ADVANCED NATURAL LANGUAGE PROCESSING AND TEMPORAL MINING FOR CLINICAL DISCOVERY

Saeed Mehrabi

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

February 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Josette F. Jones, PhD, Chair

Mathew J. Palakal, PhD

Stanley Yung-Ping Chien, PhD

Xiaowen Liu, PhD

C. Max Schmidt, MD, MBA, PhD

August 17, 2015

DEDICATION

Dedicated to my parents.

ACKNOWLEDGEMENTS

I would like to start by expressing my deepest gratitude to my advisor, Dr. Mathew Palakal, for his brilliant guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I am sincerely grateful to all IUPUI faculties especially my thesis committee chair Dr. Josette Jones for her tremendous help during my PhD studies and thesis work. I also would like to thank my thesis committee members, Dr. Xiaowen Liu and Dr. Chen for accepting to be part of my thesis committee. I would like to thank Dr. Max Schmidt, Dr. Alex and Dr. Heidi Schmidt for their contribution and help in co-authoring various papers. The clinical significance of this work is indebted to Dr. Max Schmidt help and guidance. Special thanks to Regenstrief institute especially Dr. Paul Dexter, Joe Kesterson, and Chris Beesley. I acknowledge my gratitude to Anand Krishnan for his contribution and help. Finally, I would like to thank my colleagues at Mayo clinic especially Dr. Hongfang Liu, the NLP director at Mayo clinic for providing me with additional data to strengthen my explorations, Dr. Sunghwan Sohn for his help in comparison of our negation algorithm with his DepNeg algorithm and revising our Journal of Bioinformatics manuscript and Dr. Ravikumar Komandur Elayvilli for his feedbacks and suggestions.

Saeed Mehrabi

ADVANCED NATURAL LANGUAGE PROCESSING AND TEMPORAL MINING FOR CLINICAL DISCOVERY

There has been vast and growing amount of healthcare data especially with the rapid adoption of electronic health records (EHRs) as a result of the HITECH act of 2009. It is estimated that around 80% of the clinical information resides in the unstructured narrative of an EHR. Recently, natural language processing (NLP) techniques have offered opportunities to extract information from unstructured clinical texts needed for various clinical applications. A popular method for enabling secondary uses of EHRs is information or concept extraction, a subtask of NLP that seeks to locate and classify elements within text based on the context. Extraction of clinical concepts without considering the context has many complications, including inaccurate diagnosis of patients and contamination of study cohorts. Identifying the negation status and whether a clinical concept belongs to patients or his family members are two of the challenges faced in context detection. A negation algorithm called Dependency Parser Negation (DEEPEN) has been developed in this research study by taking into account the dependency relationship between negation words and concepts within a sentence using the Stanford Dependency Parser. The study results demonstrate that DEEPEN, can reduce the number of incorrect negation assignment for patients with positive findings, and therefore improve the identification of patients with the target clinical findings in EHRs. Additionally, an NLP system consisting of section segmentation and relation discovery was developed to identify patients' family history. To assess the generalizability of the negation and family history algorithm, data from a different clinical institution was used in both algorithm evaluations.

The temporal dimension of extracted information from clinical records representing the trajectory of disease progression in patients was also studied in this project. Clinical data of patients who lived in Olmsted County (Rochester, MN) during

1966 to 2010 was analyzed in this work. The patient records were modeled by diagnosis matrices with clinical events as rows and their temporal information as columns. Deep learning algorithm was used to find common temporal patterns within these diagnosis matrices.

Josette F. Jones, RN, Ph.D., Chair

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION & BACKGROUND

1.1 Problem Statement

Electronic Health Records (EHR) contains valuable longitudinal clinical information that can be used for various applications such as Clinical Decision Support Systems (CDSS), medication reconciliation, public health emergency surveillance, quality measurements, etc. However these applications are not readily feasible due to unstructured nature of data represented in clinical documents. Natural Language Processing (NLP) has been used to extract and store clinical concepts in a structured format. Concept identification or extraction is one of the popular sub tasks of NLP in enabling secondary use of EHR. However this task is not easy because the meaning of a concept is significantly affected by modifiers such as negation (i.e. no symptom attributable to her pseudocyst) and family history (i.e. her father has diabetes). In this work, we have developed a novel negation algorithm to detect the negation status of clinical concept. The system was developed using the data of patients with pancreatic cyst and it was evaluated on any clinical disorder or sign and symptoms.

In the era of precision medicine, accurately identifying familial conditions is crucial for providing target treatment. Personalized medicine is defined as "the use of combined knowledge (genetic or otherwise) about a person to predict disease susceptibility, disease prognosis, or treatment response and thereby improve that person's health" (Redekop & Mladsi, 2013). Personalized medicine could be misinterpreted as treatment developed uniquely for each individual therefore it is replaced with precision medicine in recent days.

Identification of familial conditions requires detailed family history information. The family history information can be available in clinical notes by "documenting parents' and siblings' age and health (or age and cause of death), as well as a checklist of conditions with environmental and hereditary etiologies" (Rich, et al., 2004). We have developed a rule-based NLP system to identify patients with family history of pancreatic cancer and evaluated our system on data of two institutions to assess its portability.

The extracted concepts across longitudinal patients' records consist of phenotypic information, disease characteristics, treatment and outcome, which describe the patients'

course of disease. In order to represent the temporality of the extracted information from patients record, each patient's records was modeled as a matrix of temporal clinical events where common pattern discovery methods can be applied to matrices representing a cohort of patients. In this work we developed a deep learning algorithm for temporal pattern discovery over longitudinal healthcare records.

Chapter two describes state of art concept identification and our experiments in detection of pancreatic cyst concepts from clinical records. Chapters three and four focus on two challenges of identification of concept within a sentence namely negation and family history. If we consider the temporal order of the extracted concepts, we can find common phenotypic patterns among cohort of patients with similar conditions. In chapter five, we describe our model to represent each patient records as a diagnosis matrix and a deep Boltzmann machines that was used to find common temporal patterns on a cohort of patients.

1.2 Clinical NLP Introduction and Background

NLP allows computers to understand natural language used by humans as opposed to artificial language used by computers and is defined as the "formulation and investigation of computationally effective mechanisms for communication through natural language" (Carbonell & Hayes, 1992).

Patients' medical records (e.g. radiology reports, pathology reports, clinical notes, and discharge summaries) include wealth of information about patients that are in free text format. NLP can be very useful tool in extracting information from these free text format documents and creating structured information that can be used for further knowledge extraction by researchers. Research on processing of natural language in clinical notes has been slower in comparison to other domains such as biomedical or general English due to HIPAA privacy concerns in sharing clinical data and lack of common standards in annotation. One of the oldest and most studied clinical NLP systems is the Medical Language Extraction and Encoding System (MedLEE) developed by Carol Friedman et al at the Columbia University in the mid 90s (Friedman, Hripcsak, DuMouchel, Johnson, & Clayton, 1995). MedLEE was initially developed on chest radiology reports (Friedman, Alderson, Austin, Cimino, & Johnson, 1994), however it is

now, extended to any kind of clinical notes. Informatics for Integrating Biology and the Bedside (i2b2) is an NIH-funded national center that has organized shared tasks focusing on problems less studied in clinical NLP and sharing annotated clinical notes that removed some of barriers to the development of clinical NLP systems (Chapman, Nadkarni, Hirschman, W D'Avolio, Savova, & Uzuner, 2011). The first i2b2 challenge in 2006 was on automatic identification of smoking status of patients from information contained in discharge summaries (Uzuner, Goldstein, Luo, & Kohane, 2008). The second i2b2 challenge focused on obesity and its comorbidities (Uzuner Ö., 2009). It was a multi-label classification task that classified obesity and its comorbidities as present, absent, questionable or unmentioned in discharge summaries. The third i2b2-shared task (2009) was on medication extraction (Uzuner, Solti, & Cadag, 2010). Various NLP systems using regular expression, machine learning or hybrids of them were used to extract information such as medications, dosages, modes of administration, frequency of administration and the reason for administration. The fourth challenge (2010) was on extraction of concepts (problems, treatments or tests), assigning assertions (whether the concepts are present, absent associated with someone else, hypothetical, conditional or possible) and classifying relations between those concepts (with eight different types of relations, for instance: treatment worsened a medical problem, treatment improved a medical problem, etc.) (Uzuner, South, Shen, & DuVall, 2011). The fifth challenge (2011) was on anaphora resolution (Uzuner, Bodnari, Shen, Forbush, Pestian, & South, 2012). The sixth challenge was to correctly identify and interpret temporal relations. i2b2 provided a corpus of de-identified discharge summaries with annotated clinical events, temporal expressions and relations to the NLP community (Suna, Rumshisky, & Uzuner, 2013). The latest i2b2 challenge was on de-identification of medical records by removing protected health information (PHI) and identification of the risk factors of heart disease over time.

CHAPTER TWO: CONCEPT IDENTIFICATION

2.1 Introduction

Concept extraction or identification is a subtask of information extraction where phrases of interest are extracted from text. In clinical NLP, these concepts can be phrases referring to disorders (i.e. he had a hemiarthroplasty for osteonecrosis), sign and symptoms (i.e. His last seizure was on July), treatment or medication (hypertension was controlled on hydrochlorothiazide), etc.

Various tools have been developed for the recognition of diverse clinical entities including medications, dosages, anatomical sites, disease and disorders, etc. based on dictionaries (Eriksson, Jensen, Frankild, Jensen, & Brunak, 2013), rules (Sun & Nguyen, 2010) and machine learning algorithms (Jiang, et al., 2011) (Zhang & Elhadad, 2013) or combination of these methods (Xu, Hong, Tsujii, & Chang, 2012). However there is no single approach that is the Holy Grail for this problem. Domain adaptation is required in each approach for improved performance. Clinical terminologies are not only different across domains (cardiovascular versus orthopedics), but also distinctive from one type of clinical notes to the other, for instance the language used in radiology reports are unlike that of pathology reports. Therefore we can consider each particular domain of healthcare or type of clinical note as a sublanguage. Sublanguage is defined as "the specialized form of natural language which is used within a particular domain or subject matter" (Grishman & Kittredg, 1986). Sublanguage is characterized by a specialized vocabulary, syntax or semantic relationship.

