Matching Patient Records to Clinical Trials Using Ontologies

Chintan Patel[3], James Cimino[3], Julian Dolby[1], Achille Fokoue[1], Aditya Kalyanpur[1], Aaron Kershenbaum[1], Li Ma[2], Edith Schonberg[1], and Kavitha Srinivas[1]

[1] IBM Watson Research Center,P.O.Box 704, Yorktown Heights, NY 10598, USA
dolby, achille, adityakal, aaronk, ediths, ksrinivs@us.ibm.com
[2] IBM China Research Lab, Beijing 100094, China
malli@cn.ibm.com
[3] Columbia University Medical Center
chintan.patel, ciminoj@dbmi.columbia.edu

Abstract. This paper describes a large case study that explores the applicability of ontology reasoning to problems in the medical domain. We investigate whether it is possible to use such reasoning to automate common clinical tasks that are currently labor intensive and error prone, and focus our case study on improving cohort selection for clinical trials. An obstacle to automating such clinical tasks is the need to bridge the *semantic gulf* between raw patient data, such as laboratory tests or specific medications, and the way a clinician interprets this data. Our key insight is that matching patients to clinical trials can be formulated as a problem of semantic retrieval. We describe the technical challenges to building a realistic case study, which include problems related to scalability, the integration of large ontologies, and dealing with noisy, inconsistent data. Our solution is based on the SNOMED CT® ontology, and scales to one year of patient records (approx. 240,000 patients).

1 Introduction

This paper describes a large case study that explores the applicability of ontology reasoning to problems in the medical domain. Currently, medical ontologies are primarily used for terminology services. We explore whether it is possible to use ontology reasoning to automate common clinical tasks, such as cohort selection of patients for clinical trials, infectious disease monitoring, and clinical decision support. An obstacle to automating these tasks is the need to bridge the *semantic gulf* between raw patient data, such as laboratory tests or specific medications, and the way a clinician interprets this data. For example, a laboratory report which indicates the presence of a class of organisms implies the presence of an infectious disorder; similarly, certain types of chemotherapy drugs imply the presence of certain cancers. Using ontologies, it should be possible to automate this interpretation process and build a reusable solution. Toward this goal, we focus our case study on the problem of cohort selection for clinical trials.

Low participation in clinical trials is a significant problem in clinical and translational research, where participation rates range between 5%-10% for most trials [1]. A key deterrent to participation is that matching patients to clinical trials is currently a manual, physician-driven process. Automating this process has shown some promising results in terms of increased patient referrals from physicians [2]. However, current efforts at automation require the development of custom applications.

The SNOMED CT® ontology [3], which formally defines classes of disorders, drugs, and organisms, is well suited for our case study to see whether ontologies can help automate the problem of cohort selection. Our primary insight is that matching patients to clinical trials can be formulated as a problem of semantic retrieval, i.e., a clinical trial criterion can be expressed as a semantic query, which a reasoner can then use together with SNOMED CT to infer implicit information that results in retrieving eligible patients.

Our goal in this study is to assess the feasibility of this approach in a realistic scenario. The technical challenges fall primarily into three categories: knowledge engineering, scalability, and noisy data, each of which is described below.

Knowledge Engineering. A key challenge is to combine the legacy patient data with existing ontologies such as SNOMED CT to demonstrate the value of ontology matching for cohort selection. The following examples illustrate this problem:

- There are currently 39 clinical trials [4] that specify *warfarin medication* as an inclusion criterion. SNOMED CT has the names of generic drug concepts, which are in turn described in terms of their active ingredients, such as warfarin. However, the patient record contains only the names of vendor-specific drugs. What is needed here is a mapping from vendor-specific drug names to generic drug concepts, to allow an inference about active ingredients of drugs.
- There are 26 clinical trials that specify *Methicillin-resistant Staphylococcus aureus (MRSA) disorder* as an inclusion criterion for the trial. SNOMED CT defines *MRSA disorder* as a disorder that indicates the class of MRSA organisms as a causative agent. However, the patient record contains institution-specific laboratory tests that indicate only the presence or absence of a particular organism (e.g., MRSA organism) in institution-specific terminologies. What is needed here is a mapping of the presence or absence of the organism to whether its corresponding SNOMED CT equivalent term is a causative agent or not.
- There are currently 6240 trials that refer to disorders that involve different types of *neoplasms*. SNOMED CT classifies 1522 different types of morphologies as neoplasms. However, the patient record contains information about a specific radiology test that indicates the presence of a certain morphology in a certain body part, all coded in local terminologies. Once again local terms for body parts and morphologies need to be mapped to their SNOMED CT counterparts.

