Leveraging SNOMED CT for patient cohort identification over heterogeneous EHR data

Xubing Hao[1], Yan Huang, PhD[2], Licong Cui, PhD[1,⋆], Xiaojin Li[2,⋆]
[1]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
[2]Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX

Abstract

SNOMED CT is extensively employed to standardize data across diverse patient datasets and support cohort identification, with studies revealing its benefits and challenges. In this work, we developed a SNOMED CT-driven cohort query system over a heterogeneous Optum® de-identified COVID-19 Electronic Health Record dataset leveraging concept mappings between ICD-9-CM/ICD-10-CM and SNOMED CT. We evaluated the benefits and challenges of using SNOMED CT to perform cohort queries based on both query code sets and actual patients retrieved from the database, leveraging the original ICD-9-CM and ICD-10-CM as baselines. Manual review of 80 random cases revealed 65 cases containing 148 true positive codes and 25 cases containing 63 false positive codes. The manual evaluation also revealed issues in code naming, mappings, and hierarchical relations. Overall, our study indicates that while the SNOMED CT-driven query system holds considerable promise for comprehensive cohort queries, careful attention must be given to the challenges of falsely included codes and patients.

1 Introduction

SNOMED CT is the most comprehensive health terminology worldwide[1,2]. It provides a standardized vocabulary for representing clinical concepts and facilitating data exchange and integration across different health information systems, including Electronic Health Records (EHRs). For instance, the largest observational data network worldwide - Observational Health Data Sciences and Informatics (OHDSI)[3], has adopted SNOMED CT as a standard vocabulary for diagnoses (or conditions) in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM); and patient data originally encoded with diverse vocabularies such as the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)[4] and the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)[5] are mapped to SNOMED CT codes[6].

SNOMED CT has also been widely used to support patient cohort retrieval or identification[7,8]. Formally, cohort identification refers to computer-based matching of eligibility criteria against clinical data to identify a cohort of potentially eligible patients[9]. An eligibility criterion needs to be transformed into a computable representation (e.g., diagnosis codes) and then queried against the backend database. Web-based query tools such as i2b2[10] and ATLAS[11] have been developed for cohort identification over clinical data. Such query tools heavily rely on the backend vocabularies as the semantic backbone for query construction, expansion, and translation. For instance, in the ATLAS tool developed by the OHDSI community[11], most standard concepts related to diagnoses in the OMOP vocabulary are sourced from SNOMED CT to support structured query from heterogeneous data coded with diverse vocabularies.

Studies have shown both the benefits and challenges of using SNOMED CT as a standard vocabulary. Jung et al.[12] compared the effectiveness of using SNOMED CT, ICD-10, and Korean Classification of Diseases-7 (KCD-7) to generate epilepsy patient cohorts; and they concluded that SNOMED CT is more suitable for being used as the standard vocabulary for epilepsy patient cohort identification. According to Tavakoli et al.[13], SNOMED CT has better semantic alignment than ICD-10-CM and ICD-11 for ophthalmic infections and ophthalmic trauma. Willett et al.[14] defined clinical conditions in EHRs using SNOMED CT and highlights the simpleness, conciseness, and shareable of SNOMED CT-based diagnosis value sets in EHRs. However, in a study evaluating the content coverage of the OMOP vocabulary for kidney transplant-related concepts[15], among 1,981 concepts that cannot be covered by the OMOP vocabulary, 450 (22.72%) is in the category of condition (or diagnoses).

⋆Corresponding author. Email: xiaojin.li@uth.tmc.edu, licong.cui@uth.tmc.edu

In this study, we develop a SNOMED CT-driven cohort query system over a heterogeneous EHR data, Optum® de-identified COVID-19 Electronic Health Record data, and evaluate the benefits and challenges of using diagnosis-related concepts in SNOMED CT to perform cohort queries in this setting.

2 Methods

We leveraged the Optum® COVID-19 data with diverse vocabularies (such as ICD-9-CM, ICD-10-CM, and SNOMED CT for diagnoses) to develop the SNOMED CT-driven cohort query system. We extracted the diagnosis-related concept mappings from OMOP vocabularies to create the code mappings between ICD-9-CM/ICD-10-CM and SNOMED CT that are needed for supporting cohort queries. To assess the benefits and challenges of using diagnosis-related concepts in SNOMED CT to perform cohort queries over the Optum® COVID-19 data, we performed two types of evaluation: (1) based on the query code sets, and (2) based on the actual patient cohorts retrieved from the dataset. Figure 1 shows the overall workflow of our study.

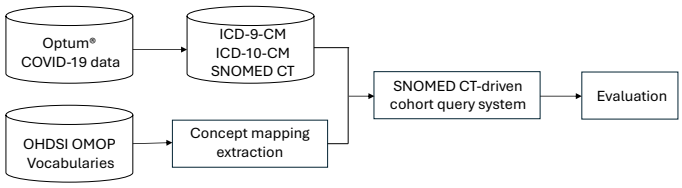

Figure 1: Workflow of our study.

2.1 Optum® COVID-19 data

We utilized the Optum® COVID-19 data, sourced from numerous healthcare providers across the United States, comprising over 700 hospitals and 7,000 clinics. As of the January 2022 release, this dataset encompassed 8.87 million unique individuals who had received documented clinical care with a diagnosed COVID-19 or acute respiratory illness after 02/01/2020 or had undergone documented COVID-19 testing regardless of the results. The dataset contains a vast array of raw clinical data, including newly identified COVID-specific clinical data points from both inpatient and ambulatory electronic medical records. This comprehensive data encompasses patient-level information such as demographics, diagnoses, procedures, laboratory tests, care settings, medications prescribed or administered, and mortality records. These data have been certified as de-identified by independent statistical experts in accordance with the Health Insurance Portability and Accountability Act (HIPAA) statistical de-identification rules. They are managed under the Optum® COVID-19 data customer data use agreement, ensuring compliance with privacy regulations and safeguarding patient confidentiality.