The pattern discovery in sublanguage analysis involves annotated corpora where concepts are marked based on their relevance or non-relevance to the extraction task. Normally pattern discovery starts with a set of initial patterns that are good candidates of the topic of interest. Figure 2.1 shows the regular expression pattern development flowchart where the annotated documents are divided into two parts of training and test sets and following steps are performed. Regular expression or regex is a pattern of characters matching specific strings of text.

1) Training set is loaded and randomized
2) The initial set of regex patterns is applied to the training set

3) The irrelevant extracted concepts and relevant missed concepts are used to modify the initial regex rules
4) Final rule set is applied to the train set (Steps 3 & 4 are repeated until the target accuracy is reached for the training data)
5) Test set is loaded and randomized
6) Final rule set is applied to the test set
7) Extracted concepts are compared to the reference standard to evaluate the accuracy of final regex rules.

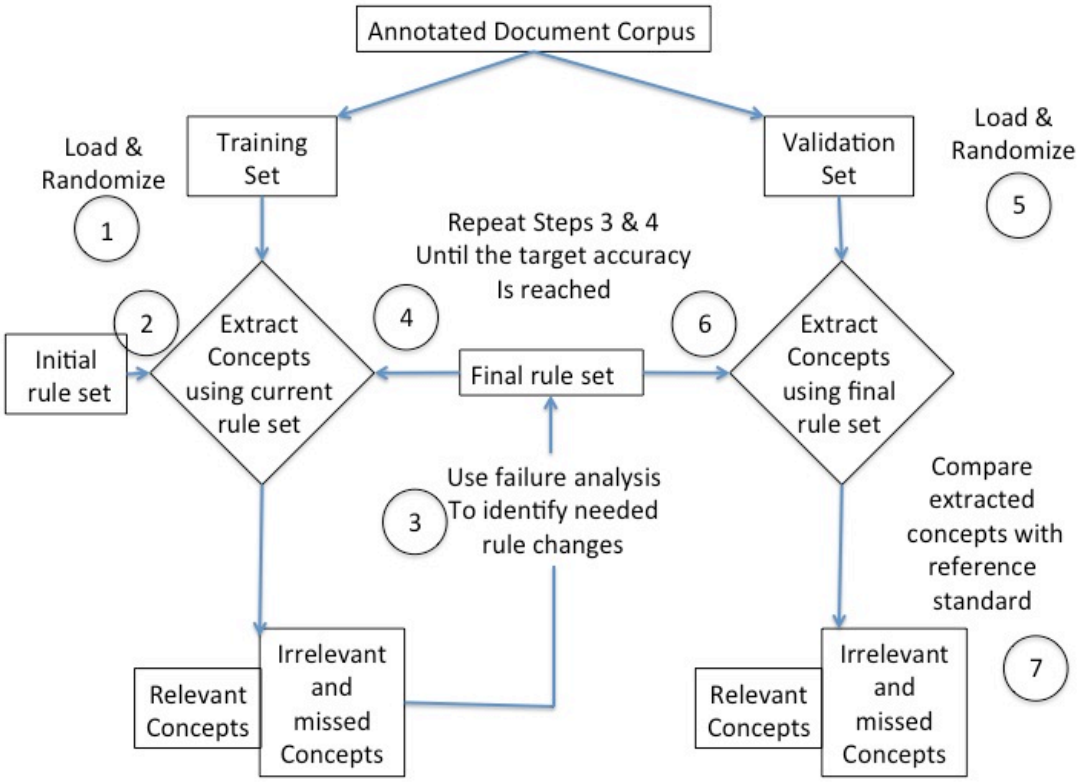


Figure 2.1 Regex Development Flowchart

In this work, we built an NLP sublanguage model to identify patients with pancreatic cyst from unstructured clinical notes. In the followings, we provide a brief introduction on why identification of patients with pancreatic cyst is important and some of the previous works along with our methodology, dataset and results of our explorations.

2.1.1 Pancreatic Cyst

Cancer of pancreas is the fourth leading cause of death in the US (Howlader, et al., 2012) and eight in the world (Cancer Research UK, 2012). It is estimated that 48,960 (24,840 men and 24,120 women) will be diagnosed with and 40,560 (20,710 men and 19,850 women) will die of cancer of the pancreas in 2015. Although considerable progress has been made in cancer survival rates over the past decades, 5-year survival rate for pancreatic cancer has hardly changed, rising from 3% in 1975 to 5.8% in 2012 (Howlader, et al., 2012). The poor prognosis of pancreatic cancer is primarily due to its late stage diagnosis with more than 80% of patients presenting with locally advanced or metastatic disease where available systemic therapies remain largely ineffective (Howlader, et al., 2012). Therefore, the only chance for improving survival is to detect pancreatic cancer at an earlier stage (Cleary, et al., 2004).

Pancreatic cancer has several risk factors such as obesity, smoking, and alcohol intake but its exact causes are not yet known. Screening of general population for early identification of pancreatic cancer is not feasible because of its low incidence and lack of effective screening tests to identify patients at earlier stages of the disease. Yet, screening high-risk populations such as patients with a family history of pancreatic cancer and patients with pancreatic cysts represent two windows of opportunity for early detection of pancreatic cancer.

Family history of pancreatic cancer increases the risk of developing pancreatic cancer (Permuth-Wey & Egan, 2009). One first-degree relative (a parent or sibling) with pancreatic cancer increases the risk to 7 to 9-fold and three or more first-degree relatives with pancreatic cancer increase the risk to 17 to 32-fold (Klein, et al., 2004). Risk is also increased if a first-degree relative was diagnosed with pancreatic cancer before age 50 (Brune, et al., 2010).

Pancreatic cysts are well-recognized precancerous lesions. Several studies report their incidence in 2.6% of abdominal CTs and in up to 20% of MRI studies. (Laffan , et al., 2008), (Lee, Kim, Choi, Hong, & Kim, 2011) (Zhang, Mitchell, Dohke, Holland, & Parker, 2002). Pancreatic cysts have degrees of malignancy based upon their type and degree of dysplasia. There are various types of pancreatic cysts, which include pseudocysts, serous cystic neoplasms (SCN), mucinous cystic neoplasms (MCN) and

intraductal papillary mucinous neoplasms (IPMN). Only MCN and IPMN may progress to invasive pancreatic adenocarcinoma (Pandol, Gukovskaya, Edderkaoui, Dawson, Eibl, & Lugea, 2012). Because these lesions are typically asymptomatic and incidentally detected, they are mostly ignored, though they still may harbor a malignant potential of 20% to 90% (Schmidt, et al., 2007), (Lennon & Wolfgang, 2013). Additionally, 10% of pancreatic cancers have a familial or hereditary component (Shi, Hruban, & Klein, 2009), (Permuth-Wey & Egan, 2009).

Accurate identification, surveillance and treatment of patients with pancreatic cysts or family history of pancreatic cancer represent an opportunity to prevent pancreatic cancer. Much information about pancreatic cysts and patient's family history can be found in free text format in various narrative medical reports including pathology, cytopathology, radiology (MRI, CT, EUS) and physician's clinical reports. Therefore NLP is required to harness the information embedded in clinical narratives.

2.2 Previous Works in Pancreatic Cyst Identification

Al-Haddad et al. (2013) used Regenstrief EXtraction tool (REX) to identify patients with a confirmed, pathological, diagnostic report of IPMN. (Al-Haddad, Friedlin, Kesterson, Waters, Aguilar-Saavedra, & Schmidt, 2010) REX is a rule-based NLP method that uses a window of words before and after a medical concept to determine if the concept is negated, affirmed, related to a patient or his family members. REX was applied to 165,000 clinical reports of 5694 patients and validated by testing its performance against a manually created surgical database of patients who had a surgical resection of IPMN at Indiana University (IU) hospital during 1985-2009. The NLP system detected 208 out of 215 patients who had a confirmed pathology report of IPMN in the surgical database, and it found an additional 37 patients that were not included in the surgical database.

Friedlin et al. (2010) compared the identification of pancreatic cancer patients using ICD9 codes and NLP processing of clinical notes using REX (Friedlin, et al., 2010). Zhao et al used the PubMed knowledge and EHR data to develop a weighted Bayesian network (BN) to predict pancreatic cancer (Zhao & Weng, 2011). A weight score was calculated for twenty selected risk factors of pancreatic cancer based on the

positive, negative and neutral association between these risk factors and pancreatic cancer using PubMed articles. The risk factors were then used in construction of a BN with a weighted inference model. The EHR data of patients with and without pancreatic cancer were used to predict risk of pancreatic cancer patients. In another study, Markov chain model was used to compare four different management strategies (pancreaticoduodenectomy, yearly non-invasive surveillance, yearly invasive surveillance, do nothing) in patients with branch duct IPMN (Weinberg, Spiegel, Tomlinson, & Farrell, 2010). PancPRO is a Bayesian modeling framework used to assess the pancreatic cancer risk of patients with family history of pancreatic cancer (Wang, Chen, Brune, Hruban, Parmigiani, & Klein, 2007). A genetic model for mutation susceptibility was specified and prevalence associated with these mutations was estimated. Bay's rule was used to convert the genotype probability given phenotypic information.

2.3 Methodology

2.3.1 Pancreatic Cyst Identification

Unstructured Information Management Architecture (UIMA) was used as the framework for developing our text-mining system. UIMA is a software architecture developed by IBM for the analysis of unstructured content that includes text, audio or video data. It is openly available through the Apache software foundation (Ferrucci & Lally, 2004). UIMA is composed of processing units called analysis engine (AE) that analyze unstructured data and infers information from them. AEs are constructed from analysis logics called annotators. An AE may contain a single annotator (called primitive AE) or multiple annotators (called aggregate AE). Figure 2.2 shows the UIMA pipeline integrated to extract the information from medical reports. The input goes through a series of tasks depicted in each block. The input for every step is the output of its predecessor task.