It is clear from these examples that a key knowledge engineering task is to map patient record terms to concepts in the SNOMED CT model. This mapping process is not simply a matter of establishing equivalences, which is itself non-trivial for large terminologies. The local terminology is often coded as a taxonomy, so there is the additional difficult problem of ontology integration [5]. Because each health care institution codes patient data using an idiosyncratic local terminology, mapping to the SNOMED CT model requires customization per health care institution. Fortunately, while this task is a significant effort, it only has to be performed once per institution, and is reusable for solving different clinical problems.

Scalability. Another key challenge is the need for reasoning over ontologies that are very large and expressive. The size of the knowledge base for the clinical trials case study far exceeds the capabilities of most reasoners. There are several reasoners that are designed to handle large Tboxes (e.g., Fact++ [6], Pellet [7], Racer [8]). Other reasoners scale to large Aboxes in secondary storage (e.g. Kaon2 [9], SHER [10]). The combination of a large Abox and a large Tbox required for this case study, however, far exceeds the size of the knowledge bases that have been tested so far with these reasoners.

Another factor is the expressivity needed for solving the clinical problem. While SNOMED CT is modeled within the EL++[11] formalism (intersections, existential restrictions, role hierarchies), negation and universal restrictions are inherent in the patient data and in the queries. As an example, negation of complex concepts is an important aspect of the patient record, e.g., when pneumonia has been ruled out on the basis of a radiology report. Similarly, clinical trials exclusion criteria are negations of complex concepts, which means that the solution requires the expressivity of OWL-DL.

Noisy, incomplete data The third challenge is that clinical data tends to be incomplete and noisy. SNOMED-CT contains complete definitions for disorders including both information needed to infer the presence of the disease, and also information to relate the disorder to other disorders. However, patient data contains only information needed to infer the disease.

Clinical data is also inconsistent from a logical perspective. It is not uncommon for a laboratory test to contain both positive and negative findings. To perform semantic retrieval, current reasoners assume that the data is consistent. Therefore, cleansing the data efficiently is another open issue.

In the rest of this paper, we present our solutions to these technical challenges, and summarize the results for matching 9 clinical trial criteria against a knowledge base with 59 million Abox assertions and 22,561 Tbox assertions. The clinical trials case study is described in more detail in Section 2. Sections 3-5 present the technical challenges and issues that we faced, and how they were resolved. Section 6 gives results and validation, and Section 7 draws conclusions.

2 Case Study Description

The architecture for retrieving patients eligible for clinical trials is shown in Figure 1. Clinical trial criteria are formulated as queries, and a reasoner matches the queries against a knowledge base to retrieve eligible patients. We use the SHER reasoner, which implements the techniques in ([12], [10]) for scalable Abox reasoning. The first steps in creating this solution are constructing a knowledge base Tbox, based on SNOMED CT, and an Abox from structured patient records. For our case study, we use one year of anonymized patient records from Columbia University Medical Center.