The Optum® COVID-19 data was released to support data-driven COVID-19 research. Like the traditional research workflow, the first step of performing COVID-19-related research often involves identifying patient cohorts and formulating scientific hypotheses. However, a key data challenge hindering patient cohort identification is the coding heterogeneity, that is, a mixed use of vocabularies such as ICD-9-CM, ICD-10-CM, and SNOMED CT for diagnoses. As a consequence, researchers need to manually collect all possible codes for a specific diagnosis or health condition from disparate coding systems before requesting data to identify eligible patients, which is time-consuming and cumbersome[16]. Therefore, there is a need to develop a SNOMED CT-driven cohort query system to support patient cohort identification by researchers without the need to curate heterogeneous codes for eligibility criteria.

We first loaded the Optum® COVID-19 data into a MongoDB database[17]. We further built the Event-level Inverted Index (ELII) initially proposed in our previous work[18], which showed remarkable query performance.

2.2 Concept mapping extraction

The Observational Health Data Sciences and Informatics (OHDSI) has generated and currently maintains mappings from over 130 source vocabularies[6]. SNOMED CT serves as a standard vocabulary for diagnoses (or conditions) in the OHDSI Standardized Vocabularies. SNOMED CT's selection is attributed to its global applicability, clinical focus, detailed granularity, comprehensive hierarchical structure, and growing adoption in clinical data entry through methods such as natural language processing and problem lists[19]. Codes that are non-standard from vocabularies such as ICD-9-CM and ICD-10-CM are mapped to SNOMED CT standard codes.

In OHDSI, the concept mapping process always attempts to reflect the semantic of a Source Concept by an Equivalent Standard Concept. Equivalent implies that the concepts share identical meanings, cover the same semantic scope, and crucially, maintain the same hierarchical relationships. In cases where an Equivalent Standard Concept is unavailable, the mapping strategy shifts to align with a more general (uphill) Standard Concept(s)[20]. For example, ICD-10-CM code "K80.0: Calculus of gallbladder with acute cholecystitis" is mapped to SNOMED CT concept "59771005: Calculus of gallbladder with acute cholecystitis (disorder)". This is an example of mapping to an Equivalent Standard Concept. Additionally, ICD-10-CM code "K80.00: Calculus of gallbladder with acute cholecystitis without obstruction" (K80.0's child) is mapped to SNOMED CT concept "197377009: Gallbladder calculus with acute cholecystitis and no obstruction (disorder)" (59771005's child). An example of mapping to a more generic Standard Concept is that ICD-9-CM code "208.11: Chronic leukemia of unspecified cell type, in remission" is mapped to SNOMED CT concept "92811003: Chronic leukemia in remission (disorder)".

In this work, we reused the code mappings from OHDSI OMOP vocabularies (used in the 1.13.0 version of Athena) to create the concept mappings needed for supporting cohort queries. Since the Optum® COVID-19 data uses three vocabularies (ICD-9-CM, ICD-10-CM and SNOMED CT) to encode its diagonosis data, we specifically extracted mappings between ICD-9-CM and SNOMED CT and mappings between ICD-10-CM and SNOMED CT.

2.3 SNOMED CT-driven cohort query system development

Figure 2 shows the general architecture design of our SNOMED CT-driven cohort query system, consisting of three core architectural elements: 1) a web-based interface (see Figure 2.A), called query builder, which is a robust and intuitive interface meticulously that has been designed and developed to empower researchers in quickly identifying the required SNOMED CT terms and performing an exploratory cohort query; 2) an advanced query engine for searching EHR records (see Figure 2.B); such a query engine expanded the user queries built from the web-based interface with mapped ICD codes and translates expanded queries into executable database query languages, and consists of three modules: an ICD codes expansion module, a query translation module a query translation module, a query execution module, and a patient information retrieval module; and 3) two backend MongoDB databases for storing SNOMED CT terms, ICD Mappings and ELII of COVID-19 EHR data (see Figure 2.C). The query system is implemented using Ruby on Rails[21], which is an agile web development framework.

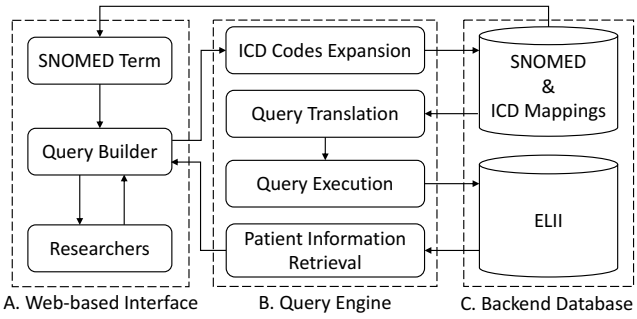

Figure 2: The system architecture of the SNOMED CT-driven cohort query system.