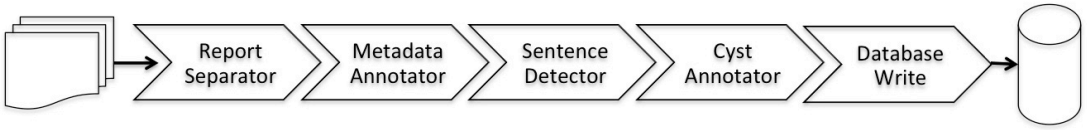


Figure 2.2 Analysis Engines Used in the UIMA Pipeline

The initial input is a text file containing all the reports for all the patients. Report separator AE in this pipeline separates each report. Each report has Metadata information such as medical record number (MRN), report id, report name, report date and report body that is the notes dictated by physicians. Metadata annotator AE extracts these Metadata from each report using regular expressions. The next AE is the cTAKES sentence detector (Savova, et al., 2010). This is a UIMA wrapper around the openNLP sentence detector (OpenNLP), which was originally used during the first pass. Because sentences were breaking on new lines, it was replaced with Ytex sentence detector. Ytex sentence detector is a modified version of cTAKES to deal with breaking sentences on new lines (Garla, et al., 2011). The cyst annotator AE extracts cyst concepts within each report using regular expression. Table 2.1 shows each concept and the regular expression developed to extract it from the patients' medical reports.

Table 2.1 Various Cyst Types and Their Associated Regular Expression Patterns

Concept	**Regular Expression**
Pancreatic Cyst	(?i)(pancreatic cyst(s)? | cyst(s)?(of| in)? the pancreas | pancreatic cystic)
Pancreatic pseudocyst	(?i)(pseudo\s?cyst(s)?)
Mucinous Cyst	(?i)(mucinous cyst(ic | ts) neoplasm | mucinous cystadenoma | intraductal papillary mucinous | \b(mcn)\b | \b(mca)\b | \b(ipmn)|\b(ipmt)\b)
Serous Cyst	(?i)(serous cyst(ic | s | adenoma) | \b(sca)\b)
Retention Cyst	(?i)(retention cyst(ic | s))
Cystic neuroendocrine tumor	(?i)(cystic neuroendocrine tumor | cystic neuroendorine | neuroendocrine cyst(ic|ts)|islet cell cyst tumor|cystic islet cell tumor)
Cystic degeneration cancer	(?i)(cystic degeneration cancer | cystic degeneration | degeneration cyst(ic | s))
Duct ectasia	(?i)(duct(al) ectasia | ectasia of the(pancreatic)? duct | ectasic duct)
Duct dilatation	(?i)(pancreatic duct(al) dilatation | dilatation of the(pancreatic)? duct | dilated(pancreatic)? duct)

Concepts that are used to identify pancreatic cysts in medical records were assembled by our medical team and additional keywords were added by searching through literature and Unified Medical Language System (UMLS) knowledge base (Al-Haddad, Friedlin, Kesterson, Waters, Aguilar-Saavedra, & Schmidt, 2010). The initial set of regular expressions based on these concepts was applied to the development dataset described in more detail in section 2.3.1. The false positives and false negative cases were analyzed and the regular expression patterns were modified accordingly. This process was repeated several times until we reached the target precision and recall on the training set. At the final step the algorithm was applied to the test set for evaluation.

The regular expressions shown in Table 2.1 only match text strings in clinical notes without considering the contextual information surrounding the concept of interest. For instance, the concept "ductal extasia" in the sentence "*Chronic mastitis or* ***ductal ectasia*** *left breast*" does not exist in pancreas or the concept "*pancreatic ductal dilatation*" identified in the following sentence "*No pancreatic ductal dilatation is visualized."* is negated.

We developed a rule-based algorithm described in the next section to remove concepts that happened in other organ than pancreas. In order to identify the negation status of concepts, we used a widely used negation algorithm called NegEx (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001). NegEx is a string-matching algorithm that looks for negation terms such as 'No', 'No evidence of', 'Rule out', etc., within the sentence containing the concept. Because NegEx failed to consider the contextual relationship between negation words and concepts, we used Dependency parser on top of NegEx to improve its performance. Dependency parser is a binary asymmetric relation that holds between a token (word) and its dependents in a sentence. Chapter three provides more detailed on our negation methodology.

2.3.2 Contextual Rule-Based Algorithm

Sentences A to D in this section represent sentences that were extracted based on patterns listed in Table 2.1. Sentence "A" was extracted because it contained the keywords "dilated duct" which could identify the presence of pancreatic duct dilatation. However, duct dilatation can also occur in the biliary system or breast tissue therefore

only clinical records that contained the term “pancrea” were selected before searching for the pancreatic dilated duct concept.

A) *“Additional views were obtained of the right breast in the retro-areolar region which demonstrate **dilated ducts** and increased nodularity as well as two coarse calcifications.”*

Sentences “B” and “C” extracted because they contained the “ductal ectasia” concept. Both of these sentences are not related pancreatic ductal ectasia, therefore ductal ectasia was only searched in reports that contained the keyword “pancrea” but did not contain breast-related terms: “breast,” “nipple,” “areola,” or “mammogram,” or biliary-related terms: “biliary,” “bile” or “biloma.”

B) *FINDINGS: The liver is of normal size without intraparenchymal mass or biliary ductal ectasia.*

C) *Chronic mastitis or **ductal ectasia** left breast.*

Mucinous and serous cysts can occur in the ovary (Sentence D) as well as in the pancreas. Thus, a filter was applied to search for keyword “ovary” to remove reports pertaining to ovarian-specific mucinous and serous cysts.

D) *POSTOPERATIVE DIAGNOSIS: Left ovarian **mucinous cystadenoma**.*

And finally, in the “Impression” section of radiology notes, the radiologist typically reports the final diagnosis, conclusions of the radiographic study and recommendations for further evaluations. Such a report may contain a phrase like “evaluate for dilated ducts”, yet this recommendation does not mean that a patient has pancreatic dilated ducts. Therefore sentences that begin with “evaluate”, “assess” or “indicate” were removed.

Figure 2.3 shows a process that was adopted to remove unnecessary reports before applying the regex patterns.

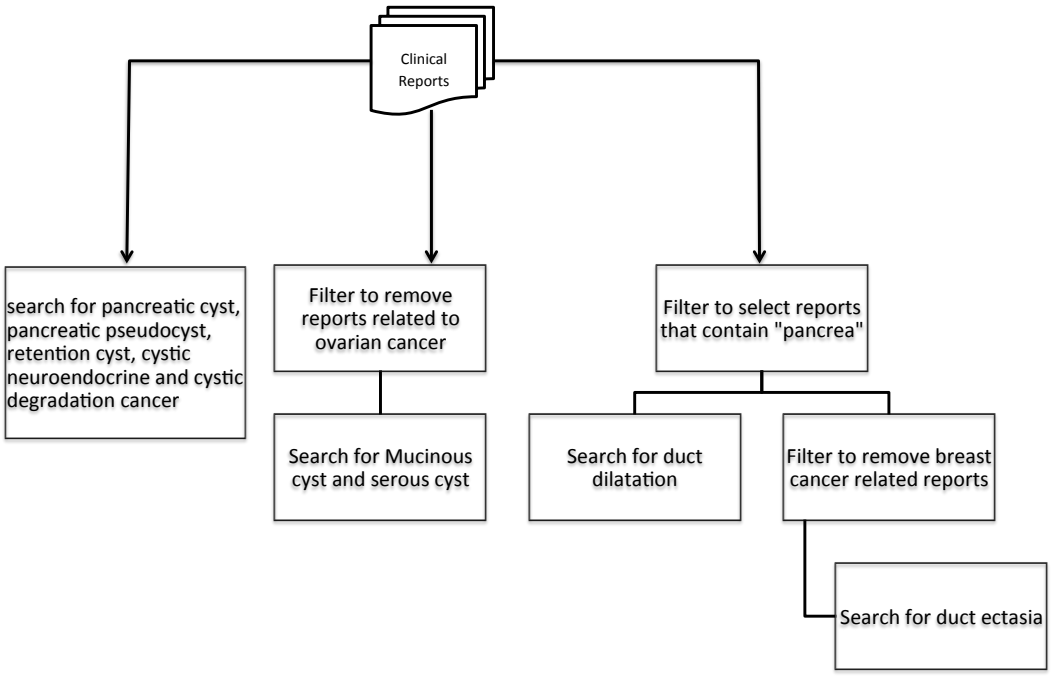


Figure 2.3 Contextual Rule-Based Flowchart

2.4 Datasets

Longitudinal health records including discharge summary, surgical pathology document, imaging reports (abdominal MRI, CT with/without contrast, Ultrasound, etc.) and other clinical notes (procedure notes, visit notes, letter, consultation, etc.) of patients who visited the Sidney & Lois Eskenazi Hospital in Indianapolis was used in this study. The Eskenazi Hospital is a 316-bed hospital providing a comprehensive range of primary and specialty care services in central Indiana. It is comprised of providers who are faculty and residents of the IU School of Medicine. This study was conducted under an approved Institutional Review Board (IRB) protocol by IU.

2.4.1 Development Dataset

A random number of 44 patients were selected that contained the terms ‘pancreas’ and ‘cyst’ somewhere in their reports. From the patients with ‘pancreas’ and ‘cyst’ in a text report, all of their text reports of any type were extracted forming a corpus with 1064 reports. The average number of reports per patient in our corpus is 23.64 with minimum of 1 and maximum of 221 reports per patient. Pancreatic cysts were present in 28 out of

44 patients and absent in the remaining 16 patients' records. The data set was randomly divided into two sets of train and test set, each containing 14 patients with pancreatic cyst and 8 patients with no pancreatic cysts. The test set is composed of 703 patient records, of which 509 belong to patients having pancreatic cysts and the remaining 194 records to patients without diagnosis of pancreatic cysts. The train set is composed of 316 patient records, of which 240 reports belong to patient having pancreatic cyst and the remaining 121 records to patients without pancreatic cysts. Two pancreatic-cyst surgeon experts created the gold standard data. The discrepancies between the two annotators were resolved by discussing the differences in annotation. Cohen's kappa was used to measure the inter annotator agreement K=88% (Carletta, 1996)

2.4.2 Monthly Dataset

All patients who visited the Sidney & Lois Eskenazi Hospital in Indianapolis, Indiana for any reason during March-December 2013 were retrieved. Table 2.2 shows the number of patients and their reports for each month.

Table 2.2 Number of Patients and Their Reports in Each Month

Month	Total Number of patients	Total Number of reports
March	7950	97535
April	6419	78451
May	6036	70100
June	7514	78110
July	7390	81991
August	7534	79072
September	7826	80973
October	7794	84868
November	7152	80916
December	7086	79957

2.5 Results

Table 2.3 shows the total number of patients identified with pancreatic cyst and their clinical records in each month.