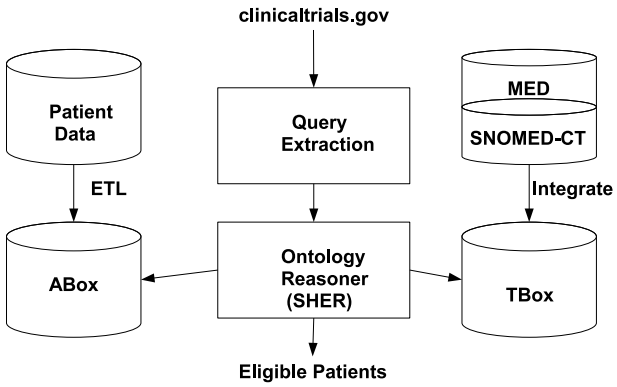

Fig. 1. Case Study Solution Architecture

Constructing the Tbox requires integrating the terminology used in the patient data and SNOMED CT terminology. The Columbia patient data are encoded in a frame-based semantic network called MED [13]. We considered only the MED taxonomy which consists of 100,212 concepts that capture the organism, disease, and medical test hierarchies. SNOMED CT has 379,630 concepts which include organism, pharmaceutical product, specimen, body structure, clinical findings, and procedures. SNOMED CT is not just a taxonomy; 217,619 of SNOMED CT concepts are defined in terms of existential restrictions. Such definitions allow the inferencing of disorders from relationships in the Abox such as associated morphology, finding site, and causative agents.

Constructing the Abox requires translating records encoded in the MED taxonomy into a set of assertions encoded in SNOMED CT in SHER's relational store. The patient database [14], which stores the raw data for the Abox, includes a single table of clinical events, where each event consists of one or two records. The events used in the case study correspond to laboratory test results, radiology findings, and drug treatment. We use an Extract-Transform-Load (ETL) process to transform the patient events into assertions compatible with SNOMED CT.

The queries themselves are extracted from clinical trial criteria found on [4], where the criteria are expressed as text. We convert the text-based queries into logical DL queries, which use SNOMED CT concepts.

Sections 3-5 describe the major technical challenges encountered in implementing the case study.

3 Knowledge Engineering

3.1 Mapping MED to SNOMED CT

To create the Tbox, the first step is to map concepts in MED to the concepts in SNOMED CT. Our goal is to achieve a high degree of accuracy and coverage through a semi-automated process:

1. **Existing Mappings:** Many of the concepts in both MED and SNOMED CT are mapped to the concepts in the Unified Medical Language System (UMLS®) [15]. Therefore, it is possible to use UMLS as an intermediary target, mapping MED to UMLS to SNOMED CT for a subset of MED concepts.
2. **NLP-based Mapping:** We next use the medical Meta Map tool (MMTx) [16] to map natural language strings associated with MED concepts to UMLS concepts, and then to SNOMED CT concepts when possible. Only mappings with a perfect score on MMTx are retained.
3. **Prefix Removal:** Some MED concept strings contain institution-specific prefixes, such as NYPH. We remove these prefixes to increase the number of perfect matches on MMTx.
4. **Manual mapping:** Vendor-specific drugs in MED do not have a mapping to a generic mapped drug concept in SNOMED CT; these 1000 concepts are manually mapped by domain experts (co-authors CP and JC).

This procedure maps 17,446 out of 100,212 MED concepts to SNOMED CT. The next step, described in Section 3.3, dramatically increases the coverage by including assertions corresponding to the MED taxonomy.

3.2 Validation of Mappings

To determine the accuracy of mapping MED to SNOMED CT, domain experts (co-authors CP and JC) analyzed the mapping results. Since both MED and SNOMED CT share a common upper level ontology (Semantic Network in UMLS), it is possible to determine whether each of the mapped concepts belong to the same conceptual category. These categories were further combined into semantic groups. An example of a valid mapping is the MED concept *fibromyalgia*, which has an upper level concept in UMLS of *Finding*, mapped to the SNOMED CT concept *Primary fibromyalgia syndrome*, which has an upper level concept of *Disease or Disorder*. Both *Finding* and *Disorder* belong to the same semantic group. By this approach, 2,534 invalid mappings were found with mismatching

source and target upper level concepts and semantic groups. Manual inspection of invalid mappings revealed that the majority (all but 11) are caused by errors in categorizing MED concepts in terms of the upper UMLS concepts, e.g. 768 are caused by a single missing parent type of *DRUG* in MED. The faulty 11 are true false positives, and are eliminated from our mapping. False negatives from unmapped concepts are discussed in the next section.

For each mapped MED concept, a subclass assertion is added to the Tbox to relate it to its mapped SNOMED concept. We use subclass rather than equivalence assertions because the current mapping between MED and SNOMED is not guaranteed to be sufficiently precise to warrant equivalence. However, without concept equivalence, negated queries fail, since we cannot infer that the negation of a MED concept is a subclass of the negation of its corresponding SNOMED CT concept. A more precise MED to SNOMED mapping will eliminate this issue, and this is an issue for future work.

