASE.prefLabel =~'(?i).*unstable angina.*' OR DIS
EASE.prefLabel =~'(?i).*allergy status.*' OR DISEASE.prefLabel =~'(?i).*hypertension.*')
RETURN DISTINCT PATIENT
Execution time for
 0.07999396324157715 seconds
Total Number of patients: 9
```

Figure 8. Non small cell lung cancer

```
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*multiple myeloma.*')
MATCH (PATIENT:Patient)-[rel_ncitClass]-(NCITCLASS:ncitClass) WHERE (NCITCLASS.prefLabel =~'(?i).*ECOG Performance Status 2.*')
RETURN DISTINCT PATIENT
Execution time for
 0.3150141239166626 seconds
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*history of heart disease.*' OR DISEASE.prefLab
el =~'(?i).*systolic.*' OR DISEASE.prefLabel =~'(?i).*diastolic.*' OR DISEASE.prefLabel =~'(?i).*arrhythmia.*' OR DISEASE.prefLabel =~'(
?i).*myocardial infraction.*' OR DISEASE.prefLabel =~'(?i).*heart failure.*' OR DISEASE.prefLabel =~'(?i).*cardiovascular disease.*')
RETURN DISTINCT PATIENT
Execution time for
 0.12281441688537598 seconds
Total Number of patients: 80
```

Figure 9. Multiple Myeloma

```
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*malignant neoplasm of colon.*')
MATCH (PATIENT:Patient)-[rel_ncitClass]-(NCITCLASS:ncitClass) WHERE (NCITCLASS.prefLabel =~'(?i).*ECOG Performance Status 2.*')
MATCH (PATIENT:Patient) WHERE (PATIENT.age >= 18)  AND  (PATIENT.age <= 99)
MATCH (PATIENT:Patient)-[rel_Labs]-(LABS:Labs) WHERE (LABS.liver_bilirubin <= 1.5)  AND  (LABS.blood_red_blood_cells_hgb >= 9)  AND  (LA
BS.kidney_creatinine <= 1.95)
RETURN DISTINCT PATIENT
Execution time for
 7.234000444122314 seconds
MATCH (PATIENT:Patient)-[rel_Disease]-(DISEASE:Disease) WHERE (DISEASE.prefLabel =~'(?i).*HIV.*' OR DISEASE.prefLabel =~'(?i).*malignant
 neoplasm of rectum.*' OR DISEASE.prefLabel =~'(?i).*pulmonary fibrosis.*' OR DISEASE.prefLabel =~'(?i).*heart failure.*' OR DISEASE.pre
fLabel =~'(?i).*systolic.*' OR DISEASE.prefLabel =~'(?i).*diastolic.*' OR DISEASE.prefLabel =~'(?i).*Unstable angina.*' OR DISEASE.prefL
abel =~'(?i).*allergy status.*')
RETURN DISTINCT PATIENT
Execution time for
 0.13400650024414062 seconds
Total Number of patients: 23
```

Figure 10. Colon cancer

References

[1] Franz Baader, Stefan Borgwardt, and Walter Forkel. Patient Selection for Clinical Trials Using Temporalized Ontology-Mediated Query Answering. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pages 1069–1074, 2018.

[2] Kemele M. Endris, Philipp D. Rohde, Maria Esther Vidal, and Sören Auer. Ontario: Federated Query Processing Against a Semantic Data Lake. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11706 LNCS:379–395, 2019.

[3] SM Shamimul Hasan, Donna Rivera, Xiao-Cheng Wu, Eric B Durbin, J Blair Christian, and Georgia Tourassi. Knowledge graph-enabled cancer data analytics. *IEEE journal of biomedical and health informatics*, 24(7):1952–1967, 2020.

[4] Tina Hernandez-Boussard, Keri L Monda, Blai Coll Crespo, and Dan Riskin. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *Journal of the American Medical Informatics Association*, 26(11):1189–1194, 2019.

[5] Zhisheng Huang, Annette Ten Teije, and Frank Van Harmelen. Semanticct: a semantically-enabled system for clinical trials. In *Process Support and Knowledge Representation in Health Care: AIME 2013 Joint Workshop, KR4HC 2013/ProHealth 2013, Murcia, Spain, June 1, 2013, Revised Selected Papers*, pages 11–25. Springer, 2013.

[6] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. *Scientific Reports*, 7(1):1–11, 2017.

[7] Stefan Schmid, Cory Henson, and Tuan Tran. Using Knowledge Graphs to Search an Enterprise Data Lake. *Lecture Notes in Computer Science (including subseries LN in AI and Bioinformatics)*, 11762 LNCS:262–266, 2019.

[8] Isabel Segura-Bedmar and Pablo Raez. Cohort selection for clinical trials using deep learning models. *Journal of the American Medical Informatics Association*, 26(11):1181–1188, 2019.

[9] Yong Shang, Yu Tian, Min Zhou, Tianshu Zhou, Kewei Lyu, Zhixiao Wang, Ran Xin, Tingbo Liang, Shiqiang Zhu, and Jingsong Li. EHR-Oriented Knowledge Graph System: Toward Efficient Utilization of Non-Used Information Buried in Routine Clinical Practice. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2463–2475, 2021.

[10] Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: N2C2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 2019.

[11] Amber Stubbs and Özlem Uzuner. New approaches to cohort selection. *Journal of the American Medical Informatics Association*, 26(11):1161–1162, 2019.

[12] V. G.Vinod et al. Vydiswaran. Hybrid bag of approaches to characterize selection criteria for cohort

identification. *Journal of the American Medical Informatics Association*, 26(11):1172–1180, 2019.

[13] Xin Wang, Qiang Xu, Le Le Chai, Ya Jun Yang, and Yun Peng Chai. Efficient Distributed Query Processing on Large Scale RDF Graph Data. *Ruan Jian Xue Bao/Journal of Software*, 30(3):498–514, 2019.