

Syntactic-Semantic Frames for Clinical Cohort Identification Queries

Dina Demner-Fushman and Swapna Abhyankar

National Library of Medicine, Bethesda, MD
{ddemner, abhyankars}@mail.nih.gov

Abstract. Large sets of electronic health record data are increasingly used in retrospective clinical studies and comparative effectiveness research. The desired patient cohort characteristics for such studies are best expressed as free text descriptions. We present a syntactic-semantic approach to structuring these descriptions. We developed the approach on 60 training topics (descriptions) and evaluated it on 35 test topics provided within the 2011 TREC Medical Record evaluation. We evaluated the accuracy of the frames as well as the modifications needed to achieve near perfect precision in identifying the top 10 eligible patients. Our automatic approach accurately captured 34 test descriptions; 25 automatic frames needed no modifications for finding eligible patients. Further evaluations of the overall average retrieval effectiveness showed that frames are not needed for simple descriptions containing one or two key terms. However, our training results suggest that the frames are needed for more complex real-life cohort selection tasks.

1 Introduction

Cohort identification is an essential phase of clinical research and an active area of medical informatics research. Researchers or clinicians first express cohort characteristics (using clinical language familiar to them) as a free text question which subsequently has to be translated into a machine-understandable query to retrieve the relevant information from electronic clinical data warehouses. The descriptions of the cohort inclusion and exclusion criteria are usually complex and multi-faceted. For example, the ClinicalTrials.gov Protocol Registration System allows up to 15,000 characters for the free-text description of the clinical trial eligibility criteria along age, gender and various conditions axes¹. Traditionally, researchers use formal query languages (directly or with the help of a computer programmer) to query structured clinical data. For example, the Biomedical Translational Research Information System (BTRIS), which is the National Institutes of Health's clinical research data repository, contains pre-defined query templates associated with general retrieval strategies and search filters. Users select the templates relevant to their research question (for example, a lab template for retrieving laboratory test results) and provide the appropriate filter values (such as age, date and specific laboratory test) for the retrieval ([3]).

¹ <http://prsinfo.clinicaltrials.gov/definitions.html>

To facilitate direct cohort selection by clinical researchers, Murphy et al. ([15]) have developed a visual approach within the i2b2 hive. The i2b2 visual query tool displays a hierarchical tree of items for the users to choose from in a “Term” panel; the “Query Tool” panels allow users to combine search terms; and the display widgets show the aggregate numbers of patients who match the query criteria. The visual approach was adopted in the Stanford Translational Research Integrated Database Environment that provides a drag and drop cohort discovery tool ([13]). Deshmukh et al. ([5]) found the visual query tools suitable for generating research cohorts based on simple inclusion/exclusion criteria provided that clinical data is structured, coded and can be transformed to fit the logical data models of the i2b2 hive.

Secondary use datasets are becoming more widely available and contain rich collections of both structured and unstructured data. In many such datasets, essential cohort characteristics are only found in the free-text reports; however, efficiently extracting relevant information from narrative text is challenging. Friedman et al. ([6]) presented a natural language processing (NLP) based method for encoding data from clinical text so that the coded data could subsequently be processed by traditional query tools. However, researchers still have to develop the formal queries using templates or visual query tools. Tu et al. ([19]) have developed a semi-automated method for annotating eligibility criteria using the Eligibility Rule Grammar and Ontology (ERGO). The ERGO annotations were then translated to SQL queries to search an electronic health record database for potentially eligible patients. Alternatively, we propose an information retrieval method that takes a researcher’s cohort selection criteria expressed in familiar clinical language and automatically extracts the relevant concepts and their relationships into a structured question frame to query narrative reports indexed with a search engine. Similar to translating frames into SQL queries needed for searching relational databases, we automatically translate our frames into a search engine query language. The relations between query concepts are preserved through a set of rules that map frame slots to specific clinical report sections and impose search limits (such as the allowed maximal distance between the terms) on predicates.