3.3 Integrating the MED Taxonomy with SNOMED CT

Although we succeeded in mapping 17,446 MED concepts, this constitutes only 17% of the MED Tbox. In terms of the 13,313 MED concepts referred to in the Abox of one year patient data, only 9% had a direct mapping to a SNOMED CT concept.This reflects the fact that the patient data is coded in institution-specific MED concepts that do not have direct mappings to SNOMED CT concepts. However, since many of their super-concepts map to SNOMED CT concepts by our mapping process, we can significantly increase coverage by adding subclass assertions corresponding to the MED taxonomy. If we include the subclasses of the mapped MED concepts, we increase coverage of the MED Tbox to 75,514 concepts. For the Abox, including subclasses of mapped MED concepts increases coverage from 9% to 88% (11,732 concepts).

3.4 Abox Construction

To construct a SNOMED CT Abox from the one year patient data, we transform the existing relational patient database with implicit relationships into membership and role assertions corresponding to SNOMED CT. As an example of such a mapping, if a patient record states that the patient is on drug *Cerner Drug: Lactulose Syrp 20G/30ml*, it needs to be transformed into the appropriate SNOMED CT role assertion between the patient and the drug. We use the attribute *administeredSubstance* as the relationship and the drug itself is mapped to *Lactulose* in SNOMED CT in the Tbox.

Patient data transformation performs several critical functions:

- In the clinical domain, negative findings for medical tests and procedures are crucial in selection for clinical trials and clinical decision support. Therefore, negative results in the patient data should be modeled using logical negation. The transformation process extracts positive and negative results from the patient record and makes them explicit.

– In the clinical domain, results of laboratory tests and findings form logical groupings (e.g., a specific laboratory test indicates an organism as well as the source specimen for the test). Disorders in SNOMED CT capture such groupings by nesting existential restrictions as illustrated by the SNOMED CT definition of $Breast\ Neoplasm$ below:
$\exists roleGroup.(\exists hasMorphology.Neoplasm \sqcap \exists hasFindingSite.Breast)$.
We therefore model groups of events using the SNOMED CT $roleGroup$ attribute, as discussed in the examples below.

The Abox construction process is driven off of set of transformation rules, derived by abstracting implicit information models for both the patient database and SNOMED CT. Fortunately, the structure of the data and these information models are relatively simple, so that the number of rules is small. Table 1 illustrates two radiology rules. These rules generate unique individuals p and e in

Radiology Event Template	Abox Assertion Templates
$?PatientID, ?TimeStamp,$ $?Morphology, ?BodyPart, ?HighCertainty$	individuals: p, e, r $assocObservation(p, e)$ $roleGroup(e, r)$ $hasTimeStamp(e, ?TimeStamp)$ $r : \exists hasMorphology.?Morphology$ $r : \exists findingSite.?BodyPart$
$?PatientID, ?TimeStamp,$ $?Morphology, ?LowCertainty$	individuals: p, e, r $assocObservation(p, e)$ $roleGroup(e, r)$ $hasTimeStamp(e, ?TimeStamp)$ $r : \forall hasMorphology.\neg ?Morphology$

Table 1. Transformation Rules for Radiology Events

the Abox, representing each unique patient and event. A unique individual r is generated to represent the grouping of the associated Morphology and BodyPart of an event. New relationship assertions are generated to associate p with e, and e with r.

The first rule transforms a positive morphology finding and associated body site, and the second rule transforms a negative morphology finding, in which case there is no associated body site. In the positive case, the first rule adds membership assertions with existential restriction concepts to the Abox, associating r with the morphology and the body site. In the negative case, the second rule adds a membership assertion with a universal restriction concept that includes negation to the Abox.