The query builder consists of three areas which correspond to three steps to perform SNOMED CT-driven queries as follows: 1) query term selection area, where users can find and select query terms (i.e., SNOMED CT terms) and add them to the query construction area; 2) query construction area, where shows the mapped ICD codes for each SNOMED CT term, 3) query results display area, where the patient list and demographic information retrieved from ELII satisfying the query criteria are returned to the user. In the query term selection area, two modes are available for finding query terms of interest: browsing and searching. The browsing mode presents query terms in a hierarchical order, enabling users to navigate through SNOMED terms along with their direct descendants. The search mode caters to users with specific knowledge, allowing them to directly search for SNOMED terms of interest. Based on the SNOMED CT terms selected by a user, the query builder automatically generates visual query widgets using a dynamic approach to show all the SNOMED CT descendants and mapped ICD codes. The query construction area is designed to be as close to natural language as possible, ensuring that the query logic is easily understandable and clear to users. The query results display area is driven by the query criteria specified in the query construction area.

As users select query terms to define queries, the builder generates an array of key-value pairs within JSON objects that represent the current state of the user interface and query criteria. These objects do not directly contain query language

but include the query terms along with additional metadata describing the query. Subsequently, the expansion module automatically retrieves the SNOMED CT descendants and mapped ICD codes based on the selected SNOMED CT terms and incorporates them into the query criteria. Then, the query translation module converts these JSON objects with expanded ICD codes into MongoDB statements to query the backend database. A general template is pre-defined and used for dynamically generating the actual MongoDB statement for query translation. The template for querying SNOMED CT term with expanded ICD codes is defined as:

db.records_collection.distinct(< $mapped_patient_identifier$ >,
{"$or": [{"DIAGNOSIS_CD": {"$in": < $snomed_ids$ >}, "DIAGNOSIS_CD_TYPE": "SNOMED"},
{"DIAGNOSIS_CD": {"$in": < $mapped_ICD_9_codes$ >}, "DIAGNOSIS_CD_TYPE": "ICD9"},
{"DIAGNOSIS_CD": {"$in": < $mapped_ICD_10_codes$ >}, "DIAGNOSIS_CD_TYPE": "ICD10"}]})

where $< mapped_patient_identifier >$ represents the variable name of the unique patient identifier in the corresponding dataset. $< snomed_ids >$ is the set of SNOMED IDs of selected SNOMED terms and their descendants. $< mapped_ICD_9_codes >$ and $< mapped_ICD_10_codes >$ represent the sets of ICD-9 codes and ICD-10 codes, respectively, that are mapped to the selected SNOMED CT terms and their descendants. For instance, to query the SNOMED CT term "38921001: Measles with complication (disorder)":

db.records_collection.distinct(< $mapped_patient_identifier$ >,
{"$or": [{"DIAGNOSIS_CD": {"$in": ["38921001", "186562009"]},
"DIAGNOSIS_CD_TYPE": "SNOMED"},
{"DIAGNOSIS_CD": {"$in": ["0558", "05579"]}, "DIAGNOSIS_CD_TYPE": "ICD9"},
{"DIAGNOSIS_CD": {"$in": ["B0589", "B058", "B0581", "B054"]},
"DIAGNOSIS_CD_TYPE": "ICD10"}]})

The query execution module sends the translated MongoDB statements to the backend database to execute the query and subsequently receives a list of eligible patients that meet the query criteria. The patient information retrieval module gathers and reorganizes the query results to facilitate the user interface display. We utilize MongoDB as the backend database due to its advantages: 1) robust query performance with large-scale datasets; 2) flexible data modeling capabilities, enabling the development of customized structures such as inverted indexes to enhance query efficiency; 3) seamless scalability, allowing efficient expansion through the addition of standard servers; and 4) developer-friendly features that facilitate expedited development while reducing the risk of errors.

2.4 Evaluation

To evaluate the benefits and challenges of using diagnosis-related concepts in SNOMED CT to perform cohort queries, we assess its implications for both the query code sets and the resultant patient cohorts retrieved using the code set from the Optum® COVID-19 data, leveraging the original ICD-9-CM and ICD-10-CM as baselines. In our system, once a standard SNOMED CT concept has ICD codes mapped to it, patients with the mapped ICD codes will also be returned. We leverage a downward query mechanism which is the code and its descendent codes will all be used when querying patients. For example, if querying patients with ICD-10-CM code "G44.04: Chronic paroxysmal hemicrania", the downward query strategy will return patients with ICD-10-CM code "G44.04" as well as all the descendants of "G44.04" in ICD-10-CM (including "G44.041: Chronic paroxysmal hemicrania, intractable" and "G44.049: Chronic paroxysmal hemicrania, not intractable"). We conduct the evaluation based on each standard SNOMED CT concept. For each SNOMED CT standard concept *X*, we would have two code sets: (1) the baseline code set C_1 which includes all ICD codes that are mapped to *X* as well as these ICD codes' descendants; and (2) the extended code set C_2 which includes all ICD codes that are mapped to *X* and *X*'s descendants, as well as these ICD codes' descendants. These two code sets will be later used for patient cohort query in the Optum® COVID-19 data.

For the code sets, there are two possible relations between C_1 and C_2: (1) $C_1 = C_2$; and (2) $C_1 \subset C_2$. If $C_1 = C_2$, then there is no additional code included in the extended code set comparing to the baseline code set and thus no additional

patient will be queried from the database. If $C_1 \subset C_2$, then there are additional codes included in the extended code set and there might be additional patient retrieved from the database comparing to the baseline code set. We refer to the ICD code in C_2 but not in C_1 as Additional Positive (AP), AP representing a concept that is a valid subtype of the original SNOMED CT concept as Additional True Positive (ATP), and AP representing a concept that is not a valid subtype of the original SNOMED CT concept as Additional False Positive (AFP).

