Interactive Cohort Identification of Sleep Disorder Patients Using Natural Language Processing and i2b2

W. Chen[1]; R. Kowatch[2]; S. Lin[1]; M. Splaingard[3]; Y. Huang[1]

[1]Research Information Solutions and Innovations; [2]Center for Innovation in Pediatric Practice; [3]Sleep Disorder Center, Nationwide Children's Hospital, Columbus, OH

Keywords

Sleep disorder, cohort identification, natural language processing (NLP), i2b2, clinical ontology

Summary

Nationwide Children's Hospital established an i2b2 (Informatics for Integrating Biology & the Bedside) application for sleep disorder cohort identification. Discrete data were gleaned from semi-structured sleep study reports. The system showed to work more efficiently than the traditional manual chart review method, and it also enabled searching capabilities that were previously not possible.

Objective: We report on the development and implementation of the sleep disorder i2b2 cohort identification system using natural language processing of semi-structured documents.

Methods: We developed a natural language processing approach to automatically parse concepts and their values from semi-structured sleep study documents. Two parsers were developed: a regular expression parser for extracting numeric concepts and a NLP based tree parser for extracting textual concepts. Concepts were further organized into i2b2 ontologies based on document structures and in-domain knowledge.

Results: 26,550 concepts were extracted with 99% being textual concepts. 1.01 million facts were extracted from sleep study documents such as demographic information, sleep study lab results, medications, procedures, diagnoses, among others. The average accuracy of terminology parsing was over 83% when comparing against those by experts. The system is capable of capturing both standard and non-standard terminologies. The time for cohort identification has been reduced significantly from a few weeks to a few seconds.

Conclusion: Natural language processing was shown to be powerful for quickly converting large amount of semi-structured or unstructured clinical data into discrete concepts, which in combination of intuitive domain specific ontologies, allows fast and effective interactive cohort identification through the i2b2 platform for research and clinical use.

Correspondence to:

Yungui Huang PhD MBA
Research Information Solutions and Innovations
Nationwide Children's Hospital
575 Children's Crossroad
Columbus, OH 43215 United States
E-mail: Yungui.Huang@nationwidechildrens.org

Appl Clin Inform 2015; 6: 345–363
http://dx.doi.org/10.4338/ACI-2014-11-RA-0106
received: November 25, 2014
accepted: February 23, 2015
published: May 27, 2015
Citation: Chen W, Kowatch R, Lin S, Splaingard M, Huang Y. Interactive cohort identification of sleep disorder patients using natural language processing and i2b2. Appl Clin Inf 2015; 6: 345–363
http://dx.doi.org/10.4338/ACI-2014-11-RA-0106

Introduction

Cohort identification is commonly used to find patients with shared characteristics. It is important for early identification of disease risks [1–4] and patient recruitment for clinical trials [5–7]. Cohort identification usually requires searching a large clinical database for a small subset of subjects; therefore, it is often time consuming. The process could be more expensive when manual chart reviews are needed to confirm diagnosis or other clinical features using unstructured data such as natural language-based clinical notes [4, 8–10]. The labor intensive nature of cohort identification creates major barriers for time-critical decision making in clinical practice [11].

Natural language processing (NLP) techniques have been widely used to quickly analyze a large amount of texts in support of humans [12]. Techniques such as part-of-speech tagging, parsing, and named entity recognition (NER) may speed up the process of identifying diagnoses, procedures and medications in clinical notes with satisfactory accuracy [13–15]. Syntactic structures obtained from parse trees provide useful indicators for extracting concepts and values from natural language sentences as well as for building queries [16]. Knowledge frameworks such as Unstructured Information Management Architecture (UIMA), clinical Text Analysis and Knowledge Extraction System (cTAKES) and MetaMap Transfer (MMTx) have been widely used to provide terminology services with a natural language processing system [17–21]. Unified Medical Language System (UMLS) often serves as the backend knowledge base for clinical text processing purposes [22]. Machine learning-based approaches have also been integrated into NLP systems for clinical NER tasks [23–25]. However, their overall efficacy may be compromised sometimes by the lack of in-domain knowledge possessed by human experts [26, 27].

Sleep study summary reports (polysomnograms) are semi-structured documents that require either manual or automated natural language analysis. These documents contain information extracted from different sources such as clinical charts, GRASS® sleep study programs, and human edits [4]. Previous methods of finding cohorts for sleep study included searching sleep study documents using keyword-based file search functions on a Windows system. Human experts later manually verified these documents and pulled useful data from validated corresponding documents. This manual error-prone process may take up to 60–70 hours to finish. It is highly necessary to automate the process so that cohorts can be identified more effectively for various sleep research purposes.

I2b2 is an interactive medical informatics system that is widely used for patient cohort identification [28, 29]. It features a user-friendly web-based interface, an ontology browser and a search tool [30]. Although i2b2 has been used in various clinical applications, it is still new to sleep study, especially with unstructured sleep report data [31, 32]. I2b2 comes with its own natural language processing component as optional but this component is limited to specific purposes and is not flexible enough to produce discrete data [33].

This study aimed to leverage i2b2's rich self-service query builder functionalities. It attempted to overcome i2b2's NLP limitations by building a customized NLP workflow to populate i2b2 tables. It also constructed in-domain ontologies to speed up cohort identification by defining queries through dragging and dropping concepts into a web-based query interface. Natural language processing techniques were employed to extract both standard and non-standard clinical terms from sleep study documents collected at Nationwide Children's Hospital for the past 11 years. The accuracy of natural language processing results was evaluated using standard precision, recall and F1 measures against results provided by human experts. Screenshots were included to illustrate the use cases of the system. The contribution of this study included a working demo of the i2b2 self-service cohort identification system for sleep study and easily adaptable source code for parsing text data for new applications.

Data

15,683 sleep study reports were collected from January 2004 to September 2014. ▶Figure 1 represents the distribution of sleep study reports over the years. Considering the fast growing number of documents, there is urgent need from our physicians to automate the cohort identification process by using informatics system to overcome the poor performance of traditional manual means.

Sleep reports were generated through a pipelined process (▶Figure 2). Basic patient information was collected from smart text in Epic™ (Epic Systems, Verona, WI), a commercial medical informatics system. These smart texts are annotations that wrap a collection of information into one short phrase. For example, typing the smart phrase *#vitals#* will pull all vitals information of a patient. Sleep study lab results were obtained from the GRASS™ system (Grass Instruments, Quincy, MA) that was used to assess the quality of sleep using sensors and monitors. These results mainly contained data such as sleep architecture, respiratory and EEG in both tabulated and free text formats. In the sleep domain, sleep architecture, for example, represents the cyclical pattern of sleep between different sleep stages, including non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. Finally, both Epic text and sleep data were transferred into a Word template file for reporting.

All documents had sections listed in ▶Table 1. Each section contained concepts that could be either text or numeric. Textual concepts were concepts of diagnosis, procedure and medication while numeric concepts are those with numeric values. NLP parser and regular expression parsers were developed respectively to extract textual concepts and numeric concepts.