Table 2 shows examples of rule instantiation. The transformation rules are engineered to match SNOMED CT definitions. For example, the SNOMED CT definition of $Breast\ Neoplasm$ above typifies SNOMED CT rules for radiology findings. Note that a query for patients testing positive for breast neoplasm will

Radiology Event	Abox Assertions
Patient43, 3.15.2006, Malignant Neoplasm, Breast, High Certainty	individuals: $p43, e1, r1$ $assocObservation(p43, e1)$ $roleGroup(e1, r1)$ $hasTimeStamp(e1, 3.15.2006)$ $r1 : \exists hasMorphology.Malignant\ Neoplasm$ $r1 : \exists findingSite.Breast$
Patient32, 12.01.2005, Malignant Neoplasm, Low Certainty	individuals: $p32, e2, r2$ $assocObservation(p32, e2)$ $roleGroup(e2, r2)$ $hasTimeStamp(e2, 12.01.2005)$ $r2 : \forall hasMorphology.\neg Malignant\ Neoplasm$

Table 2. Sample Radiology Event Transformations

match the first patient in Table 2, and a query for patients testing negative will match the second patient.

4 Scalability

4.1 Dealing with Large Aboxes

SHER embodies techniques [10],[12] which use summarization and refinement to achieve scalable Abox reasoning. Specifically, a summary Abox is constructed from the original Abox. The initial summary Abox is built by mapping all instances of the same type in the original Abox to a single instance in the summary Abox. For example, all instances of *Malignant Neoplasm* in the original Abox are represented by a single instance of *Malignant Neoplasm* in the summary Abox. SHER first checks the summary Abox for any inconsistencies in the knowledge base, using Pellet tableau-based reasoner [7] for consistency checking. If the summary is consistent, then the original Abox must be consistent (for technical detail, see [12]). However, the converse is not true. If any inconsistencies exist, then the reasoner finds their justifications (i.e., the minimal set of assertions responsible for the inconsistency), and tries to selectively refine summary instances in these justifications. Refinement is the process of splitting the summary instance by the sets of role assertions that are present in the original Abox for the individuals mapped to the given summary instance. This iterative process of refinement ends when the summary is consistent, or the justifications cannot be refined any more. If the knowledge base is inconsistent, SHER provides a set of justifications that can be used to cleanse the knowledge base of inconsistencies.

To answer a query, the negation of the query is added to the concept set of each instance in the summary Abox, and the same iterative refinement process is followed. During this process, a map from refined individuals in the summary Abox to individuals in the original Abox is maintained. When the process converges, query results are obtained from this mapping. Initially, SHER could not

scale to the case study with such a large Abox. The problem was in the refinement step: the map from refined individuals to real individuals was kept in memory. To achieve scalability, the refinement mapping is now maintained in the database. In fact, the refinement process is performed entirely by database operations.

4.2 Dealing with Large Tboxes

Even though the MED-SNOMED CT integrated Tbox has a total of 523,368 subclass or equivalence assertions, we do not need all of these for reasoning. As described in [12], the techniques used in SHER are based on taking the closure of the Abox, which informally is the set of concepts that are present in the Abox, either directly or indirectly through assertions in the Tbox. For query answering, the closure of the query concept must also be included.

More specifically, we compute a subset of the MED-SNOMED CT integrated Tbox using the following procedure: (a) We use the FACT++ [6] tableau reasoner to absorb the Tbox to produce a new set of Tbox assertions $\mathcal{T}$ that eliminates any GCIs from the original Tbox. In the case of the MED-SNOMED CT integrated Tbox, no GCIs are left after absorption, and no domain or range constraints are added to the Rbox due to role absorption. (b) We then compute the closure of the Abox $clos(\mathcal{A}, \mathcal{T}, \mathcal{R})$ and queries as defined in [12]. (c) For each concept C in the $clos(\mathcal{A}, \mathcal{T}, \mathcal{R})$, we add the assertions in $\mathcal{T}$ where that concept appears on the left hand side of the assertion. The resulting Tbox has 22,561 assertions, of which 17,319 assertions are related to MED concepts.