One challenge is that we do not have a gold standard to distinguish between ATPs and AFPs. Our investigation reveals that when mappings between ICD codes and SNOMED CT concepts are inconsistent with the respective vocabularies' hierarchical structures, it may indicate the presence of AFPs within the extended code set. If an ICD code *A* is mapped to SNOMED CT code *X*, but *A*'s ancestor in ICD is mapped to *X*'s descendent in SNOMED CT, this mapping may contradict the hierarchical structures of the vocabularies (we call such mappings as *contradictory cases*). Since systematically reviewing all the cases would be time-consuming and labor intensive, in this study, we randomly selected a subset from both the contradictory and non-contradictory cases and conducted a manual review, establishing a reference standard to identify ATPs and AFPs. By analyzing ATP and AFP, and the actual patient retrieved by these codes, we aim to assess the extent to which the SNOMED CT-driven cohort query system offers advantages or disadvantages compared to a query system that does not utilize SNOMED CT as a standardized vocabulary.

3 Results

3.1 SNOMED CT-driven cohort query system

The query system, as illustrated in Figure 3, consists of three distinct areas, each serving a specific purpose. In Figure 3. A, the query term selection area presents all SNOMED CT terms, providing researchers with an extensive list from which they can select terms of interest. Users have the option to utilize the search mode within this area, enabling them to enter text and retrieve relevant query terms quickly and efficiently. Figure 3.B shows the query construction area, which played a pivotal role in formulating queries. Within this area, one selected SNOMED CT term was displayed: "38921001: Measles with complication (disorder)." Alongside this term, its associated SNOMED CT descendants and mapped ICD codes were also presented. This comprehensive view allows researchers to construct precise and targeted queries, ensuring that the query logic remains transparent and comprehensible throughout the process. The query results display area, as shown in Figure 3.C, provided valuable insights into the outcomes of the formulated queries. Users can observe the number of patients that meet the specified query criteria, offering a clear indication of the scope and relevance of the query results. Additionally, a detailed list of patient IDs and their demographic information was presented, facilitating further analysis and interpretation of the query outcomes.

Overall, the query system provided a powerful and intuitive interface for constructing and executing queries, leveraging the SNOMED CT terminologies and associated data to generate meaningful insights and discoveries. Through its user-friendly design and comprehensive features, the interface streamlines the query process, empowering researchers to efficiently navigate complex healthcare datasets and unlock valuable knowledge in their respective fields.

3.2 Concept mapping statistics

There are a total of 358,356 concepts in SNOMED CT, 17,564 codes in ICD-9-CM, and 98,593 codes in ICD-10-CM. The mappings between ICD and SNOMED CT in OHDSI concept mappings can be one to one, one to many, or many to many. There are 10,741 SNOMED CT concepts at least having one ICD-9-CM code that maps to it. There are 17,311 ICD-9-CM codes that can be mapped to at least one standard SNOMED CT concept. There are a total of 20,555 mapping pairs between ICD-9-CM and SNOMED CT. There are 14,011 SNOMED CT concepts at least having one ICD-10-CM code that maps to it. There are 96,990 ICD-10-CM codes that can be mapped to at least one standard SNOMED CT concept. There are a total of 128,943 mapping pairs between ICD-10-CM and SNOMED CT.

3.3 Evaluation

3.3.1 Code set evaluation

We first evaluated the effectiveness of our SNOMED CT-driven cohort query system by assessing the query code sets. We conducted the evaluation based on each standard SNOMED CT concept. Figure 4 demonstrates the overall results