Methods

Document processing flow

There were several steps in processing documents (▶Figure 3). First, all Word documents were converted into text format. Second, documents were split into sections as listed in ▶Table 1. Third, for the *indication* section where previous diagnosis, procedure and medication information would be extracted, we developed a heuristic classifier to classify sentences into four predefined categories: diagnosis, procedure, medication and others. Sentence classification was necessary to make our terminology parser work correctly.

Two parsers were implemented to extract information from different sections: a regular expression parser and a NLP parser. For sections that contained only numeric concepts a regular expression (RE) parser was implemented. In this case, concept names and values were adjacent in the text. For sections that contained only textual concepts, a NLP parser was implemented. Medical terminologies were extracted using a tree parser. In this case, the concept name was the terminology name and the concept value was the same as the concept name. For sections that contained both types of concepts, both parsers were employed.

Once terminologies were extracted, they were matched against Epic terminologies, an ad hoc terminology database we used for classifying terminologies into standard and non-standard. All those with exact matches in EPIC terminology database were tagged as standard and others as non-standard.

Regular expression parser and numeric concept extraction

A regular expression parser was implemented by extracting concepts and values that followed certain patterns in a sentence. For example, in the sentence *REM apnea index is 0* the concept name is *REM apnea index* and the concept value is *0*. The following regular expression pattern was used to capture such a concept value pair:

```
(?i)rem\s+apnea\s+index\s+(is\s+)?[0-9.]+
```

This regular expression would capture any caseless text that begins with *rem* followed by *apnea* followed by *index* followed by an optional *is* and a number. For each concept, we constructed a regular expression parser using the similar pattern-based approach. All numeric concept names were predefined, as they were the same as the numeric concepts from the GRASS sleep study system. In our sleep reports, numeric concepts mainly came from the Lab Results section. As lab results were semi-structured text, we achieved over 98% accuracy of extracting numeric concepts based on a manual evalution of 100 documents.

Sentence classification

The *indication* section is the only section that contains textual concepts and requires NLP parsing analysis. Sentences in the indication section were classified into one of the four categories: diagnosis, procedure, medication and others. Below is an example of the sentences in the indication section of a sleep study reports.

"This is a 09 yrs 10 mos old white male with a history of sleep onset insomnia, nocturnal awakenings and painful legs at rest for the past two years. (Diagnosis) Previous surgeries include tracheostomy and UPPP. (Surgeries) Current medications include Protonix, Zyrtec, Bactrim and Flovent. (Medications). Wake up time: 09:30am. Bed time: 08:00 pm. BMI: 35. (Others)"

Sentences in the indication sections were classified using heuristic rules by detecting lexical features such as *with a medical history of*, *surgeries include* and *medication include*, among others. Given the limited language patterns of constructing sentences of each category, the overall classification accuracy was above 97% based on 100 manually verified samples.

Using the method above, we observed 99.9% of documents containing diagnosis sentences, 80.2% containing procedure sentences, 92.6% containing medication sentences. We have not tested against machine learning-based classifiers, as our heuristic based classifier was much easier to implement and worked well enough for our purposes.

NLP parser and textual concept extraction

To extract medical terminologies from classified sentences, we parsed the sentence using NLP parsing techniques. We utilized the Stanford Parser Java library to generate the parse tree of a sentence [34]. On the top level of the tree was the original sentence and on the leaf level were the words in the sentence. Nodes in between were phrases such as noun phrases, verb phrases or prepositional phrases. The recursive neural network (RNN) dependency parser was chosen to parse the sentences [35, 36]. This RNN parser is a new parser recently developed by researchers from Stanford University. It is shown to be scalable to variable-sized inputs and aware of sentence context for parsing. Moreover, it leverages information about the semantic structures of a sentence to deal with unseen phrases during parsing [36]. Given that medical reports include a significant amount of phrases that are not common in traditional training corpus such as Wall Street Journals, we found this parser to be a good candidate to suit our needs. The RNN parser turned out to perform better at least in our case than the other two parsers available in the Stanford NLP libraries, the PCFG parser [38] and the factored parser [39]. Parsing a sample of 100 sentences showed that the RNN parser achieved over 90% accuracy, and was about 14.3% more accurate than the PCFG parser and 12.5% more accurate than the factored parser.

Compared with the traditional named entity recognizer, the NLP parser does not have to rely on knowledge base to extract terminologies. It works by examining the syntactic structures of a parse tree; therefore it works independently from the knowledge base. For example, we assume *sleep apnea* is the exact entry in a knowledge dictionary, the phrase *diagnosed with possible sleep apnea* is different from *diagnosed with sleep apnea*. The traditional named entity recognizer relies on terminology dictionary or lexical patterns and therefore is not designed to capture the modifiers such as *possible* in this case [40]. Such modifiers, however, provide important information for clinical decision-making.

We used the following sentence from our dataset to demonstrate the NLP parsing process: *Patient is a white male with a history of asthma, mild obesity and restless sleep.* Although *obesity* matched exactly to an entry in our knowledge base, *restless sleep* did not have any exact matches. The fact that they are both noun phrases, however, helps us to capture these terminologies regardless of their existence in the knowledge database. Therefore, by extracting all noun phrases in the sentence, we were able to obtain a set of candidate terminologies. The parser tree of this sentence is shown below:

```
(ROOT
  (S
    (NP (NNP Patient))
    (VP (VBZ is)
      (NP
        (NP (DT a) (JJ white) (NN male))
        (PP (IN with)
          (NP
            (NP (DT a) (NN history))
            (PP (IN of)
                (NP (NN asthma))
                (, ,)
              (NP (JJ mild) (NN obesity))
              (CC and)
              (NP (JJ restless) (NN sleep)))))))
    (. .)))
```

▶Table 2 shows the noun phrases extracted from the sentence. As we can see, the candidate list includes some overlapping phrases. Overlapped phrases may convey different meanings. By including them, we gave the parser certain flexibility for interpretations. To filter out irrelevant noun phrases, we defined a list of skipped words. This skipped list included 98% of the most frequent noun phrases that were surely not medical terminologies such as *patient*, *history*, *male*, and so on. Even if certain phrases that were not be completely skipped, we may still be able to filter them when we search and select phrases in i2b2.

Given our Epic knowledge base, *Asthma* had an exact match but the other two did not (▶Table 3). Therefore, they were classified into the two categories for building the ontologies: standard terms and non-standard terms. Both terms will be made available for searching through i2b2.

Plurality, case, abbreviation and negation handling

Case variation and plurality may cause unnecessary differentiation of ontologies. For example, *leg pain* and *Legs pain* should be under the same ontology. Case issue was handled by converting all words to their lowercases and plurality was resolved by converting all words into their lemmas. The lemma of a word is the base form a word. For example, the lemma of a plural noun is its singular form. The lemma of a past tense verb is its present tense. Final terminologies were represented as lowercase lemmas. This way, we ended up with a much smaller list of terminologies for easier searching and browsing.