Table 2.3 Number of Patients with Pancreatic Cyst in Each Month

Month	#Identified Patients	# Identified patients' reports
March	227	443
April	199	352
May	186	325
June	165	329
July	196	403
August	197	395
September	186	308
October	225	429
November	194	364
December	192	380

Table 2.4 shows the number of positively identified patients with pancreatic cysts, with a family history of pancreatic cancer and with both pancreatic cysts and a family history of pancreatic cancer.

Table 2.4 Number of Patients that Have Pancreatic Cyst, Family History of Pancreatic Cancer or Both

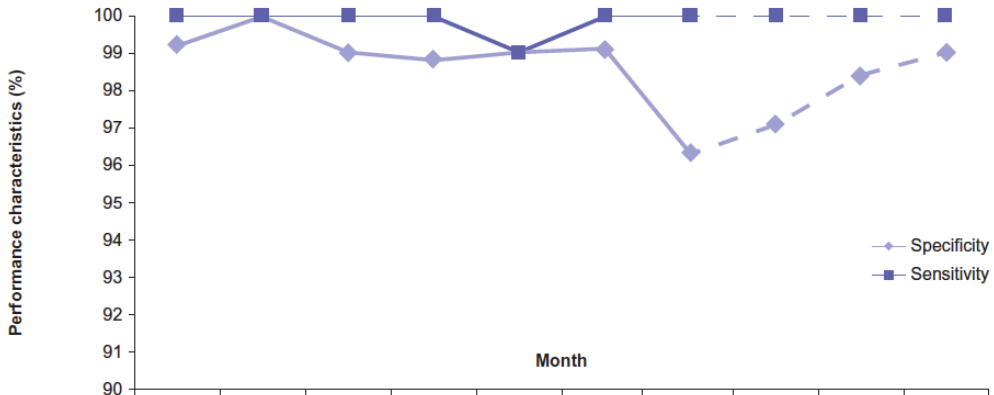
Months	Pancreatic Cyst	Family History of Pancreatic Cancer	Pancreatic Cyst and FH of Pancreatic Cancer
March	98	15	2
April	106	6	0
May	102	10	3
June	85	6	2
July	98	12	1
August	109	10	0
September	104	12	0
October	136	11	1
November	119	10	1
December	97	10	1

We used seven months of data from March to September 2013 in two rounds of evaluations to develop the pancreatic cyst detection algorithm. The last three months of data was used for final evaluation of the system. For each patient there would be either one or multiple sentences containing pancreatic cyst. The evaluation in Table 2.5 is based on *patient level* meaning that even if one sentence is affirmed, the patient is considered to be a positive case of pancreatic cyst. Conversely a patient does not have a cyst when the system correctly identifies all the sentences that are negated. We considered false positive and negative from a medical perspective (i.e., presence or absence of medical problem) meaning that if a patient has pancreatic cyst and the system considered that as a negative case (patient does not have pancreatic cyst), the result was evaluated as false negative. Similarly, if a patient does not have pancreatic cyst and the system result was affirmed (patient has pancreatic cyst), it was evaluated as false positive.

Table 2.5 Monthly Results of the Pancreatic Cyst Identification

Month		Identified Patients	True Positives	True Negatives	False Negatives	False Positives	Sensitivity	Specificity	False Positive Rate
First Evaluation	Mar.	227	98	126	0	3	100	97.67	2.33
	Apr.	199	93	104	0	2	100	98.11	1.89
	May	186	83	100	0	3	100	97.09	2.91
	Jun.	165	79	84	0	2	100	97.67	2.33
	Jul.	196	97	96	1	2	98.9	97.96	2.04
	Aug.	197	86	108	0	3	100	97.30	2.70
	Sep.	186	78	103	0	5	100	95.37	4.63
Second Evaluation	Mar.	227	98	128	0	1	100	99.22	0.78
	Apr.	199	93	106	0	0	100	100	0.00
	May	186	83	102	0	1	100	99.03	0.97
	Jun.	165	79	85	0	1	100	98.84	1.16
	Jul.	196	97	97	1	1	98.9	98.98	1.02
	Aug.	197	86	109	0	2	100	98.2	1.80
	Sep.	186	78	104	0	4	100	96.3	3.7
Oct.		225	85	136	0	4	100	97.14	2.86
Nov.		194	73	119	0	2	100	98.35	1.65
Dec.		192	94	97	0	1	100	98.98	1.02

There were 1,359 patients with at least one mention of a ‘pancreatic cyst’ from March to September. Three of those identified patients were excluded (n = 1,356) because they were detected twice (one patient identified in June had previously been identified in March and two patients identified in August and September had previously been identified in July and August, respectively). The NLP algorithm identified 623 positive (patients with pancreatic cyst or pancreatic ductal dilation) patients. Manual (physician expert) review identified 615 positive patients (nine and one patients were found to be false positives and false negative, respectively). This resulted in a calculated prevalence of pancreatic cysts of 1.2%. Over the 7-month period, the mean sensitivity of the NLP algorithm for identification of a pancreatic cyst and/or ductal dilation was 99.85% (range 98.98–100). Similarly, the mean specificity was 98.8% (range 96.3–100) (Figure 2.4).



	March	April	May	June	July	August	September	October	November	December
Specificity	99.2	100	99	98.8	99	99.1	96.3	97.1	98.4	99
Sensitivity	100	100	100	100	99	100	100	100	100	100

Figure 2.4 System Performance over the Study Period for Pancreatic Cyst Identification

2.6 Discussion

Pancreatic cancer is a deadly cancer due in part because it is typically diagnosed in advance stages when there is no effective treatment. Early detection of pancreatic cancer is possible through surveillance of patients at risk. Current strategies for early diagnosis of pancreatic cancer have focused on serum biomarkers. The most commonly used biomarker is serum carbohydrate antigen 19-9 (CA19-9). Its use as a screening tool in the general population, however, would be suboptimal because of its low sensitivity

(median 79%, range 70–90%) and specificity (median 82%, range 68–91%) (Goonetilleke & Siriwardena, 2007). Small studies have suggested that the use of a combination of biomarkers instead of an individual biomarker may improve sensitivity and specificity (Behnam & Smith, 2014) (Firpo, et al., 2009). Given the high genetic heterogeneity of pancreatic adenocarcinoma (Jones, et al., 2008), such may require a large number of biomarkers and is not likely to be routinely utilized soon. Similarly, cross-sectional imaging studies have low performance characteristics for screening pancreatic cancer in the general population (Sahani, Shah, Catalano, Boland, & Brugge, 2008) (Nabavizadeh, Greenleaf, Fatemi, & Urban, 2012).

Patients with pancreatic cysts or a hereditary predisposition to pancreatic cancer are considered to be at higher risk of developing cancer. Pancreatic cysts, especially mucinous cysts, are well-established precancerous lesions. As they have the potential to develop into invasive adenocarcinoma in a median of 5 years (range 2–20) (Tanaka, et al., 2006) (Tanaka, et al., 2012), tracking them closely for clinically relevant changes may represent a window of opportunity for pancreatic cancer prevention and early detection. This study showed that the NLP system could accurately detect patients with premalignant cysts (mucinous cysts), thus ensuring adequate management. Although 'pseudocyst' is a benign condition that alone does not require screening for pancreatic cancer, we included it in the final list of 'pancreatic cyst' concepts to be thorough and not miss patients with potentially other types of cysts. Our reasoning was based on manual review of multiple cross-sectional imaging studies reporting a pseudocyst in spite of the absence of clinical and radiological evidence of pancreatitis. Similarly, hypothetical terms, such as 'may represent', were considered affirmed to avoid missing patients with a potentially premalignant condition.

The present study, over a 10-month period, demonstrates that it is feasible and inexpensive to automate the identification of patients with pancreatic cyst(s) and/or pancreatic ductal dilation using NLP. Our algorithm allowed tracking of those patients with high sensitivity (99.9%) and specificity (98.8%). Although manual review remains an important part of the study, patient capture is easier, faster and more thorough when employing a NLP algorithm. Incidentally discovered pancreatic cysts may be inconsequential. The data currently available on the natural history of pancreatic cysts

and their malignant potential have some inherent selection/referral bias, and thus the percentage of truly consequential cysts might be lower in a general population screened by an automated process. The present study aimed to confirm feasibility of an automated process using NLP for pancreatic cyst screening. It accomplished this with high sensitivity and specificity. To analyze the potential for individual lives and cost-effectiveness, further studies including public health cost analyses would be needed to compare the cost of a preventive strategy with follow-up examinations versus the cost of pancreatic cancer treatment.

Once patients with pancreatic cysts are correctly identified, screening this 'at risk' subpopulation for pancreatic cancer may be more feasible because the pretest probability is increased, thus compensating for the suboptimal sensitivity/specificity of currently available biomarkers/imaging studies.

Beyond the identification and tracking of patients with pancreatic cysts, our system sets ground for improved pancreatic cancer screening. As this algorithm is adaptable, it can be incorporated into any hospital electronic system to help capture patients with pancreatic cysts. We anticipate the use of this algorithm as a template for other regional health information organizations (e.g. Boston, Utah, Stanford and Vanderbilt). The ultimate goal is to move towards establishment of a national pancreatic cyst registry. It may lead to a more organized national initiative for pancreatic cancer prevention and early detection, and optimal education of both healthcare providers and patients on current management and screening resources available.

2.7 Conclusion

Patient care and clinical research up until the present are largely based on retrospective or prospective collection of data into databases, which is often done manually. The increased utilization of EHR by medical centers has created new patient care and clinical research possibilities. NLP helps researcher to automate the process of patient data extraction that eventually increases the research scope (data volume, time) and statistical power while decreasing the required manpower utilization. In this chapter, we described pancreatic cyst identification based on sublanguage analysis. We briefly touched on the importance of the contextual information surrounding a concept of interest

in concept identification task. For instance, the presence of a clinical finding in narrative patient's report does not imply that the patient has the finding. In the next chapter, we describe our negation detection algorithm that is built on top of NegEx and compare its performance with NegEx on wide set of clinical concepts. Another contextual challenge is to determine if the identified concepts belong to patient or his family members. In chapter four we discuss our NLP system that is developed to identify patient with family history information.

CHAPTER THREE: NEGATION

3.1 Introduction

A study of negation has shown that clinical observations are frequently negated in clinical narratives (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001). For example, physicians often report that a condition is absent in a patient.

Negative clause is defined as “an assertion that some event, situation, or state of affairs does not hold. Negative clauses usually occur in the context of some presupposition, functioning to negate or counter-assert that presupposition” (Payne, 1997).