We hypothesized that an approach using a domain-specific search engine that considers document structure and question-answering and NLP techniques that incorporate both syntax (i.e., structure) and semantic (i.e., meaning) information would yield robust results. Complex question answering that uses: 1) question classes and named entity classes; 2) syntactic dependency information; and 3) semantic information in the form of predicate-argument structures or semantic frames has been successful in open domain question answering ([16]). In this work, we focused on one question class and combined the syntactic dependency, predicate argument structure and named entities information in a single syntactic-semantic frame for answering cohort selection questions. To fully benefit from the semantic processing of the inclusion criteria, we need to structure the patients’ data using an analogous template and unify the patient note frames and the corresponding question frames. In this study, we approximate

structuring the patients' narrative data by splitting the clinical documents into sections that correspond to the template slots (such as **past medical history** and **medications on admission**) and using complex search operators (such as the order of the terms and the distance between them). The Medical Record Retrieval track within the 2011 Text Retrieval Conference (TREC) gave us the opportunity to evaluate this information retrieval method for identifying patient cohorts based on specific inclusion and exclusion criteria ([20]). The cohort descriptions were based on the list of priority topics for comparative effectiveness research issued by the US Institute of Medicine (IOM) of the National Academies ([9]). The proposed query frames could be translated to SQL queries to search over the structured clinical data (if available), as well as to search for eligible patients in free-text reports. The ability to search across both structured and unstructured clinical data will enable complex queries that can identify the most relevant patients.

2 Methods

In this paper, we present the arguably most difficult first step in automatic cohort identification: the automatic generation of question frames from cohort inclusion and exclusion criteria expressed in natural language. We build upon the evidence-based medicine PICO method for asking a well-formed clinical question. Richardson et al. ([17]) first described the PICO method to help clinicians efficiently find the most relevant answers to their clinical questions, and it has been widely incorporated into medical training as part of the evidence-based medicine curriculum. PICO organizes each question into four main parts: 1) **P**atient or **P**roblem; 2) **I**ntervention; 3) **C**omparison intervention (if applicable); and 4) **O**utcome. Previously, we developed an automated method to extract semantic question frames in PICO format for literature-based clinical question answering ([4]). The *frame* is the overall structure that holds all of the relevant concepts for each question, and each frame has four *slots* corresponding to each of the four PICO elements. The system places the relevant concepts from each question into the appropriate PICO slot. In our original work, we modified the PICO format by splitting the **Patient/Problem** slot and adding **Anatomy** to the **Patient** slot, and we merged the **Intervention** and **Comparison** intervention, given that the distinction is not always clear in either the question or in the answer. Inspired by Boxwala et al. ([2]) and Ruiz et al. ([18]), who analyzed query requirements for cohort identification, we further developed our semantic question frame extraction method into a syntactic-semantic method by: 1) refining the basic PICO frame elements with syntactically related words; 2) capturing conjunctions and prepositional phrases; and 3) similar to Jacquemart and Zweigenbaum ([10]), augmenting the basic PICO frame with relational slots that express question elements using predicate-argument structures ([concept]–(relation)–[concept]).

We used 60 training questions created by the second author (SA) to develop our syntactic-semantic method. She based 30 of the training questions on her

Table 1. Syntactic-semantic question frame elements for capturing cohort characteristics