Some abbreviations were actual medical terminologies such as T&A (tonsillectomy with adenoidectomy) while others were not such as s/p (status post). Previous approaches compared different classic NLP systems (i.e. MetaMap, MedLEE and cTAKES) for medical abbreviation recognition and concluded with suboptimal overall results [41]. In our case, abbreviations were handled automatically by the NLP parser through the tree parsing process. In most cases, these abbreviation will be tagged as nouns or noun phrases and therefore result in candidate terminologies.

Whether a phrase should be considered as candidate terminology depends on whether it is negated in the sentence. For example, in the sentence *patient does not have leg pain*, *leg pain* should not be considered as a candidate terminology. Negation is handled by a negation library called NegEx [42]. This library can detect negation of a phrase by analyzing the relationship between a negation

trigger (e.g. *does not have*) and a phrase (e.g. *leg pain*). Previous experiments showed that this library could reach above 94% of accuracy [42].

Populating i2b2 data repository

i2b2 (Informatics for Integrating Biology & the Bedside) is widely used to access clinical data for knowledge discovery [29]. i2b2 backend data repository is based on a special relational database schema called the star schema. A star schema makes concepts and observations more extensible [29]. To use i2b2 for our sleep disorder project, we mapped all extracted concepts and their values to i2b2's star schema tables.

A central fact table and a few dimension tables represent the schema (▶Figure 4). The central table is the observation fact table that contains all concepts and their values. Five surrounding tables, called dimension tables, provide supplemental information to the concepts in the fact table. The concept table contains all concepts extracted from all documents. The patient table contains all patient information the same as the provider table contains all provider information. The encounter table contains the encounter number field which uniquely identifies each encounter. A visit may have multiple encounters. i2b2 tables relate to each other through foreign keys such as concept code, medical record number, and so on.

All extracted concepts are mapped to the concept table while values of the concepts are mapped to the fact table. In the fact table, there is a type field to indicate whether the concept is a textual or numeric concept. The value of a numeric concept will be number. The value of a textual concept will be the concept itself.

Typically, i2b2 is popular using discrete data collected from a clinical system, clinical encounters or other relational databases. Compared with relational databases, the star schema of i2b2 is famous for its simplified query logic and improved performance against aggregation operations. For further information, one may take a look at the detailed documentations (www.i2b2.org).

Ontology development

The i2b2 cohort identification process relies heavily on the design and development of ontologies. The i2b2 ontology is hierarchically organized collection of concepts. It is represented by a concept path and a concept code [28]. A concept path is a "\" delimited string that separates the concept by different levels of ontologies. From left to right, the ontology becomes more specific. The concept code could be any string that can uniquely identify the concept for example, the ICD9 or ICD10 code.

Concepts in i2b2 also have values that could be either numeric or textual. Consequently, we defined two types of concepts that were extracted from the text: textual concept and numeric concept. Both types of concepts were extracted from sleep documents to build the ontologies for sleep disorder study.

Sleep disorder concepts are grouped into categories such as patient information, indications, sleep architecture, respiratory, ventilation, cardiovascular, EEG and sleep disorder diagnosis. Each category corresponded to the fixed headings of the document and constituted the top-level hierarchy of the ontology path. Under each top ontology were nested ontologies that corresponded to more granulated concepts. The general form for all ontology paths was defined as

- *\[Section Header]\[Concept Category]\[Concept Class]\[Concept Name]*

In the *indication* section, the section header was *Patient Indication and Identification*. But since the indication section mainly includes medical history information of a patient, we updated the section header name to *Medical History* in i2b2 to make it more intuitive to browse.

Ontology building

The knowledge base we used was a terminology database from Epic systems. The knowledge base was used during ontology mapping but not during ontology extraction. The reason we used such an

ad hoc knowledge base rather than ones in the open domain (such as SNOMED or UMLS) was because physicians in our hospital are all familiar with Epic systems. In addition, we found the terms in our sleep documents are better matches with those in the Epic database. Therefore, using the Epic knowledge base we may develop the most useful ontologies for our patients.

The Epic knowledge base also includes generic information such as classes of diagnosis, procedures and medications. Using this information, we further developed the concept path required by i2b2 as follows:

```
\\Medical History\\Diagnosis\standard\diagnosis type\diagnosis name
\\Medical History\\Diagnosis\non-standard\diagnosis name
\\Medical History\\Procedure\standard\procedure type\procedure name
\\Medical History\\Procedure\non-standard\procedure name
\\Medical History\\Medication\standard\medication type\medication name
\\Medical History\\Medication\non-standard\medication name
```

Given the above concept path structure, users are able to query any concept on the path using both concept browser and search functions provided by i2b2. The type information is not part of the parser but part of the ontology. The typing information comes directly from Epic systems. Once the program detects a diagnosis term, all the ontological information related to that term would automatically come from the Epic knowledge base.

Results

Top terminologies

From 15,683 sleep documents we extracted 13,095 patients, 26,550 concepts, and 1.03 million facts. Both textual and numeric concepts were extracted from different sections of sleep documents. From the *indication* section, previous textual concepts of diagnosis, procedures and medications were extracted. ▶Figure 5 and ▶Figure 6 show the top standard and non-standard medications found among all documents. *Albuterol* and *Singulair* are the most common standard medications that can be found in our Epic knowledge base while *Periactin* and *Elavil* are the most common non-standard medications.

In the sleep disorder *diagnoses* section, only the sleep disorder related diagnoses were extracted. Results in ▶Figure 7 showed that *snoring* and *periodic limb movement* (plm) were two leading sleep disorder diagnoses. Numeric results were mainly extracted from the *lab results* sections. For example, *highest plm index* found is 146 and waso (wake after sleep onset) time ranges from 1 to 596 minutes.

Terminology extraction evaluation

Our evaluation process was based on the comparison of NLP parsed results against human expert annotated results. Standard precision, recall and F1 measures were calculated for a random sample of 100 documents. Our evaluation contained two types of accuracy metrics: exact match accuracy and partial match accuracy. This applies to both standard and non-standard terminologies.

Exact and partial matches were defined against human experts rather than entries in the knowledge base. For example, if obstructive *sleep apnea* should be extracted according to human experts, the extracted term *sleep apnea* will only result in a partial match. In most cases, the boundary was clear and easy to identify consecutive words that consisted of a noun phrase.

- **Exact match**: If the term is the same as the human gold.
- **Partial match**: If only a few words from the term match the human gold.

String similarity could be measured using metrics such as edit distance [43] but we found the distance-based similarity measure was not as useful here as in other cases. For example, a larger overlap between a candidate term and human gold does not guarantee a higher rate of correctness. Accord-

ing to our human experts, as long as the partial term retained the same major meaning as the gold, it is a valid partial match. For example, *sleep apnea* and *apnea* are both valid partial matches to *obstructive sleep apnea* but *obstructive* is not a partial match. A partial match was decided by the head noun of the term according to our human experts. Given the availability of our human expert resources, we asked two domain experts to manually annotate concepts in the sentences together at the same time. If they have any disagreement, they will resolve it to their best-agreed result possible. From 100 human annotated sample documents, we obtained 576 diagnosis concepts, 182 procedue concepts and 416 medication concept in total.