Negation detection in clinical language tends to be very trivial in sentences such as "no fracture", "patient denies headache", and “she does not have marked dysmenorrhea”. Therefore simplistic approaches such as NegEx (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001) that use negation cue words without considering the semantic of a sentence perform well. However, the simplistic approaches sometimes fail to correctly identify the negation status of clinical concepts in sentences with complex structure. We have faced with this problem while using NegEx in our NLP system that automates the identification and tracking of patients with pancreatic cysts described in the previous chapter (Roch, et al., 2015). Table 3.1 shows some examples of such sentences where NegEx incorrectly negates pancreatic cyst concepts.

Table 3.1 Example of Sentences Where NegEx Failed to Capture the Correct Negation Status of Concepts Denoted by Bold Letters

Record Type	Sample Sentence
Discharge Summary	Additionally, there was no evidence of extension of his infected **pseudocyst** into the psoas muscle.
Abdomen CT	There is no significant interval change in the 2 large pancreatic **pseudocysts**.
OPERATIVE REPORT	We confirmed no evidence of epithelium consistent with a **pseudocyst**.
Consultation	Acute pancreatitis with pseudocyst, with no obvious complications of the **pseudocyst** at this point in time.
Liver CT W Contr	Although there is no discretely visualized or abnormal enhancing pancreatic mass, there is marked **pancreatic duct dilatation** with side duct ectasia and abrupt cutoff of the pancreatic duct within the pancreatic head.

Aiming to reduce the number of missing pancreatic cyst patients in our NLP system inspired us to improve the negation assignment of NegEx by incorporating dependency parsing into NegEx. Dependency relation is a binary asymmetric relation between tokens within a sentence that has been shown to improve various NLP tasks including information extraction (Fundel, Küffner, & Zimmer, 2007), negation detection (Wilson, Wiebe, & Hoffmann, 2005), entity disambiguation (Finkel, Dingare, Nguyen, Nissim, Manning, & Sinclair, 2004) and many others (Cohen & Elhadad, 2012). We developed and tested our negation identification algorithm focusing on only pancreatic cyst concepts using a single institution data set. In order to evaluate its performance on other clinical concepts and dataset, we applied our system on 159 clinical notes from Mayo Clinic where clinical findings such as disorders and signs/symptoms have been annotated. We compared the performance of our algorithm on Mayo Clinic dataset with NegEx.

3.2 Previous Works

Negation detection has been the main or sub task of several challenges in NLP. Assertion classification was one of the three tasks in the 2010 i2b2/VA shared task where each medical concept had to be classified into one of six categories of "*present*", "*absent*", "*possible*", "*conditional*", "*hypothetical*", and "*not associated with the patient*" (Uzuner, South, Shen, & DuVall, 2011). Processing modality and negation was the main task of Question Answering or Machine Reading Evaluation (QA4MRE) lab at CLEF 2011 (Morante & Daelemans, 2011). Negation and speculation in NLP (NeSp-NLP 2010) (Morante & Sporleder, 2010), identifying hedges and their scope in CoNLL-2010 shared task (Farkas, Vincze, Mora, Csirik, & Szarvas, 2010), and SEM 2012 shared task of resolving the scope and focus of negation (Morante & Blanco, 2012) are few other initiatives that show the growing importance of negation processing in the NLP research community.

Corpora used in 2010 i2b2/VA and CoNLL-2010 shared tasks are available to researcher with signing a data use agreement to facilitate the development and evaluation of clinical NLP algorithms. BioScope corpus that was used as part of the CoNLL-2010 shared task has been created by annotating negation and uncertainty in biomedical texts is

also publicly available (Vincze, Szarvas, Farkas, Móra, & Csirik, 2008). BioScope corpus consists of clinical text, abstract and full text of scientific articles. The free text clinical notes of BioScope corpus are the radiology reports from the 2007 ICD9 challenge of the Cincinnati children hospital (Pestian, et al., 2007). NegEx has released a de-identified physician annotated test set of 2,376 sentences from 120 clinical reports. Also an instruction on how to produce an annotation guideline for biomedical corpus with negation layer is available (Morante, 2010).

In negation detection, rule based techniques have been shown to be effective and widely used in many NLP systems (Savova, et al., 2010) (Friedman, Hripcsak, Shagina, & Liu H, 1999). Rule based negation systems can be token-based (e.g., NegEx (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001), NegExpander (Aronow, Feng, & Croft , 1999), NegFinder (Mutalik, Deshpande, & Nadkarni, 2001), NegHunter (Gindl, Kaiser, & Mik, 2008;)) ontology-based (Elkin, et al., 2005), or utilize syntactic parsing results (e.g., DepNeg (Sohn, Wu, & Chute, 2012), ChartIndex (Huang & Lowe, 2007), Ballesteros et al (Ballesteros, Francisco, Díaz, Herrera, & Gervás, 2012)). For example, NegEx processes one sentence at a time by finding negation and termination terms. Termination terms are conjunctions such as "*but*" that end the scope of negation terms. There are three types of negation in NegEx algorithm, pseudo negation terms that are similar to negation terms but do not negate clinical conditions, pre-condition negation terms that appear before the clinical findings, and post-condition negation terms that appear after the clinical findings. If a pseudo negation term is found, NegEx skips to the next negation term in the sentence and uses corresponding regular expressions based on pre/post negation terms. NegEx has been extended into an algorithm called ConText in order to determine if a clinical condition of interest is hypothetical, historical or experienced by someone other than patient in addition to negation identification (Harkema, Dowling, Thornblade, & Chapman, 2009). Both NegEx and ConText have been translated into other languages (Afzal, Pons, Kang, Sturkenboom, & Schuemie, 2014) (Skeppstedt, 2010).

There are some attempts to incorporate syntactic parsing to improve the negation detection (Sohn, Wu, & Chute, 2012) (Ballesteros, Francisco, Díaz, Herrera, & Gervás, 2012). For example, DepNeg is a dependency parser-based negation algorithm that

utilizes the dependency structure of a target named entity in the sentence instead of a fixed negation scope (Sohn, Wu, & Chute, 2012). DepNeg uses manual negation rules based on the patterns of dependency paths between the focus (i.e., named entity) and the potential negation terms in the text that enables correctly identifying problematic negations in the traditional negation algorithm, such as NegEx. Similarly, Ballesteros et al used Minipar dependency parser to determine the scope of negation terms by traversing the dependency path from sentence's verb towards the end of the sentence. They could detect negation terms and their scope in clinical text of BioScope corpus with precision and recall of 0.958 and 0.906 respectively (Ballesteros, Francisco, Díaz, Herrera, & Gervás, 2012).

Machine learning has also been applied in negation detection. For instance, there are twenty-one systems developed for i2b2/VA assertion classification task where majority of them applied various machine learning algorithms including support vector machines (SVMs). The best system achieved 0.9326 micro-averaged F-measure using a 2-step approach. Where, in the first step, each word was represented as a feature vector consisting of n-gram, token category, and window of four tokens before and after the word, etc. and then a set of different classifiers were used to predict a score per class for each concept. In the second stage a multi-class SVM was used to predict the final assertion prediction for each token (de Bruijn, Cherry, Kiritchenko, Martin, & Zhu, 2011). Similar 2-step approach was applied to BioScope corpus by Diaz et al where each token in a sentence was classified as negation/speculation signal and a second classifier was used at a sentence level to determine the negation status of concept (Cruz Díaz, Maña López, Vázquez, & Álvarez V, 2012). Goldin and Champan compared Naïve Bayes and decision trees with default NegEx rule on 207 sentences of clinical records with negation "not". The default NegEx rule negates any UMLS concept within six-word window of "not." Naïve Bayes performed better than decision tree and baseline method with F-Measure of 0.90 (Goldin & Chapman, 2003).

Features used in machine learning algorithms may include results from rule-based systems as well as syntactic parsing results. For example, Grouin et al used SVM with NegEx and ConText dictionaries before or after a concept in a 5-word window (Grouin, et al., 2010). Wu et al (Wu, et al., 2014) also used SVM with following list of features, 1)

binary feature indicating if a given word appeared in a window size of 3,5 or 10 from the named entity 2) token in an exact distance from the named entity 3) negation terms 4) DepNeg dependency rules indicating whether a named entity is on the same dependency path as the negation word 5) constituency tree fragments to represent if a named entity is inside a phrase. They trained and test their system on four different corpora of SHARP NLP (Rea, et al., 2012), 2010 i2b2/VA, MiPACQ (Cairns, et al., 2011), and NegEx test sets and compared their system with YTEX (Garla, et al., 2011) implementation of NegEx algorithm. Their results were mixed and non conclusive, NegEx performed very well on NegEx test set (F-measure= 0.953) but the performance declined on other corpora with lowest F-measure of 0.623. Using a single versus all corpora for training the SVM has also generated mixed results that can be contributed to the diversity of their corpora.

As majority of the systems reviewed above are not publicly available, it is not feasible to compare various systems reported in the literature. Determining the scope of negation is a main challenge in most of rule based methods such as NegFinder that use a context free grammar parser especially when the distance between negation term and concept is more than a few words. For instance in the sentence "*Based on this, he required no operative intervention for his pseudocyst."* Because of the negation term *"*no*"* NegEx will consider the concept "pseudocyst" as negated while *"*no*"* is associated with *"*operative intervention*"* and not the *"*pseudocyst*"*. DepNeg attempts to remove this deficiency using dependency parser and shows promising preliminary results while using a limited set of rules on 159 Mayo clinical notes. DepNeg was compared with cTAKES adoption of NegEx, which is customized to Mayo Clinic data. cTAKES is an open source natural language processing tool for information extraction from medical records developed by Mayo Clinic and released under Apache license (Savova, et al., 2010). DepNeg focused on improving the precision of NegEx therefore it decreased the number of false positives in comparison to cTAKES negation (cTAKES negation -FP: 34, DepNeg-FP: 6) but increased the number of false negatives (cTAKES negation-FN: 47, DepNeg-FN: 61) (Sohn, Wu, & Chute, 2012).

There are two approaches of graph-based and transition-based in dependency parser. DepNeg uses ClearParser (Choi & Palmer, 2011), which is a graph-based dependency parser to determine whether the negation words are on the same path as

clinical concepts and therefore negated. Unlike DepNeg, we use a transition-based dependency parser to find if there is any dependency relation between negation words and concepts. And because NegEx had low number of false negatives (high recall) in our training set, we only applied the dependency parser to concepts that are considered negated by NegEx unlike DepNeg that applies dependency parser to all sentences containing negation tokens.