Refined Frame Slot (Basic PICO)	Example of a Refined Frame Slot	Original Question
Age (Patient)	<Age>under50</Age>	patients younger than 50 with hearing loss
Gender (Patient)	<Gender>F</Gender>	women admitted for myocardial infarction who are on hormone-replacement therapy
Population (Patient)	<Population>athletes</Population>	patients seen in the ER with concussion who were athletes
PastMedHx (Problem / Intervention)	<PMH><Prblm> hepatitis</Prblm> <Cause>blood transfusion</Cause></PMH>	patients with a history of hepatitis related to blood transfusion, now with liver cancer
SocialHx (Problem)	<SocialHx>smoking</SocialHx>	patients with a history of smoking as well as personal and family history of lung cancer
AdmitProblem (Problem)	<AdmitPrblm>stroke</AdmitPrblm>	patients admitted for stroke who arrived too late for tPA administration
DischargeProblem (Problem)	<DischPrblm>wound infection</DischPrblm>	patients who developed a wound infection during the current hospital stay
Problem (Problem)	<Prblm>concussion</Prblm>	patients seen in the ER with concussion who were athletes
Finding (Problem)	<Finding>hearing loss</Finding>	patients younger than 50 with hearing loss
Complications_Of (Problem)	<ComplicationsOf> <Prblm>pneumothorax</Prblm> <Cause>VATS</Cause> </ComplicationsOf>	patients who developed a pneumothorax as a complication of VATS
Allergies (Problem)	<Allergies>drug allergy</Allergies>	patients with a known drug allergy who received a drug in the same allergy class
Anatomy (Patient)	<Anatomy>cervical spine</Anatomy>	patients admitted for surgery of the cervical spine for fusion or discectomy
MedBeforeAdm (Problem / Intervention)	<MedBeforeAdm><Drug> </Drug> <Prblm> osteoporosis OR osteopenia</Prblm> </MedBeforeAdm>	women with hip or vertebral fracture despite being on medication for osteoporosis or osteopenia
MedOnDisch (Problem / Intervention)	<MedOnDisch><Drug> </Drug> <Prblm> COPD </Prblm> </MedOnDisch>	patients with COPD who were not discharged on inhaled steroids
MedForProblem (Problem / Intervention)	<MedForPrblm><Drug> ritalin</Drug><Prblm> depression</Prblm> </MedForPrblm>	patients on Ritalin for depression
ProcBeforeAdm (Intervention)	<ProcBeforeAdm>dialysis</ProcBeforeAdm>	patients admitted for complications due to renal failure despite being on dialysis
ProcForProblem (Problem / Intervention)	<ProcForPrblm> <Proc>ablation</Proc> <Prblm>atrial fibrillation</Prblm> </ProcForPrblm>	patients with atrial fibrillation treated with ablation
Procedure (Intervention)	<Procedure>surgery<MOD> robotic-assisted </MOD></Procedure>	patients who had robotic-assisted surgery
FamilyHx (Problem / Intervention)	<FamilyHx><Prblm>lung cancer</Prblm></FamilyHx>	patients with a history of smoking as well as personal and family history of lung cancer
DischDest (Outcome)	<DischDest>skilled nursing facility</DischDest>	patients with dementia who were discharged to a skilled nursing facility or other institutional setting
Encounter Location ()	<Location>ER</Location>	patients seen in the ER for low back pain who were not admitted to the hospital

patient encounters and on interesting topics in recent issues of the General Medicine Journal Watch ([11]), and the other 30 on the IOM priority topics (the Medical Records Retrieval track organizers told the track participants in advance that the test topics would be based on the IOM priority topics and did not restrict access to those topics during the development period). Together we expanded the four basic PICO slots into more than twenty refined slots; for example, the original *Patient* slot was split into *Age*, *Gender*, *Population*, and *Anatomy*. Both authors then manually encoded 30 training questions each using the refined frame slots, and subsequently they reviewed all 60 training questions and finalized the refined frames together. We developed frames capable of capturing nuances of the question, such as temporal relations and specific groups of patients. For example, we defined three medication slots (*medications before admission*, *on discharge*, and the fallback, *medication for problem*). These distinctions are needed to encode (and answer) temporal questions, such as *Find patients with HIV admitted for a secondary infection who were not on prophylaxis for opportunistic infection*. Table 1 presents the final set of the frame slots. Note that we chose the surface representation of our frame slots in XML format for the convenience of then automatically translating the frames to the query syntax of the search engines, Lucene² and Essie ([8]), used for retrieval. Both search engines rank results according to the likelihood of their relevance and provide complex syntax that allows the user to structure queries beyond simple keyword searches.

2.1 Frame Extraction System

Our automatic system extracts the frames in four steps. In the first step, the system submits the question to MetaMap ([1]) with the default settings to extract the Unified Medical Language System[®] (UMLS[®]) ([12]) concepts. For each concept, the system stores the lexical match with offset and length, negation status and semantic types in a lookup table.