Precision, recall and F1 measures were calculated to evaluate the accuracy of exact and partial matches (▶Table 4). Precision was calculated as the percentage of correct terms over all the terms that were actually found. Recall was calculated as the percentage of correct terms over all terms that should be found. F1 was calculated as 2*(precision*recall)/(precision + recall). The results were evaluated using 100 new samples.

We also calculated the accuracy of extracted numeric concepts. As they were mainly machine-generated snippets from the GRASS sleep study system, a high accuracy of 99% was achieved based on 100 samples.

I2B2 sleep disorder portal

The ontology browser in i2b2 could be used to select concepts extracted from documents (▶Figure 8). Each top node was named after the corresponding section heading. Nested nodes corresponded to different levels of ontologies in the concept path. For illustration purpose, a numeric concept was marked with (N) while textual concepts were marked with (T).

The concept search function of i2b2 allowed searching concepts by either names or ICD9 codes. ▶Figure 9 shows the function of searching concepts containing "sleep apnea". Keyword-based concept search is much faster than using the ontology browser.

▶Figure 10 shows the interactive query builder for finding any patients with any type of *sleep apnea* and whose *BMI* is greater than 25 and *PLM index* is greater than 0. This resulted in 21 patients in our dataset. These results were the same as those done by manual methods but it was simply much faster. ▶Figure 11 and ▶Figure 12 represent the views of results summary and previous queries. To protect the privacy of our patients, we did not expand their medical record numbers in the view.

Discussion

Cohort identification is an important task for time-critical decision-making. This paper developed an i2b2 application to speed up the process of cohort identification by converting textual information from sleep study documents into discrete concepts. By using our system, physicians may browse different sleep study ontologies and find patient cohorts using an interactive query builder. The system greatly reduced the labor cost of cohort identification and made data more accessible than using traditional keyword-based methods of document search. . A lot of "what if" questions can be answered directly in real time instead of waiting for weeks for manual chart review.

Two types of parsers were developed to extract data from documents and were demonstrated to be effective. Regular expression parser was developed to parse semi-structured numeric concepts. It achieved very high accuracy given the fact that they were mainly machine generated. By comparison, the NLP parser parsed sentences into tree structures and each node in the tree corresponded to a meaningful text segment. Based on the part of speech tag of the node, one may filter out all noun phrases as candidate terminologies. A major benefit of using the NLP parser was that it allowed capturing of both standard and non-standard terms, which by default are not supported by traditional dictionary-based named entity recognition methods.

The overall accuracy of the terminology parsing was above 80%. For cohort identification purposes, this was within an acceptable range if physicians did not aim to accurately find all eligible patients. False positives can be further reduced by manual review of a much smaller cohort; although, false negatives are harder to improve. The accuracy was subject to several factors. First, sentences

may be misclassified; however, given the semi-structure nature of our document the misclassification cases were rare. Second, terminologies of different classes may be in the same sentence. This may be the major issue that decreased the accuracy. If a medication name was mentioned in a diagnostic sentence, it was not detected given our current implementation. Finally, sentences that were classified in the *others* categories may also include diagnoses, medications or procedures. In this case, we entirely missed the terminologies in those sentences. Despite these limitations, we still found the NLP parser to be useful and it could be further improved by implementing more sophisticated classifiers.

Conflicts Of Interest

The authors declare that they have no conflicts of interest.

Human Subjects Protections

The project is not qualified as Human Subjects research, as defined by the United States Department of Health and Human Services and Food and Drug Administration. Therefore, an application was not submitted to the IRB per our institution's policy on Quality Improvement projects.

Acknowledgement

The authors would like to acknowledge Megan Reynolds, our domain expert, for providing the gold standard, Richard Hoyt for the discussion on i2b2, and Florine Shivers for the editing. We also appreciate the constructive comments from four anonymous reviewers. We also thank the continuous support from Research Information Solutions and Innovations department at Nationwide Children's Hospital.