5 Noisy, Incomplete Data

The data in the patient records is incomplete with respect to SNOMED CT definitions. For example, suppose a clinical trial criterion is Methicillin resistant Staphylococcus aureus infection (MRSA). The SNOMED CT definition of MRSA is the intersection of three terms:

$\exists hasCausativeAgent.Methicillin\ resistant\ Staphylococcus\ aureus$,
$Infection\ due\ to\ antimicrobial\ resistant\ bacteria$, and
$Infection\ due\ to\ Staphylococcus\ aureus$.

If a patient record contains a positive test for an MRSA organism, a clinician would likely say the patient matches the eligibility criteria. However, the patient record matches only the first term of the SNOMED CT definition. Since there is not information in the patient record that matches the second two conjuncts, which provide definition completeness, the patient will not be retrieved as eligible. We therefore support users specifying which terms of the definition are required, allowing them to tailor the query to match the data that they have. We refer to this as *query weakening*. The patients retrieved are then ranked based on the number of matching terms.

As discussed earlier, clinical data is inconsistent. As an example, two different laboratory tests for the same disorder can result in contradictory results. SHER

is designed to detect multiple inconsistencies in the data efficiently (for technical details, see [12]). We use these algorithms to eliminate inconsistent data before querying it.

6 Evaluation

In this section, give the experimental results of our case study. Our experiments were conducted on a 2-way 2.4GHz AMD Dual Core Opteron system with 16GB of memory running Linux, and we used IBM DB2 V9.1 as our database. Our Java processes were given a maximum heap size of 8GB.

6.1 Validation with a 100 Patient dataset

To validate the clinical correctness of results we first performed an experiment with a randomly selected dataset of 100 patients from a 20 year clinical dataset from the Columbia Medical Center. The 100 patient dataset has 7,451 Tbox subclass assertions, 98,956 type assertions, and 119,206 role assertions.

ClinicalTrials.gov ID	Description
$NCT00084266$	Patients with MRSA
$NCT00288808$	Patients on warfarin
$NCT00393341$	Patients with breast neoplasm
$NCT00419978$	Patients with colon neoplasm
$NCT00304382$	Patients with pneumococcal pneumonia where source specimen is blood or sputum
$NCT00304889$	Patients on metronidazole
$NCT00001162$	Patients with acute amebiasis, giardisis, cyclosporiasis or strongloides...
$NCT00298870$	Patients on steroids or cyclosporine
$NCT00419068$	Patients on corticosteroid or cytotoxic agent

Table 3. Clinical Trial Requirements Evaluated

We selected 9 clinical trials from [4] that query for different types of clinical information (see Table 3). These queries were chosen to cover the domains of laboratory, drug and radiology data. Table 4 shows the DL version of the queries, along with the concepts that were weakened to find solutions, because of the partial information present in the clinical record. The order of the queries Table 4 reflects the order in Table 3. Table 5 shows the queries, the number of patients matched to the queries, whether matches reflect matches to weakened queries, and time to process the queries in seconds. For query $NCT00001162$, the results shown are for the union of 7 different disorders, only 4 of which are illustrated in Table 4.

The matched patients for the 9 clinical trials were manually evaluated by an analysis of the original Columbia database records by one of the authors (CP).

DL Query	Weakened Concept
$\exists associatedObservation.MRSA$	MRSA
$\exists associatedObservation.$ $\exists roleGroup.$ $\exists administeredSubstance.$ $\exists roleGroup.\exists hasActiveIngredient.Warfarin$	None
$\exists associatedObservation.BreastNeoplasm$	Breast Neoplasm
$\exists associatedObservation.ColonNeoplasm$	Colon Neoplasm
$\exists associatedObservation.$ $\left(\begin{array}{l} PneumococcalPneumonia \\ \sqcap \\ \exists hasSpecimenSource.Blood \sqcup Sputum \end{array}\right)$	Pneumococcal Pneumonia
$\exists associatedObservation.$ $\exists roleGroup.$ $\exists administeredSubstance.$ $\exists roleGroup.\exists hasActiveIngredient.Metronidazole$	None
$\exists associatedObservation.$ $\left(\begin{array}{l} acuteamebiasis\sqcup \\ giardisis\sqcup \\ cyclosporiasis\sqcup \\ strongloides\sqcup \\ \ldots \end{array}\right)$	acute amebiasis giardisis cyclosporiasis strongloides ...
$\exists associatedObservation.$ $\exists roleGroup.$ $\exists administeredSubstance.$ $\exists roleGroup.\exists hasActiveIngredient.cyclosporine \sqcup steroids$	None
$\exists associatedObservation.$ $\exists roleGroup.$ $\exists administeredSubstance.$ $\exists roleGroup.\exists hasActiveIngredient.corticosteroid \sqcup cytotoxicAgent$	None