In the second step, the system uses regular expressions to extract patient demographics and social history. The patterns for age include a small vocabulary of age-related terms (for example, preemie, toddler, tween) and a library of regular expressions for identifying specific ages and age ranges (for example, $(\d+)\W*\text{years}\W*\text{of}\W*\text{age}$). Our entire gender look-up list is very small (female, girl, gravida, her, lady, ladies, she, woman, women, boy, he, his, male, man, men, gentleman, gentlemen). The **Population** slot is currently limited to occupations and ethnicities defined by the UMLS semantic types **Professional or Occupational Group** and **Population Group**, respectively. The patterns for social history are currently limited to identifying smoker status, alcohol consumption and illicit drug use. We used lexico-semantic patterns to extract the **Complications_of** relation. Our patterns combine semantic categories and lexemes, for example, we created the $[concept_Problem]s/p[concept_any|word_noun]$ pattern based on the training question *patients admitted for injuries s/p fall*.

² <http://lucene.apache.org/core/>

Table 2. Rules and examples for the syntactic-semantic question frames

Refined Frame slot	Rules and constraints	Example
UMLS concept augmented with modifiers	If dependency ∈ modifier & Q: ... MRSA endocarditis... governor ST ∈ (Problem Intervention) ⇒ add modifiers to the term	Q: ... MRSA endocarditis... endocarditis amod MRSA & endocarditis[dsyn] ⇒ <Prblm> endocarditis<MOD>MRSA </MOD></Prblm>
Conjunction	If dependency ∈ (AND OR) & governor dependent ∈ (Problem Intervention) ⇒ terms	Q: ... staging or monitoring of cancer... staging conj_or monitoring & monitoring[hlca] ⇒ <Proc>staging OR monitoring</Proc>
Admit_problem	If dependency ∈ (prep_with prep_for) & governor admit & dependent Problem ⇒ admit_problem	Q: ... admitted with an asthma exacerbation prep_with asthma exacerbation & asthma exacerbation[findg] ⇒ <AdmitProblem>asthma exacerbation </AdmitProblem>
Med_for_problem Proc_for_problem	If dependency path contains a treatment indicator, Intervention and Problem ⇒ if Intervention ∈ Drug ⇒ Med_for_problem else ⇒ Proc_for_problem	Q: ... monoclonal antibody treatment for inflammatory bowel disease amod monoclonal antibody treatment prep_for inflammatory bowel disease & inflammatory bowel disease[dsyn] monoclonal antibody[aapp,imft] ⇒ <MedForPrblm><Drug>monoclonal antibody</Drug><Prblm> inflammatory bowel disease</Prblm></MedForPrblm>

In the third step, the system processes the question sentences using the Stanford dependency parser ([14]). To prevent the parser from breaking-up multi-word concepts, the system first concatenates all of the words in a concept. We focused on extracting a limited set of typed dependency relations, conjunctions, and modifiers. The system only populates the frame slot if the semantic and syntactic constraints are satisfied. Table 2 shows the examples of the rules. If a rule is applied, the concepts used in that rule are marked as “used” in the look-up table.

After completing iterations over the dependency paths, in the final (fourth) step, all of the remaining unused concepts that could populate the four basic PICO slots are added to the frame. That is, if the lookup table for the question concepts contains concepts in the semantic groups **Disorders (Problems)**, **Interventions** or **Anatomy** that are not already marked as used, the concepts populate the traditional PICO frame slots. Although every question passes through the four modules that implement the four steps, the question frame may pass through a module without changes if none of the rules or patterns applies. Table 3 illustrates each step of the automatic frame extraction for one of the test questions.