3.3 DEpEndency ParsEr Negation (DEEPEN)

We have developed an algorithm called DEpEndency ParsEr Negation (DEEPEN) that uses a chain of nested dependency relations between the clinical findings and negation terms. DEEPEN evaluate concepts that are considered negated by NegEx algorithm; so if a concept is considered affirmed by NegEx, no action is taken. Stanford dependency parser (SDP) (de Marneffe, MacCartney, & Manning, 2006) is applied to sentences containing the negated concept. SDP comprises of 53 grammatical relations (det: determiner, infmod: infinitival modifier, etc.) that will be generated for words within a sentence (de Marneffe & Manning, 2008). The SDP output consists of dependency relation, governor term and dependent term. Dependency relation is the grammatical relation between dependent term and governor term. Governor term is the word in the sentence that the dependency relation is reported for and dependent term is the word that is dependent of the governor term. For instance as shown in Figure 3.1, in the sentence "*Based on this, he required no operative intervention for his pseudocyst.*", *det(intervention-9, no-7)* *"*det*"* is the dependency relation, *"*intervention*"* is the governor term and *"no"* is the dependent term.

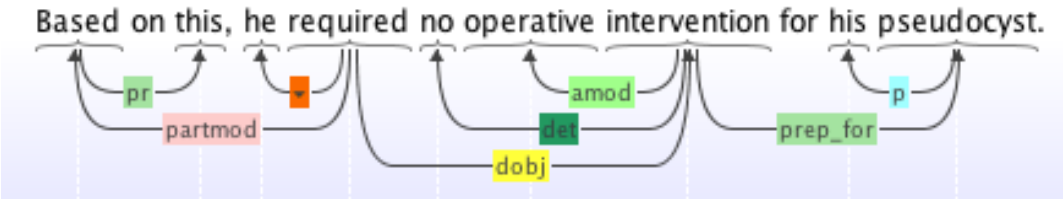


Figure 3.1 Dependency Relations Between Tokens in a Sentence

For every sentence with a concept that is considered negated by NegEx, a dependency chain is generated that is composed of three levels of token dependencies. First level token is governor of negation term, "*evidence*" in *det (evidence-2, No-1)*. Second level tokens are dependents of first level tokens, "*of*" in *prep (evidence-2, of-3)*. Third level tokens are dependents of second level tokens, "*dilatation*" *pobj (of-3, dilatation-6)*. Dependency chain is the concatenation of these three levels of token dependencies, "*evidence of dilatation*". If the concept is found in the dependency chain, it is negated otherwise it is affirmed. The concept "*pancreatic duct dilatation"* in the sentence "*No evidence of pancreatic duct dilatation or common bile duct stones*." is in the dependency chain, therefore it is negated. For concepts that are noun phrase such as "*pancreatic duct dilatation*", even if part of the noun phrase is in the dependency chain (dilatation), the concept is negated.

This basic rule fails in sentences with certain structures and therefore negated concepts are falsely identified as affirmed (i.e., false negative). We developed a set of rules to address the false negative results of applying DEEPEN on the IU training set.

In the previous chapter, we considered false positive and negative from a medical perspective (i.e., presence or absence of medical problem) meaning that if a patient has pancreatic cyst and the system considered that as a negative case (patient does not have pancreatic cyst), the result was evaluated as false negative. Similarly, if a patient does not have pancreatic cyst and the system result was affirmed (patient has pancreatic cyst), it was evaluated as false positive. In information retrieval focusing on negation status, however, we evaluate True positive—both system and the gold standard negates the term; True negative—both system and the gold standard does not negate the term; False positive—System negates the term but the gold standard does not negate the term; False negative—System does not negate the term but the gold standard negates the term.

DEEPEN was developed with the mindset of decreasing the number of false positives, nonetheless we attempted to decrease the number of false negatives by addressing most common sentence structures seen in our IU training data set. Figure 3.2, shows the flowchart of the algorithm used in development of DEEPEN.

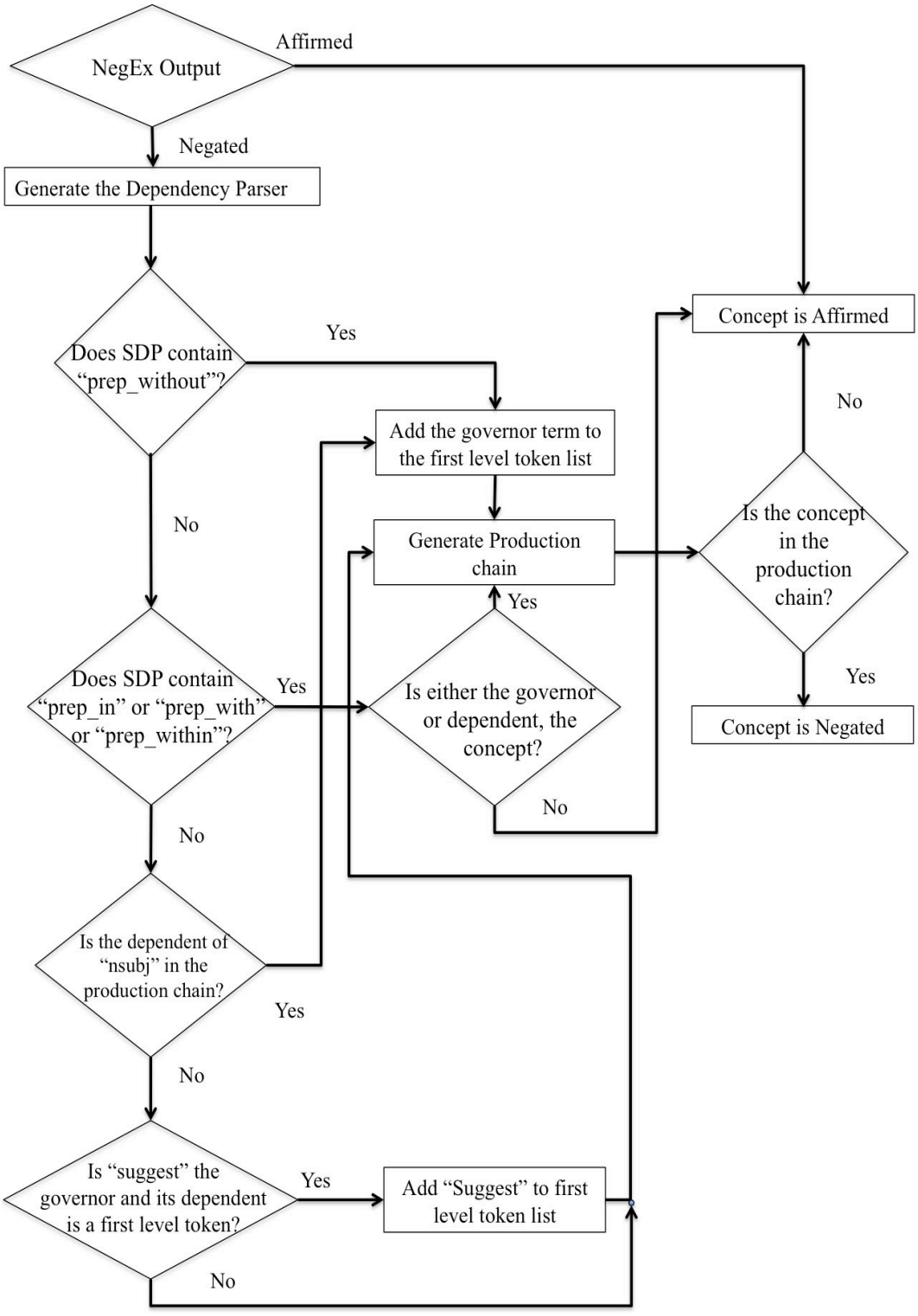


Figure 3.2 Detailed Flowchart of the DEEPEN Algorithm

Table 3.2 shows some examples of various rules developed in DEEPEN. More details and examples of DEEPEN rules are provided in the appendix A. DEEPEN is written in java and is freely available for researchers to use[1].

Conjunction And Rule: If there is a conjunction "*and*" in a sentence, it will be divided into two sub-sentences and negation is examined for both sub-sentences.

Preposition Within Rule: DEEPEN uses the collapsed representation of SDP where dependencies that involve propositions or conjunction are merged to create a direct dependency between content words. For instance, the dependencies involving prep (size-5, without-6) and pobj (without-6 inflammation-8) are collapsed into one single relation prep-without (size-5, inflammation-8). As we mentioned earlier first level token is the governor of negation term. In sentences where the negation term "*without*" is merged into the dependency relation, the governor of the relation "*prep-without*" is considered as first level token.

Preposition With/In/Within Rule: For propositions "*in*", "*within*", and "*with*" the SDP is only run when the concepts in these relations are part of the dependent or governor terms otherwise the concept is considered as "affirmed".

Nominal Subject Rule: Nominal subject in SDP is a relationship in which the subject is a noun phrase such as "*No abnormally*". If the governor of this relationship is a first level token then its dependent is added to the dependency chain.

Suggest Rule: in sentences that contain the term "*suggest*" if the dependent of the term "*suggest*" is a first level token then "*suggest*" will also be considered as a first level token.