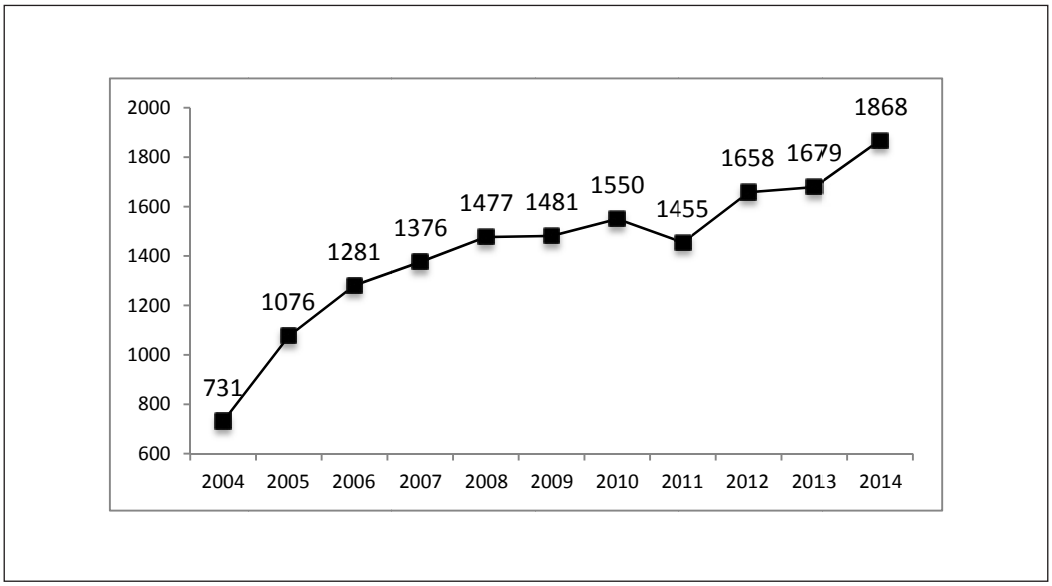

Fig. 1 Number of Reports From 2004 to 2014

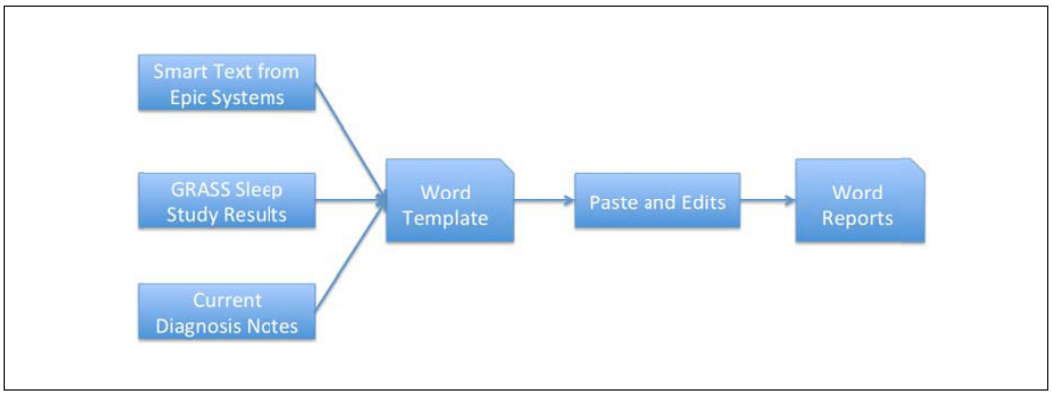

Fig. 2 Sleep Study Reports Generation Pipeline

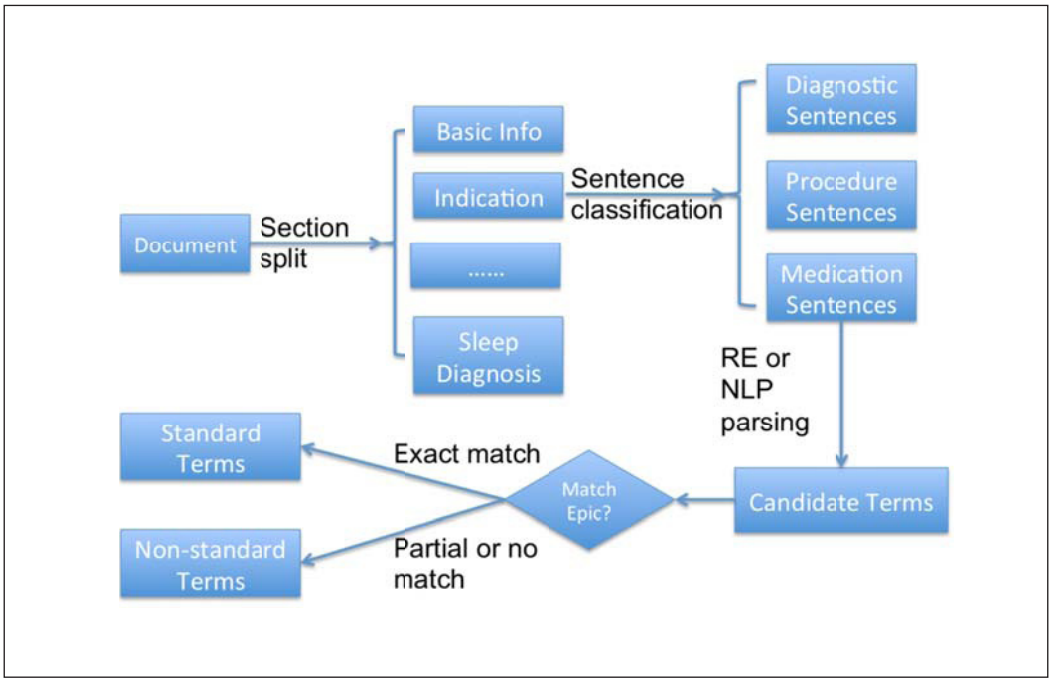

Fig. 3 NLP Parsing Pipeline

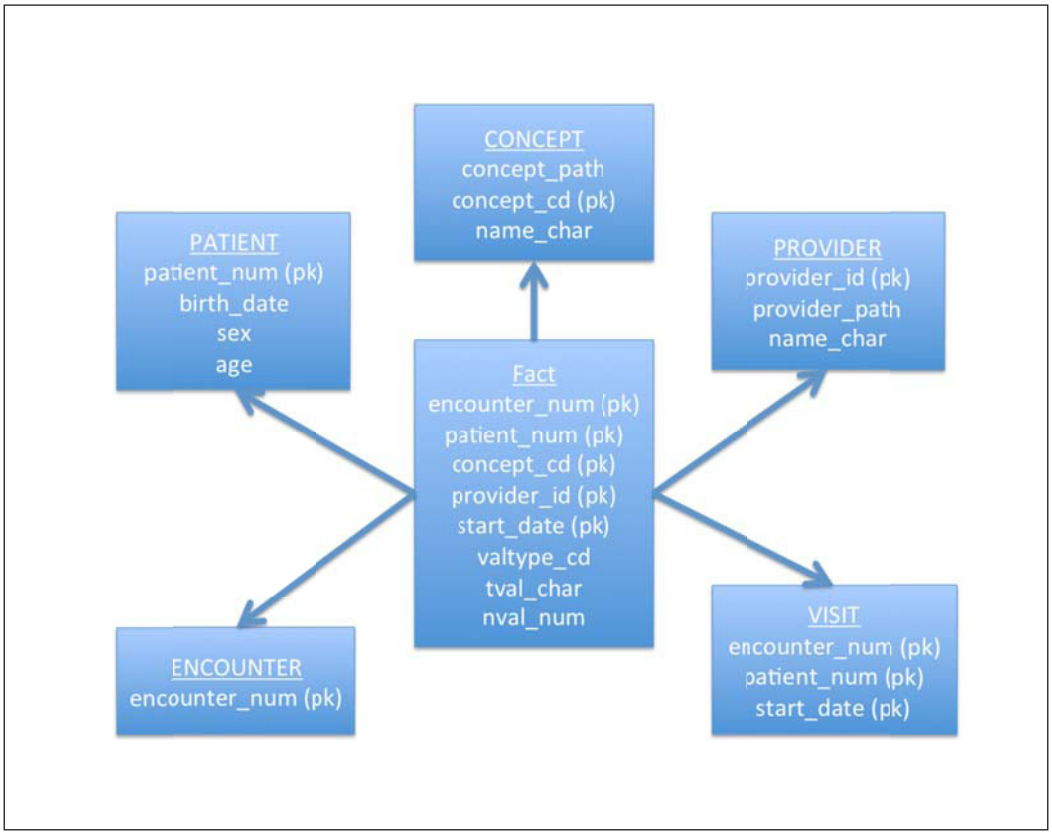

Fig. 4 Star Schema for Sleep Project

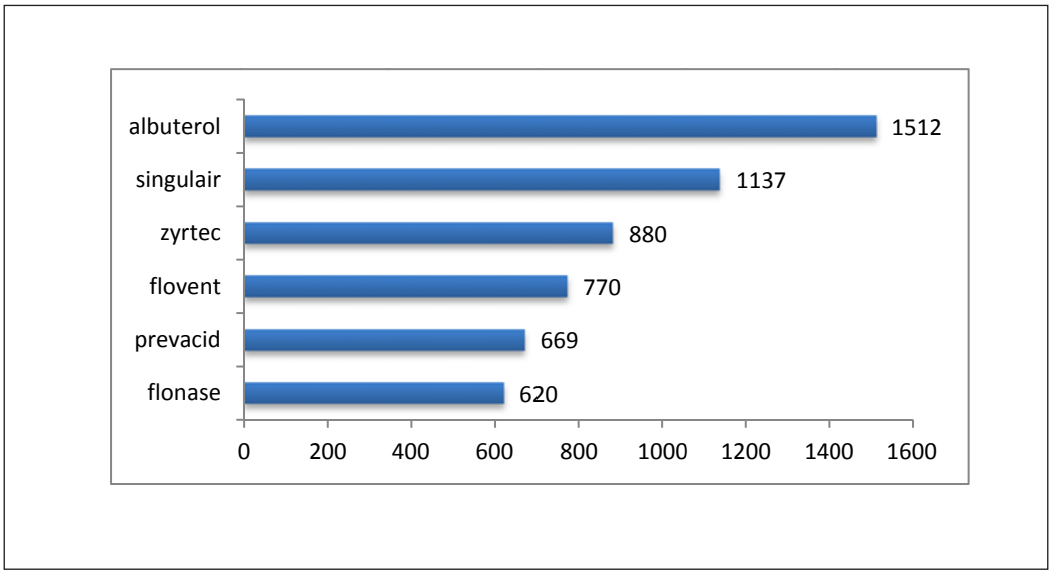

Fig. 5 Top standard medications

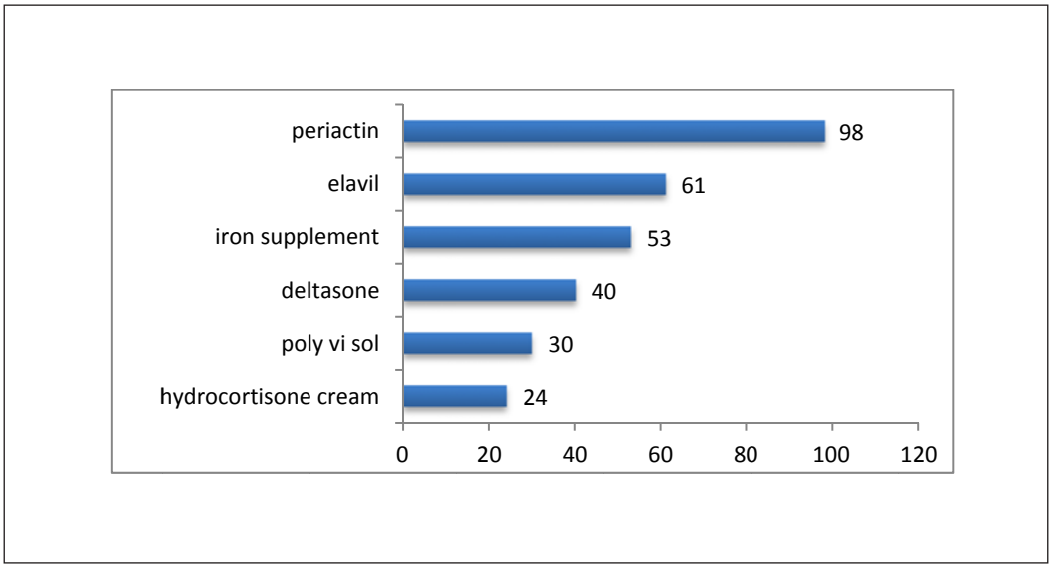

Fig. 6 Top non-standard medications

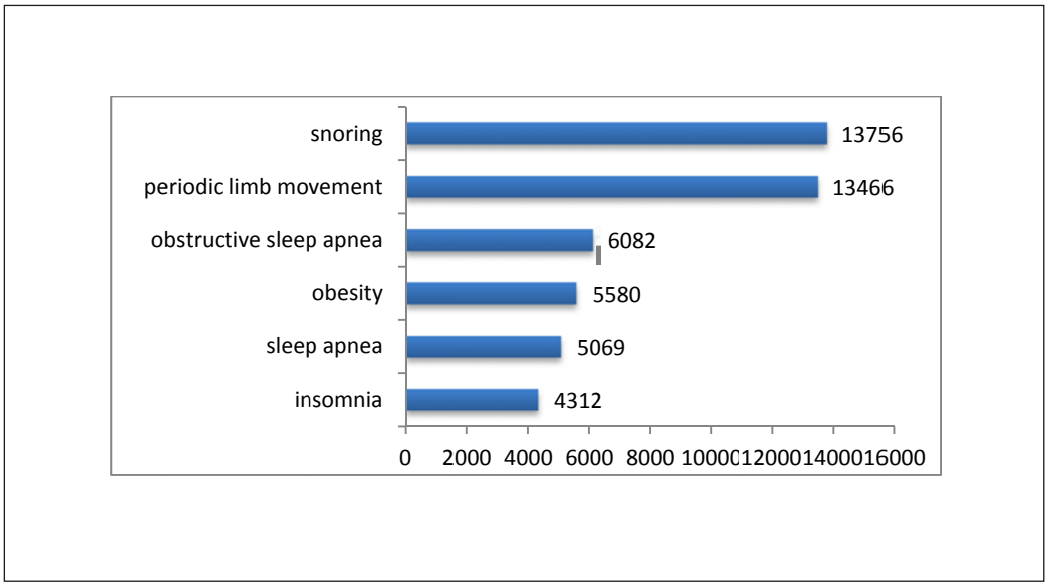

Fig. 7 Top sleep disorder diagnoses

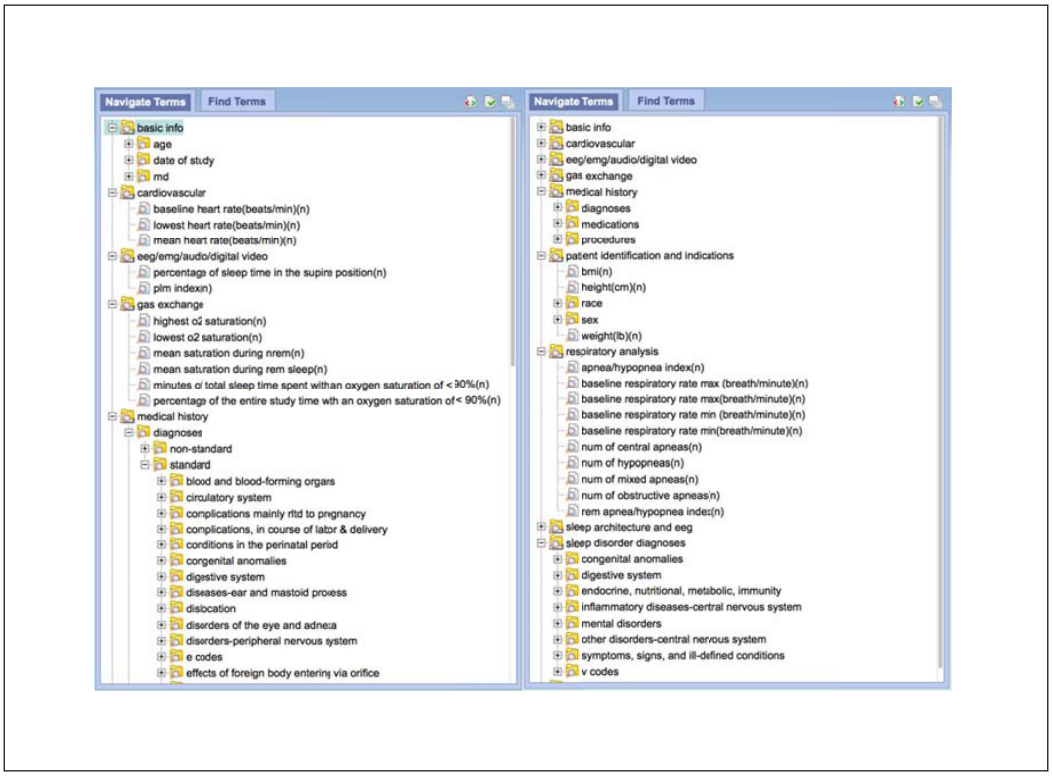