2.2 Evaluation

We conducted both an intrinsic and an extrinsic evaluation of the test frames automatically generated from the TREC Medical Record Retrieval track test topics. Since DDF implemented the frame extraction system, only SA conducted

Table 3. Four question-processing steps

Step	Process	Result
0	Original TREC test question	Patients with complicated GERD who receive endoscopy
1	UMLS concept extraction via MetaMap 2010	Disease or Syndrome: GERD Diagnostic Procedure: endoscopy <i>Explanation:</i> "GERD" and "endoscopy" are identified as relevant concepts and assigned the semantic types "Disease or Syndrome" and "Diagnostic Procedure," respectively
2	Using regular expressions to extract demographics and social history	n/a for this question
3	Extracting dependencies using Stanford Dependency Parser	Modifier: complicated the Proc_for_problem: endoscopy GERD <i>Explanation:</i> the word "complicated" is identified as a modifier of "GERD" and the relationship between "GERD" and "endoscopy" is identified as procedure for problem
4	Assign concepts that have not yet been used to the four basic PICO slots	n/a for this question
Final frame result: <ProcForPrblm><Proc>endoscopy</Proc><Prob>GERD<MOD>complicated</MOD></Prob></ProcForPrblm>		

the intrinsic evaluation to judge the accuracy of the automatic frame extraction. She judged a frame to be correct if the system extracted all of the test question elements and each one was placed into its correct slot.

Even the perfectly generated frames are not guaranteed to find eligible patients. There are three possible reasons for failure: 1) there are no eligible patients in the dataset; 2) the algorithm for converting question frames to complex search engine queries is incorrect; and 3) the query needs to be more complex than correctly combining the key terms provided in the description of the inclusion criteria in complex searches. In our extrinsic evaluation, we focused on the third reason: the complexity of the query and the modifications to the terms, template and complex searches needed for near perfect precision at top ten potential cohort participants.

For the extrinsic evaluation, we used the Essie corpus analysis and mining tool ([8]) and the Medical Record Retrieval track document collection. Essie is a domain specific search engine with built-in UMLS-based synonymy expansion that developers at the National Library of Medicine created to support NLM's ClinicalTrials.gov. Essie has been used for that purpose since 2001. The TREC document collection contained over 100,000 reports from over 17,000 patient

visits. Each visit was associated with one or more reports; for example, if the visit was to the **Emergency Department (ED)**, only the ED note was associated with that visit, but if the visit was a multiple-week hospital stay, dozens of documents of various types (e.g., progress notes, radiology reports, discharge summary) were associated with that single visit. For every test question, each author independently used the Essie interface to manually generate a query, reviewed the top ten patients' visits that Essie returned, and refined the query until she had created the ideal question to return the most relevant visits. Both authors then compared the manual queries developed using Essie to the automatic frames to evaluate the differences between the manual and automatic queries. We also compared the visits returned by each method to see how the query differences impacted the relevance of the visits that were retrieved. Here, we only evaluate how many automatic frames needed modifications and the nature of the modifications.

Finally, we compared the average performance of the baseline queries to that of the frame-based queries. For the baseline queries, the original free-text descriptions of the inclusion criteria were submitted to a search engine without any modifications. Clinical documents are not likely to contain exact matches of these long descriptions, therefore, we allow the search engine to arbitrarily break the baseline queries into phrases and remove low-frequency query terms (“lossy expansion”).

3 Results

In the intrinsic evaluation, SA evaluated the accuracy of the automatically generated frames created from the test questions provided by the TREC Medical Records track organizers. Out of the 35 frames 34 were accurate and one was incorrect. The reason for the error in the frame (shown in Table 4) is the lack of the appropriate cue in our pattern set for the *Complications_of* slot. Once the *secondary to* cue was added to the patterns (in a system that was not used in the TREC evaluation), the correct frame was extracted.

In the extrinsic evaluation, we evaluated the usefulness of the automatic frames for cohort identification. We could evaluate only 34 frames because one

Table 4. Automatic and manually corrected question frames for the question “Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes”

Automatically generated frame	Correct frame
<Age>adult</Age>	<Age>adult</Age>
<Prblm>insulin dependent diabetes	<ComplicationsOf> <Prblm> anion
<MOD>secondary </MOD></Prblm>	gap acidosis </Prblm> <Cause> in-
<Prblm>anion gap acidosis</Prblm>	sulin dependent diabetes </Cause>
<Location>emergency room</Location>	</ComplicationsOf>
	<Location>emergency room</Location>

Table 5. Modifications needed for near perfect precision. Questions for which the original frames retrieved few relevant documents in the top ten are marked with an asterisk.