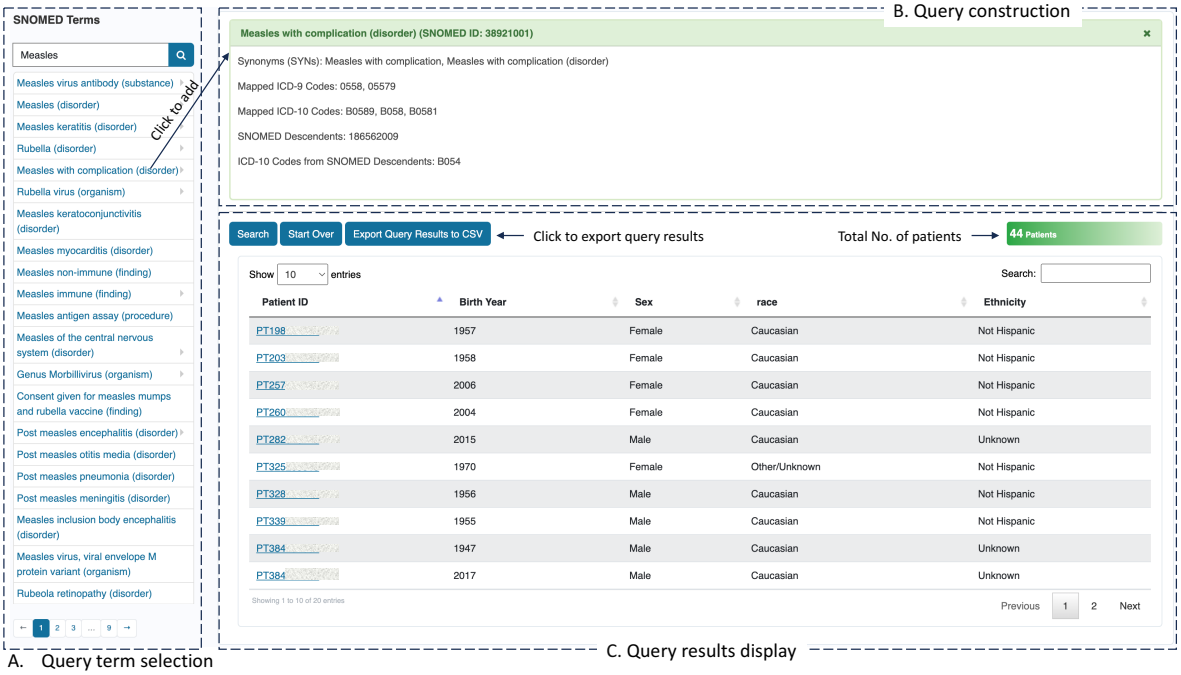

Figure 3: Screenshot of the SNOMED CT-driven cohort query system. This example queries the SNOMED term: "38921001: Measles with complication (disorder)."

of the code set evaluation. For ICD-9-CM, for the 10,741 cases where a SNOMED CT code at least has one ICD-9-CM code that maps to it, there are 7,959 (74.10%) cases where $C_1 = C_2$; while in 2,782 (25.90%) out of 10,741 cases, C_1 is a proper subset of C_2. Out of these 2,782 cases, there are 2,568 non-contradictory cases and 214 contradictory cases. For ICD-10-CM, among the 14,011 cases where a SNOMED CT code at least has one ICD-10-CM code that maps to it, there are 10,255 cases (73.19%) where the baseline code set C_1 is identical to the extended code set C_2; while in 3,756 (26.81%) out of 14,011 cases, C_1 is a proper subset of C_2. Out of these 3,756 cases, there are 3,407 non-contradictory cases and 349 contradictory cases.

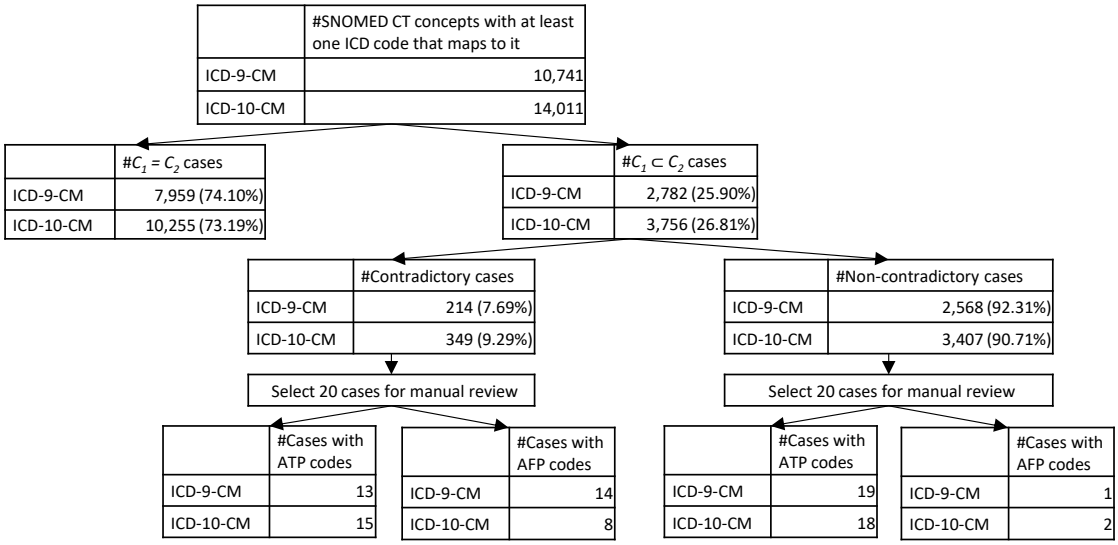

Figure 4: Results of code set evaluation.

We randomly selected 20 contradictory cases and 20 non-contradictory cases in ICD-9-CM, as well as 20 contradictory cases and 20 non-contradictory cases in ICD-10-CM and conducted a manual review to establish our reference standard identifying ATPs and AFPs. The results demonstrate that contradictory cases is often indicative of AFP codes in the extended code set. For ICD-9-CM, only 1 out of 20 non-contradictory cases indicated AFPs in the extended code set while 14 out of 20 contradictory cases indicated AFPs. For ICD-10-CM, only 2 out of 20 non-contradictory cases indicated AFPs in the extended code set while 8 out of 20 contradictory cases indicated AFPs.