Fig. 8 Ontology browser of sleep i2b2 portal

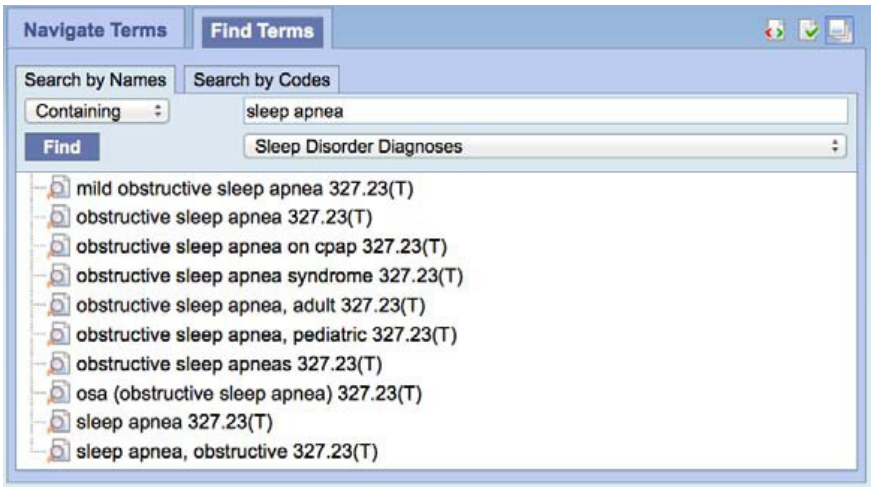

Fig. 9 Search Sleep Apnea Ontologies

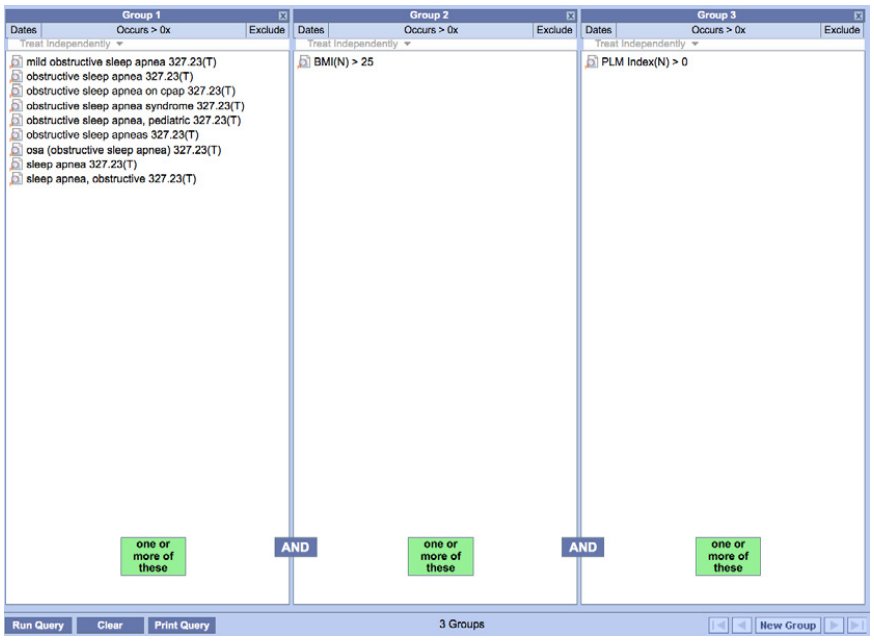

Fig. 10 i2b2 Query Interface

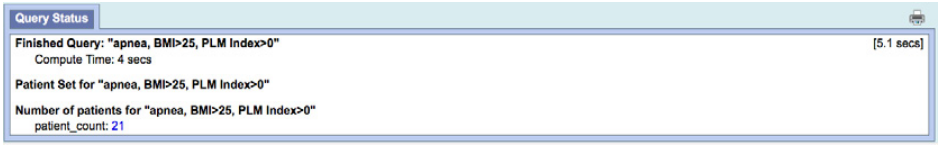

Fig. 11 Result view

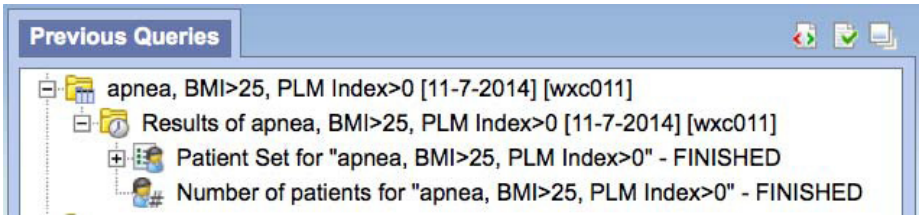

Fig. 12 Previous query view

Table 1 Sleep report sections

Section name	Structure	Concept types	Example concepts
Basic Info	Semi-structured	Text and numeric	patient name, date of birth, medical record number, referring md, date of study.
Indication	Unstructured	Text and numeric	previous diagnoses, medication, procedures, bmi, bedtime, wakeup time.
Architecture	Semi-structured	Numeric	lights out time, lights on time, total record time, number of REM (rapid eye movement) episodes.
Respiratory	Unstructured	Numeric	respiratory rate, number of central, obstructive and mixed apneas.
Gas Exchange	Unstructured	Numeric	O2 saturation.
Ventilation	Unstructured	Numeric	end tidal, ph
Cardiovascular	Unstructured	Numeric	heart rate.
EEG	Unstructured	Numeric	plm (periodic limb movement) index.
Sleep Diagnosis	Unstructured	Text	current sleep disorder diagnoses.