Test question	Automatic frame	Modifications [Modification type]
1 Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis	<Prblm>endocarditis <MOD>MRSA </MOD></Prblm>	Prblm: (SBE OR endocarditis) AND (staph OR MRSA) [Domain knowledge of the disease]
2* Patients with ductal carcinoma in situ (DCIS)	<Prblm>ductal carcinoma</Prblm> <Prblm>DCIS </Prblm>	Prblm: Ductal carcinoma in situ OR (breast cancer AND in situ) [Domain knowledge of the disease]
3* Patients treated for vascular claudication surgically	<ProcForPrblm> <Proc>surgically </Proc> <Prblm>claudication <MOD> vascular</MOD> </Prblm> </ProcForPrblm>	Prblm: vascular claudication OR (“peripheral vascular disease” AND calf) Proc: endarterectomy OR popliteal [Domain knowledge of the disease and the procedures]
4* Patients with chronic back pain who receive an intraspinal pain-medicine pump	<ProcForPrblm> <Proc>pump<MOD> intraspinal</MOD> <MOD>pain- medicine</MOD> </Proc> <Prblm> chronic back pain </Prblm> </ProcForPrblm>	Drug: (Intrathecal OR subarachnoid OR intraspinal OR epidural) AND (“morphine pump” OR “intrathecal pump” OR “pain pump” OR “dilauid pump” OR “opioid pump” OR “epidural pain”) [Domain knowledge of pain medications and administration routes]
5 Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes	See Table 4	Prblm: anion gap acidosis OR DKA OR ketoacidosis OR metabolic acidosis [Domain knowledge of the disease]
6* Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op	<AdmitProblem> hip OR knee </AdmitProblem> <Drug>anti-coagulant </Drug>	Drug: heparin OR warfarin OR Clopidogrel OR Ticlopidine OR Enoxaparin OR anticoagulants [Domain knowledge of medications]
7* Patients who underwent minimally invasive abdominal surgery	<Procedure> abdominal surgery<MOD>invasive </MOD> <MOD>minimally </MOD> </Procedure>	Proc: (Laparoscopic OR minimally invasive OR MIS) NEAR (abdominal OR bariatric OR gastrojejunostomy OR appendectomy OR colectomy OR sigmoidectomy OR cholecystectomy) [Domain knowledge of procedures]
8* Patients admitted for care who take herbal products for osteoarthritis	<MedForPrblm> <Drug>herbal products</Drug><Prblm> osteoarthritis </Prblm> </MedForPrblm>	Drug: Capsaicin OR capzasin OR arthritis formula OR soy OR Boswellia OR licorice OR cohosh OR hawthorn OR castor OR “cranberry capsule” OR “cranberry tablet” OR echinacea OR Glucosamine OR ubiquinone OR thistle OR Gingko-Biloba OR primrose OR aloe OR cinnamon OR flaxseed [Domain knowledge of medications]
9 Patients admitted with chronic seizure disorder to control seizure activity	<Problem> seizure disorder <MOD>chronic </MOD> </Problem>	Prblm: seizure disorder OR “status epilepticus” [Domain knowledge of the disease]

of the original cohort descriptions had no relevant documents in the collection. Based on manual review of how relevant the top ten visits retrieved by the automatic frames were to the cohort criteria, we found that 25 test question frames did not need any modifications, and only nine test question frames did. Of those nine, six would have failed to find most relevant documents without modifications. For the remaining three questions, modifications targeted recall

and improving precision. In all cases, modifications required domain knowledge beyond the UMLS synonymy. For the questions that would have failed, the drug classes or high level descriptions of the procedures needed to be expanded with specific instances. See for example, the expansion for *herbal products* (question 8 in Table 5). In the comparison of the overall average performance, the frame-based queries did not provide the anticipated advantage compared to the baseline queries.