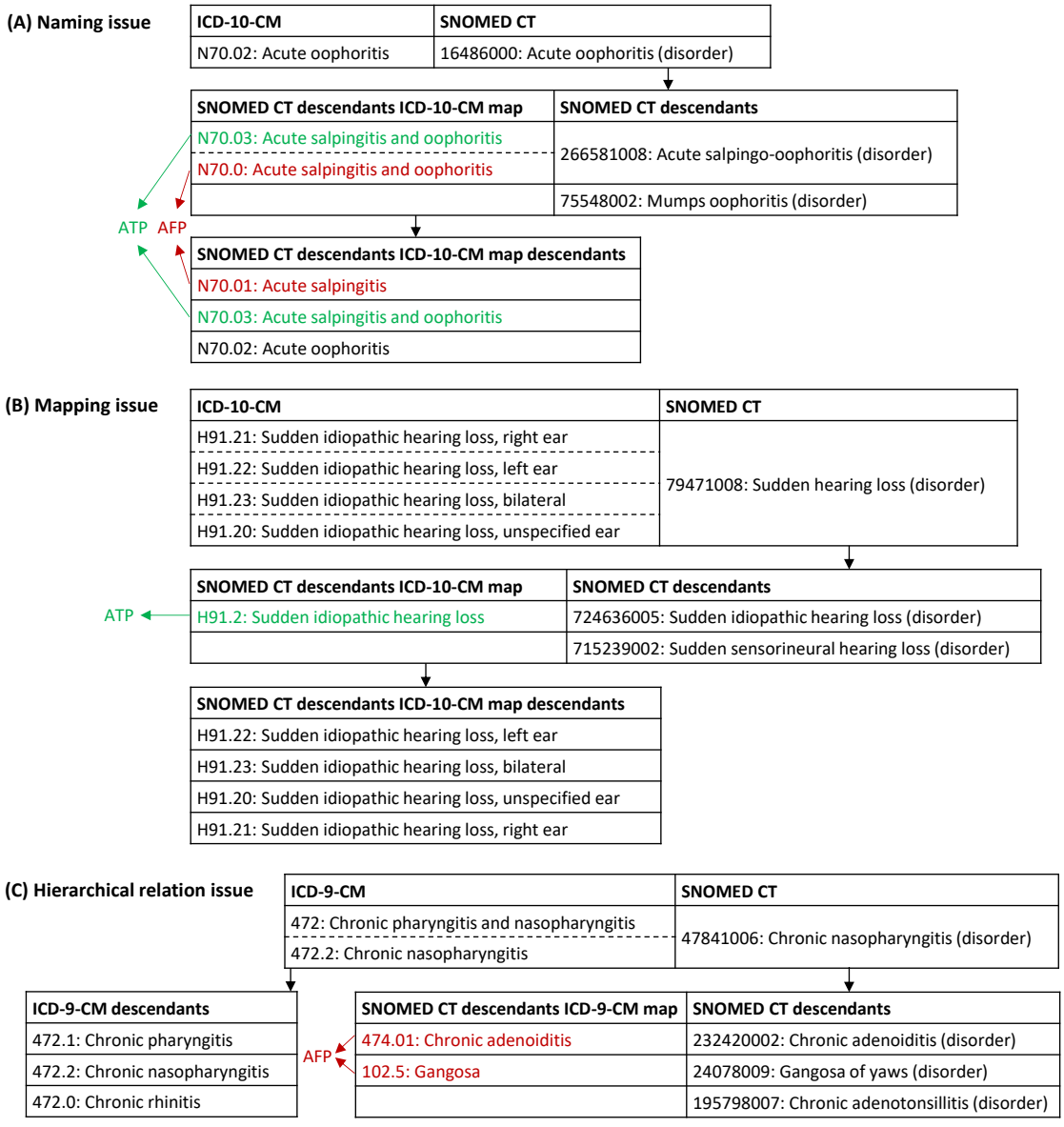

Figure 5: Three types of issues identified in the vocabularies and mappings.

In addition, the manual review also demonstrates the advantage of incorporating additional true codes for a more comprehensive query using a SNOMED CT-driven cohort query system. For ICD-9-CM, 19 out of 20 non-contradictory cases indicated ATPs in the extended code set while 13 out of 20 contradictory cases indicated ATPs. For ICD-10-CM, 18 out of 20 non-contradictory cases indicated ATPs in the extended code set while 15 out of 20 contradictory cases indicated ATPs. Note that one case could contain both ATPs and AFPs in the extended code set. From the perspective

of the number of codes in these 80 cases, overall there are 276 ICD codes in the baseline code set C_1, and 487 ICD codes in the extended code set C_2. For the 211 APs in C_2 comparing to C_1, there are 148 ATPs and 63 AFPs.

Our manual evaluation also revealed issues existing in the vocabularies and mappings. We summarized them into three types: (1) code naming issue; (2) mapping issue; and (3) hierarchical relation issue. For example, Figure 5(A) demonstrates a code naming issue in ICD-10-CM. Code "N70.0: Acute salpingitis and oophoritis" could be more precisely named as "N70.0: Acute salpingitis and/or oophoritis", since "N70.01: Acute salpingitis", "N70.02: Acute oophoritis", and "N70.03: Acute salpingitis and oophoritis" are all its subtypes. Precisely naming it would also prevent incorrect mapping to the standard SNOMED CT concept and thus prevent incorrect patient cohort query. Figure 5(B) demonstrates a potential mapping issue. Code "H91.2: Sudden idiopathic hearing loss" is mapped to SNOMED CT concept "724636005: Sudden idiopathic hearing loss (disorder)", while its four descendants are mapped to concept 724636005's ancestor "79471008: Sudden hearing loss (disorder)" in SNOMED CT. This is a contradictory case where the mapping contradicts the hierarchical structures of the vocabularies. Although it did not cause any AFP, it reveals a potential mapping issue as the four descendants of "H91.2" should be mapped to "724636005". Figure 5(C) demonstrates a potential hierarchical relation issue in SNOMED CT, since "232420002: Chronic adenoiditis (disorder)" and "24078009: Gangosa of yaws (disorder)" are not valid subtypes of "47841006: Chronic nasopharyngitis (disorder)". Such incorrect hierarchical relations may cause AFP in the code set and therefore lead to incorrect patient identification.