Table 4. DL Queries for Evaluated Clinical Trials

Such an analysis revealed no false positives in the reported matches. In terms of recall, we missed 8 patients on the steroid/corticosteroid queries because the manual mapping of drugs to SNOMED CT missed these mappings. We missed 1 patient for the Metronidazole case. Here, the miss occurred because there were duplicate MED concepts Metronidazole and Metronidazole Preparations, with only the former concept being mapped to SNOMED CT. The missed patient for Metronidazole was because some drugs such as *Cerner Drug: Metronidazole Tab 500 mg.* were subclasses of the unmapped Metronidazole concept. For the breast neoplasm query, our transformation process did not distinguish between disorders and imaging findings in the radiology data, hence the relevant MED concept *Malignant Neoplasm of Breast (Female) Unspecified* was asserted as a finding. We rectified the problem by including a query extension that also looks at the associated findings:

$\exists associatedObservation.\exists associatedFinding.BreastNeoplasm.$

Query	Matched Patients	Time (s)	Weakened Query
$NCT00084266$	1	54	yes
$NCT00288808$	4	78	no
$NCT00393341$	0	29	yes
$NCT00419978$	1	51	yes
$NCT00304382$	0	39	yes
$NCT00304889$	0	29	no
$NCT00001162$	4	225	no
$NCT00298870$	6	117	no
$NCT00419068$	6	118	no

Table 5. Patient Matches for Trial DL Queries for 100 Patients

6.2 Results with the 1 year dataset

The 1 year patient dataset had records for 240,269 patients with 22,561 Tbox subclass assertions, 26 million type assertions, and 33 million role assertions. In the 1 year patient dataset, we had 15 instances of inconsistencies in the data. These inconsistencies were due to (a) laboratory tests that produced contradictory information, for example, positive and negative assertions about organism respiratory syncytial virus by laboratory tests of direct immunofluorescence assay (DFA) and enzyme immunoassay (EIA), (b) modeling errors in MED that resulted in certain MED concepts that were classified as both negative and positive information, for example, the MED concept *Rule Out Specific Organism* was an indirect subclass of both *Positive Organism Comment Result* and *Negative Culture Result*. Since MED is a taxonomy, and does not contain an assertion that *Positive Organism Comment Result* is disjoint with *Negative Culture Result*, this inconsistency was only found when we transformed the Abox to contain assertions about the presence or absence of an organism based on these concepts. The inconsistent data were detected by SHER, and we manually deleted records that resulted in the inconsistencies.

Table 6 shows the queries, the number of patients matched to the queries, the time to process the queries in minutes, and whether the query needed to be weakened to find solutions. Table 6 demonstrates the scalability of reasoning in the SHER engine for a combination of a large Tbox and a large Abox. We do not present any comparison results because no other reasoner we know of can query this dataset. For the use of clinical trial matching, which is currently a manual process, our results show that using ontology matching to automate this task is practical.

7 Discussion

We have presented a feasibility study for an ontology-based approach to match patient records to clinical trials. Using a real world patient dataset, we described various modeling and engineering challenges that we solved, including:

Query	Matched Patients	Time (m)	Weakened Query
$NCT00084266$	1018	68.9	yes
$NCT00288808$	3127	63.8	no
$NCT00393341$	74	26.4	yes
$NCT00419978$	164	31.8	yes
$NCT00304382$	107	56.4	yes
$NCT00304889$	2	61.4	no
$NCT00001162$	1357	370.8	no
$NCT00298870$	5555	145.5	no
$NCT00419068$	4794	78.8	no

Table 6. Patient Matches for Trial DL Queries for 240,269 Patients

- Mapping the MED terminology to SNOMED CT terminology.
- Integrating the MED taxonomy with SNOMED CT.
- Transforming the 1-year patient database into a SNOMED CT Abox.
- Reasoning over a realistic dataset.
- Identifying and eliminating noise in the patient data.
- Dealing with incomplete patient information.