4 Discussion

Formally representing the essence of the cohort characteristics for comparative effectiveness studies can potentially streamline the cohort identification process. Previous analysis of the basic PICO frames, which were designed to formally represent clinical questions, showed that the framework is best suited for representing therapy questions and considerably less suitable for diagnosis, etiology, and prognosis questions. The basic framework cannot capture the fine-grained relationships between frame elements, or model temporal/state information and anatomical relations ([7]), which are exactly the elements needed to accurately capture the study cohort characteristics. We hypothesized that expanded syntactic-semantic PICO frames would compensate for the shortcomings of the basic PICO frames, potentially at the cost of being more brittle.

One benefit of our method is that the syntactic-semantic query frame is generated completely independently of the patient records to be queried, so the same frame can be used to search multiple disparate databases, regardless of the database structure. Different database sources might process patient notes in different ways, either by applying a frame structure similar to the query frames, encoding the data using NLP as described by Friedman et al ([6]), dividing the note into sections as we did for TREC 2011, or taking the text of the note without any modification. Given that the query frame is independent of the clinical note structure, the same query frame can be matched with different local note structures, which is key for data integration from multiple sites. The common query frame method can be used to query data from different providers and hospital systems not only for cohort identification, but also for assessing quality metrics and for health information exchange.

Our first concern in developing the frames was determining the minimal set of slots capable of capturing all necessary fine-grained details. For example, we define medications on discharge to capture medications administered only during the hospital encounter and distinguish those from medications on admission and take-home medications, but we have only one slot for procedures (to capture procedures performed during the hospital encounter). We assume that procedures performed before the current encounter are associated with past medical history, and all procedures not associated with the past medical history occurred during the encounter. Only 10 of our 21 slots were needed to encode the TREC test questions: *Age*, *Gender*, *Anatomy*, *Complications_of*, *AdmitProblem*, *Finding*, *Problem*, *Procedure*, *ProcForPrblm*, and *MedForPrblm*. We thoroughly verified

that the basic frame slots (used more extensively in the test questions than in the training questions) were appropriate for capturing the question information. Indeed, the test questions seemed less complex than our training set, and if they were used for actual cohort selection, it is likely that more patients would have to initially be screened and then excluded in subsequent selection steps. For example, eight questions had the disorder as the only selection criterion. This seeming simplicity of the test questions is, however, understandable. Traditionally, the first year of a TREC track (as this was) gives the participants an opportunity to focus on the document set and accomplishing the test task within a short timeframe. The complexity of the task increases in the subsequent evaluations. Future work on more complex questions with a larger set of criteria for retrospective study cohorts will determine if our current set of 21 slots is capable of capturing all of the necessary details and if the syntactic-semantic frames will provide the hypothesized advantages over the baseline queries.

Our second concern was the potential brittleness of the approach that relies heavily on the typed dependency parse tree. In this evaluation, all syntactic-semantic extraction rules that fired during the extraction of the test question frames were triggered correctly and populated the correct frame slots.

The limitation of our study is that the relatively small number of the syntactic-semantic extraction rules was developed using a relatively small set of training questions. We need to test if the number of slots and rules can be kept at a manageable size with a larger set of questions. This opportunity will present itself in the TREC 2012 Medical Records evaluations. Note, that in the subsequent evaluations we do not have to rely solely on translating the frames to the search engine query languages. Instead, we could apply the proposed method (alone or after an initial search step) to the patients' notes. Then, frame unification or a constrained frame matching approach could be applied to both the query and the patients' cases frames. In fact, we tested the constrained frame matching approach in answering clinical questions with extracts from MEDLINE[®] citations and found it to be more accurate than the information retrieval approaches alone ([4]). We anticipate similar results could be achieved for the cohort identification task. Generation of the patients' cases frames, however, will require additional research.