3.3.2 Patient query evaluation

For our manually evaluated 80 cases, we identified 211 APs within our extended code set comprising 148 ATPs and 63 AFPs. However, these figures alone do not fully capture the impact of the SNOMED CT-driven system on actual patient cohort queries. Thus, we assessed how the ATP and AFP codes influenced patient queries within the Optum® COVID-19 data. Table 1 illustrates ten examples of patient cohort query of cases among the 80 cases we manually reviewed. For instance, in the extended code set {572, 572.0, 572.1, 572.2, 572.3, 572.4, 572.8}, ICD-9-CM codes "572.1", "572.2", "572.3", and "572.4" were identified as ATPs, whereas "572" and "572.0" were identified as AFPs, contrasting with the baseline code set {572.8}. Leveraging the baseline code set alone resulted in 3,429 patients being retrieved, while the application of the extended code set led to 19,978 patients retrieved. Notably, the ATP codes within the extended set were responsible for accurately identifying 15,576 additional true patients, whereas the AFP codes resulted in the retrieval of 1,002 additional false patients. This underscores the effect of ATP and AFP codes in the extended code set generated by the SNOMED CT-driven query system on patient cohort queries in this EHR database. The results of the 80 randomly reviewed code sets and their corresponding patient cohort query results are available at https://github.com/XubingHao/AMIA2025Summit-SCTQuery.

Table 1: Ten examples of patient cohort queries.

Baseline code set C_1	**#Patients with C_1**	**Extended code set C_2**	**#Patients with C_2**	**ATP codes**	**#Patients with ATPs**	**AFP codes**	**#Patients with AFPs**
{513.0}	1,597	{513, 513.0, 513.1, 006.4}	2,117	{006.4}	1	{513, 513.1}	519
{P11, P11.0, P11.1, P11.2, P11.3, P11.4, P11.5, P11.9}	341	{P11, P11.0, P11.1, P11.2, P11.3, P11.4, P11.5, P11.9, P10.8, P10.9}	356	{P10.8, P10.9}	15	{}	0
{N41.0, N41.00, N41.01}	18,071	{N41.0, N41.00, N41.01, A59.02}	18,107	{}	0	{A59.02}	36
{813.8, 813.80, 813.81, 813.82, 813.83}	12,405	{813.8, 813.80, 813.81, 813.82, 813.83, 813.08, 813.23, 813.44}	18,090	{813.08, 813.23, 813.44}	5,685	{}	0
{G43.01, G43.011, G43.019}	33,890	{G43.01, G43.011, G43.019, G43.71, G43.711, G43.719}	56,121	{G43.71, G43.711, G43.719}	22,231	{}	0
{N70.02}	49	{N70.0, N70.01, N70.02, N70.03}	897	{N70.03}	282	{N70.0, N70.01}	591
{799.8, 799.81, 799.82, 799.89}	54,621	[{780.7, 780.71, 780.72, 780.79, 799.8, 799.81, 799.82, 799.89}	673,473	{780.7, 780.71, 780.72, 780.79}	618,852	{}	0
{572.8}	3,429	{572, 572.0, 572.1, 572.2, 572.3, 572.4, 572.8}	19,978	{572.1, 572.2, 572.3, 572.4}	15,576	{572, 572.0}	1,002
{681.0, 681.00, 681.01, 681.02}	28,687	{681, 681.0, 681.1, 681.9, 681.00, 681.01, 681.02, 681.10, 681.11}	69,820	{}	0	{681, 681.1, 681.9, 681.10, 681.11}	41,133
{J35.02}	2,739	{J35.0, J35.01, J35.02, J35.03}	33,738	{J35.03}	8,388	{J35.0, J35.01}	24,686

4 Discussion

In this work, we developed a SNOMED CT-driven cohort query system over the heterogeneous Optum® COVID-19 data leveraging concept mappings between ICD-9-CM/ICD-10-CM and SNOMED CT from OMOP vocabularies. Our system simplifies the patient cohort identification process and improves the efficiency of data uses in EHR data. Our study illustrates the process of mapping between different terminology standards adds complexity to healthcare data management. The identification of falsely included codes and patients emphasizes the need for continuous efforts to refine these processes and improve the integration of diverse coding systems. While SNOMED CT-driven query systems hold considerable promise for comprehensive patient cohort queries, careful attention must be given to the challenges of falsely included codes and patients. In the future, we plan to incorporate additional coding vocabularies and expand our system to a broader range of EHR data sets. We also plan to conduct a detailed use case study to further demonstrate the practical applications and benefits of the system in real-world clinical and research settings.

In this study, we adopted the OHDSI Standard Vocabularies because of their regular updates, open accessibility, and strong support for the interoperability across healthcare data sources. The OHDSI mappings, where SNOMED CT consistently serves as the standard concept, ensure the accuracy of our queries by aligning ICD codes with SNOMED CT concepts. While alternative mappings such as those from the National Library of Medicine (NLM) or the Unified Medical Language System (UMLS) provide valuable resources, they are less suitable for our downward query strategy as SNOMED CT may not always serve as the standard concept. This discrepancy could potentially lead to the inclusion of inaccurate records in our queries.

In the future, we plan to conduct a statistical analysis to explore the relationship between the level of the concept in its vocabulary and the number of patients retrieved in the EHR data. The level of a concept is defined by the number of hops along the longest path from the root of the vocabulary to the concept itself[22]. In our study, we leveraged a downward query strategy, wherein general concepts with a small level in the vocabulary having many descendants, are anticipated to yield a larger patient cohort upon query execution. We hypothesize that the more general a concept is, the greater the influence of APs on the size of the patient cohort derived from the query, thereby affecting the total number of patients retrieved.