[1] http://svn.code.sf.net/p/ohnlp/code/trunk/DEEPEN

Table 3.2 DEEPEN Rules with Relevant Sentence Examples and Their SDP Relations, Concepts are shown in Bold and Negation Terms in Italic (See Appendix A for Detailed Dependency Relations)

Rule	**Sentence**	**Relevant Dependency Relations** Dependency relation (governor token-index, dependent token-index)
Conjunction and	The main pancreatic duct does *not* appear disrupted and in continuity by a bridging **pseudocyst**	**pseudocyst** is affirmed in the sub-sentence "in continuity by a bridging **pseudocyst"** therefore SDP has not been applied.
Preposition without	The pancreas is normal size *without* perpancreatic inflammation or **pancreatic ductal dilatation**.	First level token: prep (size-5, *without*-6) Second Level tokens: prep_without (size-5, inflammation-8) nsubj (size-5, pancreas-2) cop (size-5, is-3) amod (size-5, normal-4) Third level tokens: det (pancreas-2, The-1) conj_or (inflammation-8, **dilatation**-12)
Preposition in, with, and within	An abdominal CT showed a normal pancreas and gallbladder with *no* **dilated ducts**.	First level token: det (**ducts**-5, *no*-3) Second Level tokens: amod (**ducts**-5, **dilated**-4)
Nominal Subject	No abnormally **dilated pancreatic duct.**	First level token: det (abnormally-2, *No*-1) nsubj (**dilated**-3, abnormally-2)
Suggest	*No* associated fluid collection to suggest **pseudocyst** or abscess.	First level token: det (collection-4, *No*-1) nsubj (suggest-6, collection-4) aux (suggest-6, to-5) dobj (suggest-6, **pseudocyst**-7) dobj (suggest-6, abscess-9) Second Level tokens: amod (collection-4, associated-2) nn (collection-4, fluid-3) Third level tokens: conj_or (**pseudocyst**-7, abscess-9)

These additional rules were added to the basic algorithm to decrease the number of incorrect assignment of present to concepts that were negated by NegEx.

We stopped the development of the algorithm as we reached acceptable precision and recall of 0.9839 and 0.9983 respectively on the training set and tested the final algorithm on the IU test set and Mayo Clinic dataset. Identified concepts and their negation status stored in the database were exported as spreadsheet to be reviewed by two domain experts independently at IU. The inter annotator agreement between the two reviewers was 95.6%. Any discrepancies regarding the negation status of a concept was discussed with the third medical expert by looking at the complete patient report. At Mayo Clinic, we used a gold-standard dataset that has been already annotated by four annotators, further details on annotation task and schema on this dataset can be found elsewhere (Ogren, Savova, & Chute, 2008).

3.4 Data Sources

This study was conducted under approved institutional review board at each institution.

3.4.1 Indiana University Dataset

The IU data was divided into two sets of training data of 664 patients consisting of 1136 reports with 1728 sentences with pancreatic cyst concept and test set of 452 patients with 793 reports and 1462 sentences.

3.4.2 Mayo Clinic Dataset

In order to evaluate the generalizability of our negation system, a set of 159 clinical notes with manual annotation of named entities and their negation status by four domain experts was used (Ogren, Savova, & Chute, 2008). There are total of 1,007 disorders with 426 unique UMLS concepts and 439 signs and symptoms with 129 unique UMLS concepts.

3.5 Evaluation

The system output was compared to the gold standard annotations to calculate the systems' precision, recall, and F-measure. Table 3.3 shows the relationship between the

system output and manually annotated sentences as defined by True positive—both system and the gold standard negates the term; True negative—both system and the gold standard does not negate the term; False positive—System negates the term but the gold standard does not negate the term; False negative—System does not negate the term but the gold standard negates the term.

Table 3.3 Comparison of the System's Result with Manually Annotated Sentences

		System Output	
		True (Negated)	**False (Affirmed)**
Gold Standard	**True (Negated)**	True Positive (TP)	False Negative (FN)
	False (Affirmed)	False Positive (FP)	True Negative (TN)

Performance of the system is measured by precision, recall, and F-Measure as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (3-1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3-2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3-3)$$

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Recall} + \text{Precision}} \quad (3-4)$$

3.6 Results

Table 3.4 shows the results of NegEx and DEEPEN applied to the IU and Mayo Clinic dataset. IU dataset contains 438 negated pancreatic cyst concepts (418 TPs + 20 FNs and 422 TPs + 16 FNs through NegEx and DEEPEN respectively) out of 1461 total concepts, which accounts for 30% of the data. Similarly 15.79% of disorders and 29.35% of sign and symptoms are negated in Mayo Clinic dataset. DEEPEN decreased the number of both false positives and false negatives when tested on IU dataset while it only decreased the number of false positive on Mayo Clinic dataset.

Table 3.4 Comparison of DEEPEN and NegEx Algorithm on IU and Mayo Clinic Dataset

IU Dataset	Pancreatic Cyst Concepts	Method	TP	TN	FN	FP	Precision	Recall	F-Measure	Accuracy
		NegEx	418	983	20	40	0.91	0.95	0.93	0.95
		DEEPEN	422	1008	16	15	0.96	0.96	0.96	0.97
Mayo Clinic Dataset	Disorders	NegEx	135	736	10	37	0.78	0.93	0.85	0.94
		DEEPEN	107	760	38	13	0.89	0.73	0.80	0.94
	Sign & Symptoms	NegEx	113	276	10	20	0.84	0.91	0.88	0.92
		DEEPEN	95	287	28	9	0.91	0.77	0.83	0.91

We also compared DEEPEN with DepNeg that uses dependency relations for negation detection. As the exact replication of the experiment reported in the DepNeg paper is not feasible, we compared DEEPEN's performance on the example sentences reported in the DepNeg paper. These sentences represent typical cases of DepNeg's capability of complicated negation detection as well as its limits. Table 3.5 shows the performance of three negation algorithms on the example sentences reported in the DepNeg paper.

Table 3.5 Comparison of DEEPEN, DepNeg, and NegEx, on Sentences Reported in the DepNeg Paper (the bold words in the sentence column denoted concepts that were examined for negation status; gray cells denote correct cases for each column).

Sentence	Negation Status			
	Gold Standard	**DEEPEN**	**DepNeg**	**NegEx**
He felt that no specific therapy was available regarding **Moebius sequence**.	Affirmed	Affirmed	Affirmed	Negated
I do not recommend drug treatment for **stone** prevention.	Affirmed	Affirmed	Affirmed	Negated
If her **pain** should not have been resolved by that time, there is the possibility of repeating facet rhizotomy.	Affirmed	Affirmed	Affirmed	Affirmed
However, I suspect that her **pain** is not due to an underlying neurologic disorder.	Affirmed	Affirmed	Affirmed	Affirmed
She denies any ear pain, sore throat, odynophagia, hemoptysis, shortness-of-breath, dyspnea on exertion, chest discomfort, anorexia, nausea, weight-loss, mass, adenopathy or **pain**.	Negated	Negated	Negated	Negated
Molecular fragile-X results reveal no apparent PMR-1 gene **abnormality**.	Negated	Affirmed	Affirmed	Negated
Mrs. Jane Doe returns with no complaints worrisome for recurrent or **metastatic oropharynx cancer**.	Negated	Affirmed	Affirmed	Negated
She is not having any incontinence or suggestion of **infection** at this time.	Negated	Affirmed	Affirmed	Negated
She denies any **blood in the stool.**	Negated	Negated	Affirmed	Negated

DEEPEN and DepNeg could correctly identify all affirmed concepts, while DEEPEN had one less false negative than DepNeg. NegEx, however, had higher number

of false positives than both DEEPEN and DepNeg while it had lower number of false negatives compared to DEEPEN and DepNeg. It should be noted that the major aim of DEEPEN and DepNeg is on having a high precision (i.e., reducing false positives).

3.7 Discussion

DEEPEN had higher precision and recall than NegEx on the IU dataset. However, when applied to the Mayo Clinic dataset, DEEPEN decreased false positives (i.e., higher precision) at the expense of increasing false negatives (i.e., lower recall), which resulted in lower F-measure than NegEx. This fact shows an interoperable issue on using heterogeneous data between institutions. NegEx uses a dictionary of negation terms that is not comprehensive. We added "lack of", "failed", "negative", "resolving" and "resolution" to NegEx's negation phrases dictionary based on observations in our training set to capture more negated concepts.

3.7.1 Error Analysis

In what follows, we discuss some of the reasons contributed to the increasing number of false negatives.

1) Errors due to sentence detection: Detecting the correct boundary of a sentence is a very important step in negation detection algorithm. Sentence detection in clinical notes is very challenging due to lack of end of sentence punctuation and random line breaks. Sentence detection can affect negation identification, for instance when "HOSP NO" and "Diagnosis*:* Pancreatic pseudocyst" in two lines were detected as one sentence the concept "pancreatic pseudocyst" is falsely considered negated because of the "NO" in "HOSP NO" that matches "no" in NegEx's negation terms. Also when multiple lines of text are considered as one sentence, dependency parser fails to correctly identify the relation between tokens in the sentence containing the concept and therefore the final negation detection result is compromised.

2) Errors due to variations in the two institutions' corpora:

DEEPEN was developed focusing on a single concept within the IU dataset although it performed well on Mayo Clinic dataset by decreasing the number of false positive in comparison with NegEx it could not maintain the same performance consistency as tested on IU data. One of the major sentence structures in the Mayo Clinic false negatives were

sentences with a negation word followed by multiple concepts separated with "comma" and "or" such as "*No associated shortness-of-breath, nausea, vomiting, diaphoresis, or light-headedness.*". All five concepts within this sentence are falsely considered affirmed by DEEPEN. More than 20 of the false negatives in sign and symptoms and 12 of false negatives in the disorders from Mayo dataset had the same structure.

3) Conditions developed previously

Sentences that mention a condition that was previously developed in a patient but are not considered a current medical problem could be very complex and require deep contextual analysis. Following is example of two such sentences A and B from Mayo clinic and IU datasets respectively.

A) "*Mr. X is doing very well from the standpoint of his **sarcoma** with no evidence of recurrent disease on physical examination."*

*B)"No lesion seen at the prior site of the mid pancreatic body lesion, which was previously to represent a **pseudocyst**."*

Based on dependency relations, "*sarcoma*" and negation word "*no*" are not related in sentence A, however it can be inferred from the context that the concept is considered as a history and therefore negated. Likewise in sentence B, the concept "*pseudocyst*" is affirmed by DEEPEN because there is no relation between negation term "No" and the concept "*pseudocyst*", however previously seen pseudocyst does not mean that the patient currently has pseudocyst.

3.7.2 Limitations

As DEEPEN does not address the present (i.e., affirmed) concepts by NegEx. The number of concepts considered incorrectly present by DEEPEN are inherited from NegEx or due to incorrect dependency relations of SDP parsing. SDP has been created using the corpus of English web Treebank that consists of sentences from weblogs, newsgroups, etc. Therefore its performance would be lower on clinical texts that lack proper grammatical structure in comparison to general English in news and weblogs.

3.7.3 Future Work

We are planning to address the false negative cases in Mayo Clinic dataset and also address the concepts that are affirmed by NegEx in the next release version of DEEPEN.