Table 2 Candidate terminology list

Terminology	Skipped or not
patient	yes
a white male with a history of asthma, … sleep	yes
a white male	yes
a history	yes
asthma	no
mild obesity	no
restless sleep	no

Table 3 Final Candidate Terminology List

Terminology	Standard or non-standard
asthma	standard
mild obesity	non-standard
restless sleep	non-standard

Table 4 Terminology extraction accuracy

Terminology	Precision (Exact)	Recall (Exact)	F1 (Exact)	Precision (Exact and Partial)	Recall (Exact and Partial)	F1 (Exact and Partial)
Diagnosis	75.10%	71.18%	73.09%	82.10%	77.33%	79.64%
Procedure	77.72%	73.21%	75.40%	85.72%	78.10%	81.73%
Medication	86.12%	79.91%	82.90%	93.95%	85.34%	89.44%
Overall	79.65%	74.77%	77.13%	87.26%	80.26%	83.60%

References

1. Profile C. Cohort profile: the Swiss HIV Cohort study. International journal of epidemiology 2010; 39: 1179-1189.
2. Hoang PD, Cameron MH, Gandevia SC, Lord SR. Neuropsychological, Balance, and Mobility Risk Factors for Falls in People With Multiple Sclerosis: A Prospective Cohort Study. Archives of physical medicine and rehabilitation 2014; 95(3): 480-486.
3. Oh J, Kang S-M, Hong N, Youn J-C, Park S, Lee S-H, Choi D. Comparison of pooled cohort risk equations and Framingham risk score for metabolic syndrome in a Korean community-based population. International journal of cardiology 2014; 176(3): 1154–1155.
4. Marcus CL, Moore RH, Rosen CL, Giordani B, Garetz SL, Taylor HG, Mitchell RB, Amin R, Katz ES, Arens R. A randomized trial of adenotonsillectomy for childhood sleep apnea. New England Journal of Medicine 2013; 368(25): 2366–2376.
5. Müller F, Christ-Crain M, Bregenzer T, Krause M, Zimmerli W, Mueller B, Schuetz P. Procalcitonin Levels Predict Bacteremia in Patients With Community-Acquired PneumoniaA Prospective Cohort Trial. CHEST Journal 2010; 138(1): 121-129.
6. Shibasaki M, Nakajima Y, Shime N, Sawa T, Sessler DI. Prediction of optimal endotracheal tube cuff volume from tracheal diameter and from patient height and age: a prospective cohort trial. Journal of anesthesia 2012; 26(4): 536-540.
7. Hahn U, Krummenauer F, Kölbl B, Neuhann T, Schayan-Araghi K, Schmickler S, von Wolff K, Weindler J, Will T, Neuhann I. Determination of valid benchmarks for outcome indicators in cataract surgery: a multicenter, prospective cohort trial. Ophthalmology 2011; 118(11): 2105-2112.
8. Jain M, Harrison L, Howe G, Miller A. Evaluation of a self-administered dietary questionnaire for use in a cohort study. The American journal of clinical nutrition 1982; 36(5): 931-935.
9. Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Andersen A-MN, Taxbøl D, Hansen KD, Juhl M, Schow TB. The Danish National Birth Cohort-its background, structure and aim. Scandinavian journal of public health 2001; 29(4): 300-307.
10. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. Epidemiology 1991: 155-158.
11. Schneeweiss S, Stürmer T, Maclure M. Case–crossover and case–time–control designs as alternatives in pharmacoepidemiologic research. Pharmacoepidemiology and drug safety 1997; 6(S3): S51-S59.
12. Jurafsky D, James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. 2000.
13. Bekhuis T, Kreinacke M, Spallek H, Song M, O'Donnell JA. Using natural language processing to enable in-depth analysis of clinical messages posted to an Internet mailing list: a feasibility study. Journal of medical Internet research 2011; 13(4): e98.
14. Wu ST, Sohn S, Ravikumar K, Wagholikar K, Jonnalagadda SR, Liu H, Juhn YJ. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Annals of Allergy, Asthma & Immunology 2013; 111(5): 364-369.
15. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association 2010; 17(1): 19–24.
16. Chen W, Fosler-Lussier E, Xiao N, Raje S, Ramnath R, Sui D, editors. A Synergistic Framework for Geographic Question Answering. Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on; 2013: 94–99.
17. Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, Shah NH. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. Journal of the American Medical Informatics Association 2012; 19(e1): e149–156.
18. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, Pickard S, Heidenreich P, Shen S, Weir C. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. Journal of the American Medical Informatics Association 2012; 19(5): 859-866.
19. Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods in molecular biology (Clifton, NJ) 2013; 1168: 275-294.
20. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010; 17(3): 229-236.
21. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 2010; 17(5): 507-513.

22. Osborne JD, Lin S, Zhu LJ, Kibbe WA. Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS). Gene Function Analysis: Springer; 2007. p. 153–69.
23. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the American Medical Informatics Association 2011; 18(5): 601-606.
24. Tang B, Cao H, Wu Y, Jiang M, Xu H, editors. Clinical entity recognition using structural support vector machines with rich features. Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics; 2012: ACM.
25. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC medical informatics and decision making 2013; 13(Suppl. 1): S1.
26. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 2012; 13(6): 395–405.
27. Zhu D, Wu S, Carterette B, Liu H. Using large clinical corpora for query expansion in text-based cohort identification. Journal of biomedical informatics 2014; 49: 275–281.
28. Murphy SN, Wilcox A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). eGEMs (Generating Evidence & Methods to improve patient outcomes) 2014; 2(2): 7.
29. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association 2010; 17(2): 124-130.
30. Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, Marsolo K, McMurry AJ, Sandborg CI, Schanberg LE. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. Journal of the American Medical Informatics Association 2013; 20(1): 172–179.
31. Moser R, Boyer E, Lupinski D, Darer J, Anderer T, Villareal A, Berger P. C-B4–02: Enhancing the Quality and Efficiency of Obstructive Sleep Apnea Screening Using Health Information Technology: Results of a Geisinger Clinic Pilot Study. Clinical medicine & research 2011; 9(3–4): 170-171.
32. Zhang G-Q, Cui L, Teagno J, Kaebler D, Koroukian S, Xu R. Merging Ontology Navigation with Query Construction for Web-based Medicare Data Exploration. AMIA Summits on Translational Science Proceedings 2013; 2013: 285.
33. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making 2006; 6(1): 30.
34. Chen D, Manning CD. A fast and accurate dependency parser using neural networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014: 740-750.
35. Socher R, Lin CC, Manning C, Ng AY. Parsing natural scenes and natural language with recursive neural networks. Proceedings of the 28th International Conference on Machine Learning (ICML-11) 2011: 129-136.
36. Socher R, Manning CD, Ng AY. Learning continuous phrase representations and syntactic parsing with recursive neural networks. Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop 2010: 1–9.
37. Chen W, editor Context-based Natural Language Processing for GIS-based Vague Region Visualization. Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science; 2014: Association for Computational Linguistics.
38. Klein D, Manning CD, editors. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1; 2003: Association for Computational Linguistics.
39. Klein D, Manning CD, editors. Fast exact inference with a factored model for natural language parsing. Advances in neural information processing systems; 2002.
40. Cohen WW, Sarawagi S, editors. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining; 2004: ACM.
41. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H, editors. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. AMIA Annual Symposium Proceedings; 2012: American Medical Informatics Association.
42. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics 2001; 34(5): 301-310.
43. Ristad ES, Yianilos PN. Learning string-edit distance. Pattern Analysis and Machine Intelligence, IEEE Transactions on 1998; 20(5): 522-532.