We are continuing to work on making it easier to integrate large ontologies, improving the integration of MED and SNOMED CT, and tuning SHER for scalability.

An interesting problem with respect to clinical data and clinical trials queries is that of open versus closed world reasoning. Description logics and OWL use an open world assumption i.e. if a fact is not explicitly asserted, no assumption is made about the fact, as opposed to a closed world assumption which assumes a fact is negative if not explicitly asserted. In the clinical domain, we need open world reasoning in radiology and laboratory data, because, for example, unless a lab test asserts a negative finding we cannot make arbitrary assumptions about the results. However, in pharmacy data, we can use the closed world assumption to infer that a patient is not on a medication if it is not asserted. Integrating open world with closed world reasoning is a key issue for future consideration [17].

SNOMED CT plays a critical role in the clinical domain; it has been adopted as a national health care standard in the United States and was recently acquired by International Health Terminology Standards Development Organization thereby making it a truly global clinical standard in healthcare. Representing patient data using SNOMED CT has benefits that go beyond the clinical trials matching application. Currently, several decision support systems, infection control systems, public health organizations and regional healthcare information organizations use SNOMED CT merely for terminology services. Our approach provides a means to reuse the knowledge already represented in SNOMED CT to perform semantic retrieval for different biomedical applications.

Acknowledgements

We gratefully acknowledge the help that Bishwaranjan Bhattacharjee provided in tuning our database queries.

References

1. Simes, R.: Clinical trials and real-world medicine. trial evidence best informs real-world medicine when it is relevant to the clinical problem. Med J Aust. **177(8)** (2002) 410–411
2. Embi, P., Jain, A., Clark, J., Bizjack, S., Hornung, R., Harris, C.: Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med. **165(19)** (2005) 2272–2277
3. SNOMED-CT: (http://www.snomed.org)
4. Clinical Trials: (http://clinicaltrials.gov/)
5. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. (2004)
6. Horrocks, I.: Using an expressive descriptive logic:fact or fiction? Proceedings of the 6th Int. Conf. on principles of Knowledge Representation and Reasoning (KR'98) (1998) 636–647
7. Sirin, E., Parsia, B.: Pellet: An owl dl reasoner. In: Description Logics. (2004)
8. Haarslev, V., Moller, R.: Racer system description. Conf. on Automated Reasoning (IJCAR 2001) (2001) 701–705
9. U.Hustadt, Motik, B., Sattler, U.: Reducing shiq description logic to disjunctive datalog programs. (Proc. of 9th Intl. Conf. on Knowledge Representation and Reasoning (KR2004)) 152–162
10. J.Dolby, A.Fokoue, A.Kershenbaum, L.Ma, E.Schonberg, K.Srinivas: Scalable semantic retrieval through summarization and refinement. Proc. of the AAAI Conf (2007) xxx
11. F.Baader, Brandt, S., Lutz, C.: Pushing the el envelope. Technical report, Chair of Automata Theory, Institure for Theoretical Computer Science, Dresden University of Technology (2005)
12. A.Fokoue, A.Kershenbaum, L.Ma, E.Schonberg, K.Srinivas: The summary abox:cutting ontologies down to size. Proc. of the Int. Semantic Web Conf. (ISWC 2006) (2006) 136–145
13. Cimino, J., Clayton, P., Hripcsak, G., Johnson, S.: Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. **1(1)** (1994) 35–50
14. Johnson, S.: Generic data modeling for clinical repositories. J Am Med Inform Assoc. **3(5)** (1996) 328–339
15. DA, D.L., Humphreys, B., McCray, A.: The unified medical language system. Methods Inf Med. **32(4)** (1993) 281–291
16. MMTx: (http://mmtx.nlm.nih.gov/)
17. Grimm, S., Motik, B.: Closed world reasoning in the semantic web through epistemic operators. In: Proc. of the Workshop on OWL: Experiences and Directions (OWLED 2005). (2005)