Our SNOMED CT-driven cohort query system offers the distinct advantage of covering patient diagnosed with certain condition, irrespective of which coding vocabulary the patient is recorded in the EHR data. To facilitate the comparison between the baseline code set and extended code set, in this paper, we only included ICD codes, and we have reported patient counts based solely on these ICD codes. Given our system's capability to also integrate SNOMED CT concepts, we intend to expand our future analysis by including SNOMED CT codes within our code sets. This will enable us to execute a more thorough and comprehensive comparison.

In addition, our analysis revealed that 253 ICD-9-CM codes lack mapped standard SNOMED CT concepts in the OMOP concept mappings. Among these, 199 codes have ancestors in ICD-9-CM that also lack standard SNOMED CT mappings, excluding them from coverage by our SNOMED CT-driven query system when employing a downward query strategy. Similarly, 1,593 codes from ICD-10-CM were identified as not having standard SNOMED CT mappings, with 387 of these codes' ancestors in ICD-10-CM also lacking such mappings and thereby not covered by the system. In the future, we plan to quantify the impact of these unmapped codes on patient cohort queries. Additionally, we plan to delve into the reasons behind the absence of standard SNOMED CT mappings for these concepts in the OHDSI concept mappings.

5 Conclusion

In this work, we developed a SNOMED CT-driven cohort query system over the heterogeneous Optum® COVID-19 data leveraging concept mappings between ICD-9-CM/ICD-10-CM and SNOMED CT from OMOP vocabularies. We evaluated the benefits and challenges of using SNOMED CT to perform cohort queries over this dataset based on both the query code sets and the actual patient cohorts retrieved from the dataset. Manual review of 80 random cases revealed 65 cases containing 148 true positive codes and 25 cases containing 63 false positive codes in the code sets. In summary, our research demonstrates that SNOMED CT-driven query systems is promising for conducting comprehensive patient cohort queries. However, it is crucial to address the challenges associated with the inclusion of incorrect codes and patients, ensuring the accuracy and reliability of query results, which are fundamental for

advancing clinical research and patient care practices.

Acknowledgment

This work was supported by the National Science Foundation (NSF) through grant 2047001 and National Institutes of Health (NIH) through grant R01NS116287. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

References

1. Overview of SNOMED CT;. (Online; accessed September, 2024). https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html.
2. Donnelly K, et al. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. 2006;121:279.
3. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574.
4. International Classification of Diseases,Ninth Revision, Clinical Modification (ICD-9-CM);. (Online; accessed September, 2024). https://archive.cdc.gov/#/details?url=https://www.cdc.gov/nchs/icd/icd9cm.htm.
5. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM);. (Online; accessed September, 2024). https://www.cdc.gov/nchs/icd/icd-10-cm/?CDC_AAref_Val=https://www.cdc.gov/nchs/icd/icd-10-cm.htm.
6. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association. 2024;31(3):583-90.
7. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. Journal of the American Medical Informatics Association. 2014;21(e1):e11-9.
8. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. Journal of the American Medical Informatics Association. 2021;28(9):2017-26.
9. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. Journal of biomedical informatics. 2014;52:78-91.
10. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. In: AMIA annual symposium proceedings. vol. 2007. American Medical Informatics Association; 2007. p. 548.
11. ATLAS;. (Online; accessed September, 2024). https://github.com/OHDSI/Atlas.
12. Jung H, Lee HY, Yoo S, Hwang H, Baek H. Effectiveness of the Use of Standardized Vocabularies on Epilepsy Patient Cohort Generation. Healthcare Informatics Research. 2022;28(3):240.
13. Tavakoli K, Kalaw FGP, Bhanvadia S, Hogarth M, Baxter SL. Concept coverage analysis of ophthalmic infections and trauma among the standardized medical terminologies SNOMED-CT, ICD-10-CM, and ICD-11. Ophthalmology Science. 2023;3(4):100337.
14. Willett DL, Kannan V, Chu L, Buchanan JR, Velasco FT, Clark JD, et al. SNOMED CT concept hierarchies for sharing definitions of clinical conditions using electronic health record data. Applied clinical informatics. 2018;9(03):667-82.
15. Cho S, Sin M, Tsapepas D, Dale LA, Husain SA, Mohan S, et al. Content coverage evaluation of the OMOP vocabulary on the transplant domain focusing on concepts relevant for kidney transplant outcomes analysis. Applied Clinical Informatics. 2020;11(04):650-8.
16. Kim Y, Zhu L, Zhu H, Li X, Huang Y, Gu C, et al. Characterizing cancer and COVID-19 outcomes using electronic health records. PLoS One. 2022;17(5):e0267584.
17. MongoDB;. (Online; accessed August, 2024). https://www.mongodb.com/.
18. Huang Y, Li X, Zhang GQ. ELII: A novel inverted index for fast temporal query, with application to a large Covid-19 EHR dataset. Journal of Biomedical Informatics. 2021;117:103744.
19. Hripcsak G, Levine ME, Shang N, Ryan PB. Effect of vocabulary mapping for conditions on phenotype cohorts. Journal of the American Medical Informatics Association. 2018;25(12):1618-25.
20. OHDSI: Mapping of Concepts;. (Online; accessed August, 2024). https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:mapping.
21. Bächle M, Kirchberg P. Ruby on rails. IEEE software. 2007;24(6):105-8.
22. Zheng F, Shi J, Cui L. A lexical-based approach for exhaustive detection of missing hierarchical IS-A relations in SNOMED CT. In: AMIA Annual Symposium Proceedings. vol. 2020. American Medical Informatics Association; 2020. p. 1392.