5 Conclusion

The secondary use of clinical data for cost- and comparative- effectiveness studies is a burgeoning area of clinical research. An automatic system capable of identifying patients based on a textual description of the inclusion and exclusion criteria will potentially speed up the process of cohort identification, as well as enable exchange of health information and evaluation of clinical quality metrics. Our evaluation of capturing the inclusion/exclusion criteria of a comparative-effectiveness study cohort expressed as a natural language question shows that syntactic-semantic frames can accurately capture the desirable patients' characteristics. Further evaluation of the frames' retrieval effectiveness showed that a

third of the frames needed modifications because of the mismatch between the high-level descriptions of the criteria in the question and the more specific terms that define the diagnoses, procedures, and medications in the clinical records. Future work will involve further evaluation of the overall average retrieval effectiveness of the automatic syntactic-semantic frames and refining the frame extraction system.

Acknowledgments. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.* 17(3), 229–236 (2010)
2. Boxwala, A., Kim, H., Choi, J., Ohno-Machado, L.: Understanding data and query requirements for cohort identification in clinical research studies. In: *AMIA Annu. Symp. Proc.*, p. 95 (2011)
3. Cimino, J.J., Ayres, E.J.: The clinical research data repository of the US National Institutes of Health. *Stud. Health Technol. Inform.* 160(Pt 2), 1299–1303 (2010)
4. Demner-Fushman, D., Lin, J.: Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics* 33(1), 63–103 (2007)
5. Deshmukh, V.G., Meystre, S.M., Mitchell, J.A.: Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med. Res. Methodol.* 9, 70 (2009)
6. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc.* 11(5), 392–402 (2004)
7. Huang, X., Lin, J., Demner-Fushman, D.: Evaluation of PICO as a knowledge representation for clinical questions. In: *AMIA Annu. Symp. Proc.*, pp. 359–363 (2006)
8. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: a concept-based search engine for structured biomedical text. *J. Am. Med. Inform. Assoc.* 14(3), 253–263 (2007)
9. Institute of Medicine of the National Academies (IOM): 100 initial priority topics for comparative effectiveness research, <http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx>
10. Jacquemart, P., Zweigenbaum, P.: Towards a medical question-answering system: a feasibility study. *Stud. Health Technol. Inform.* 95, 463–468 (2003)
11. JournalWATCH® General Medicine: <http://general-medicine.jwatch.org/> (updated August 16, 2011, accessed August 16, 2011)
12. Lindberg, D.A.B., Humphreys, B.L., McCray, A.T.: The Unified Medical Language System. *Meth. Inform. Med.* 32, 281–291 (1993)
13. Lowe, H.J., Ferris, T.A., Hernandez, P.M., Weber, S.C.: STRIDE—An integrated standards-based translational research informatics platform. In: *AMIA Annu. Symp. Proc.*, pp. 391–395 (2009)
14. de Marneffe, M.-C., MacCartney, B., Manning, C.D.: Generating Typed Dependency Parses from Phrase Structure Parses. In: *LREC 2006* (2006), http://nlp.stanford.edu/pubs/LREC06_dependencies.pdf (accessed August 16, 2011)

15. Murphy, S.N., Barnett, G.O., Chueh, H.C.: Visual query tool for finding patient cohorts from a clinical data warehouse of the Partners HealthCare system. In: Proc. AMIA Symp., p. 1174 (2000)
16. Narayanan, S., Harabagiu, S.: Question Answering based on Semantic Structures. In: International Conference on Computational Linguistics COLING 2004, Geneva, Switzerland (2004)
17. Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S.: The well-built clinical question: a key to evidence-based decisions. *ACP J. Club* 123, A12–3 (1995)
18. Ruiz, E.E., Chilov, M., Johnson, S.B., Mendonça E.A.: Developing multilevel search filters for clinical questions represented as conceptual graphs. In: AMIA Annu. Symp. Proc., p. 1118 (2008)
19. Tu, S., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D., Sim, I.: A practical method for transforming free-text eligibility criteria into computable criteria. *J. Biomed. Inform.* 44(2), 239–250 (2011)
20. Voorhees, E., Tong, R.: Overview of the TREC 2011 Medical Records Track. In: The Twentieth Text REtrieval Conference Proceedings TREC 2011, Gaithersburg, MD. National Institute for Standards and Technology (2011)