3.8 Conclusion

In this chapter, we described one of the challenges in contextual detection, detecting the negation status of a concept in a sentence. DEEPEN used a nested dependency relation to find out the relation between negation words and concepts to decrease the number of falsely negated concepts (i.e. false positives). It could effectively decrease the number of false positives in both the IU and Mayo Clinic dataset in comparison with NegEx. DEEPEN shared the idea of using a dependency parser with DepNeg to find out the relation between negation words and concepts. Our approach is different from DepNeg in: 1) DepNeg does not use NegEx to find the negation status of concepts and 2) DepNeg uses rules to find out if concepts and negation words are on the same dependency path. However, DEEPEN is built on top of NegEx and only uses dependency relation rules for concepts that are negated by NegEx. The comparison of DEEPEN with DepNeg on example sentences reported in DepNeg paper showed the capability of DEEPEN in correctly identifying negation status of complicated cases.

In the next chapter, we will describe our family history detection, another important contextual information extraction.

CHAPTER FOUR: FAMILY HISTORY

4.1 Introduction

It has been shown that a wide range of adult conditions such as diabetes, cardiovascular diseases, Alzheimer's and cancers have hereditary roots (Wilson, et al., 2009). Accurate family history information can be very helpful in precision medicine that tailors the treatment to the individual characteristics of patients. For instance, the risk of having colon cancer for individuals with family history of colon cancer is two fold, which makes individuals with positive family history of colon cancer the best candidates for genetic testing and preventive screening (Yoon, Scheuner, Peterson-Oehlke, Gwinn , Faucett , & Khoury, 2002) (Behnam, Waterman, & Smith, 2013). The family history information can be available in clinical notes by "documenting parents' and siblings' age and health (or age and cause of death), as well as a checklist of conditions with environmental and hereditary etiologies" (Degowin & Degowin, 1969).

Information extraction (IE) attempts to structure and encode the information buried in free text clinical notes. Statistical machine learning and rule-based approaches have been used in the development of IE techniques. Machine learning approaches require annotated training examples and lacks portability. Rule-based approaches on the other hand perform very well when a task involves a specific subdomain or a limited number of named entities (Liu, et al., 2013). Although, rule-based approaches are cumbersome to implement, they have been widely used in clinical NLP. In this chapter we developed a rule-based method to identify patients with family history of pancreatic cancer. In order to evaluate the generalizability of the algorithm, it was evaluated on a different institution's records that it was originally developed.

4.2 Related Works

Family history identification consists of various steps including section segmentation, relation discovery between family members and diagnosis, and negation detection. Automatic identification of section headers in clinical notes is an important preprocessing step in the family history extraction. Argumentative zoning is a closely related task that attempts to classify each sentence of a scientific article into one of seven

sections of “background”, “other” (other researchers’ work), “own” (author’s work), “aim”, “textual” (textual organization of the paper), “contrast” (other’s work weaknesses) and “basis” (authors’ work based on others’ work) (Teufel & Moens, 2002). Sequential tagging approaches such as Naïve Bays (NB) and maximum entropy (MaxEnt) model have been used in solving this problem. MaxEnt model of Merity et al., achieved 96.88% F-Score (Merity, Murphy, & Curran, 2009). Another closely related task is classification of sentence in abstracts of scientific articles into sections such as introduction, methods, results and conclusion. Machine learning algorithms such as SVM (McKnight & Srinivasan, 2003) (Rastegar-Mojarad, Boyce, & Prasad, 2013), Hidden Markov Model (HMM) (Lin, Karakos, Demner-Fushman, & Khudanpur, 2006) (Rastegar-Mojarad, 2013), and Conditional Random Field (CRF) (Hirohata, Okazaki, Ananiadou, & Ishizuka, 2008)have been used with performances ranging from 90-94.3% accuracy. CRF has several advantages over widely used probabilistic models such as HMM, Maximum Entropy Markov models (MEMM) and stochastic grammars in labeling sequential data. Lafferty et al compared the performance of CRF to HMM and MEMM models on synthetic and natural language data (Lafferty, McCallum, & Pereira, 2001). CRFs (Lafferty, McCallum, & Pereira, 2001) have been successfully applied to medication event extraction (Li, Liu, Antieau, Cao, & Yu, 2010), named entity recognition (Leaman & Gonzalez , 2008), information extraction (Fuchun & McCallum, 2006) and event causality identification (Fu, Liu, Liu , & Guo, 2011).

In clinical domain, researchers at university of Vanderbilt developed an algorithm called SecTag that uses a combination of NLP techniques, rules based and naïve Bayesian scoring methods to identify note section headers (Denny, Spickard 3rd, Johnson, Peterson, Peterson, & Miller, 2009). Section header terminology was developed using Quick Medical Reference (QMR) knowledge base and Logical Observation Identifiers Names and Codes (LOINC) and various other resources with data model similar to UMLS (Denny, Miller, Johnson, & Spickard 3rd, 2008). Similar to argumentative zoning sequential tagging algorithms have also been used in clinical section segmentation. Li et al., used HMM to label sections in clinical notes to one of 15 possible known section types achieving per section accuracy of 93% and per note accuracy of 70% (Li, Gorman, & Elhadad, 2010). Tepper et al, used two methods, one-

step approach that segment and classify sections in one step and two-step approach that uses two different models for section segmentation and classification. In one-step approach they used MaxEnt sequential tagging model to identify if a line begins, inside or outside a section category. In two-step approach they used MaxEnt sequential tagging to first label each line with begin, inside and outside tags and then a separate classification algorithm was used to label each section with appropriate section categories. The two-step approach outperformed the one-step approach with precision/recall/F-measure of 90.0-97/90.4-96.7/89-96.8 (%) on three different datasets (Tepper, Capurro, Xia, Vanderwende, & Yetisgen-Yildiz, 2012).

Once a family history section is identified and sentences within this section are parsed, the next step is to associate the diagnosis with the correct family members. Both rule-based and dependency parsers have been used to associate family members with diagnoses concepts. Goryachev et al. developed a rule-based algorithm using tokens such as “comma”, “and”, “dot”, “patient has”, “patient had” to assign diagnosis concepts to family members (Goryachev, Kim, & Zeng-Treitler, 2008). Their method achieved higher precision and recall in comparison to a dependency parser based algorithm used in another study (Lewis, Gruhl , & Yang, 2011).

4.3 Methods

Clinical reports are organized into sections with headers such as “Physical examination,” “Medication,” “family history,” etc. Usually patient’s family history is reported under the family history section of the narrative reports. Classifying clinical texts into sections can be helpful in family history extraction. However family history does not always appear under the family history section. It is sometimes mentioned with the patient’s history, diagnosis or other sections of the report. Based on this understanding, we divided the family history identification problem into two parts. In the first part, the patients’ family history, which is reported under family history section, were identified. In the second part, the family history section is removed from the clinical note and any mentions of family history in other sections were identified. The first part consists of three sub-parts: 1) section header detection, 2) family member and diagnosis identification, and 3) relation discovery between family member and diagnosis.

4.3.1 Section Header Detection

Clinicians typically use a template with pre-defined sections to write their observations but they can freely modify these sections. Furthermore there is no fixed terminology for section headers; therefore accurate identification of section headers is a challenging task.

We used Conditional Random Field (CRF), which is a sequential tagging algorithm to identify the section boundaries in clinical notes. CRFs are undirected graphical models that model the conditional distribution p (x | y) rather than joint probability distribution p (y, x) and trained to maximize the conditional probability of outputs given the inputs (Ye, Sun, Chieu, & Wu, 2009) (Swaminathan, et al., 2013).

A probability distribution of p (x, y), over a set of random variables V=x ∪ y, can be represented by a product of distributions that represent a smaller set of the full variable set (Sutton & McCallum, 2011).

$$P(x, y) = \frac{1}{z} \prod_{a \in F} \Phi_a(x_a, y_a) \quad (4\text{-}1)$$

Where, a is a subset of V (F=a ⊆ V), x=<x_1, x_2,…, x_n> is the set of input variables for instance a sequence of tokens and y=<y_1, y_2,.., y_n> is a set of output variables that can be BIO tags with section label X (B-X, I-X & O-X indicate the begin, inside and outside of a section with category X respectively). And Z in equation (3-2) is a constant that normalize distribution (3-1) to one.

$$Z = \sum_{x,y} \prod_{a \in F} \Phi(x_a, y_a) \quad (4\text{-}2)$$

$\Phi(x_a, y_a)$ can be written as k feature functions where λ is the learned weights for each feature function.

$$\Phi(x_a, y_a) = \exp\{\sum_K \lambda_{ak} f_{ak}\ (x_a, y_a)\} \quad (4\text{-}3)$$

The weights will be learned in a training procedure to positively reinforce the feature functions that are correlated with the output labels or weaken the feature functions that are not correlated with the output labels, weights for uninformative feature functions will have a zero value or neutral effect.

Reports have section relevant to their type, for instance “impression” and “indication” are section headers that only exist in radiology reports. Therefore, the training set of 400 clinical notes was randomly selected such that at least one report from

every type of report was included in the dataset. A guideline on annotation of clinical reports was created and reports were manually annotated with BIO (begin, inside and outside of a section) using knowtator (Ogren, 2006), tokens other than the section header were tagged as "out". Two annotators independently annotated the data and a third reviewer resolved any discrepancies. The inter annotator agreement between the two reviewers was 95.9%.

Table 4.1 and equations 3-4 to 3-9 show the calculation of the inter annotator agreement.

Table 4.1 Inter-Annotator Agreement for the Manual CRF Training Data Annotation

2nd Annotator / 1st Annotator	B	I	O	Total
B	1976	0	54	2030
I	0	2124	78	2202
O	51	149	87105	87305
Overall Total	2027	2273	87237	91537

$$\Sigma a = 1976+2124+87105=91205 \quad (4\text{-}4)$$

$$Ef_1=\frac{2030*2027}{91537} = 44.95 \quad (4\text{-}5)$$

$$Ef_2=\frac{2202*2273}{91537}=56.67 \quad (4\text{-}6)$$

$$Ef_3=\frac{87305*87237}{91537}=83203.8 \quad (4\text{-}7)$$

$$\Sigma ef = 83303.42 \quad (4\text{-}8)$$

$$K=\frac{91205-83303.42}{91537-83303.42}=\frac{7901.58}{8234}=95.9 \quad (4\text{-}9)$$

The Mallet implementation of the CRF algorithm was used in this study (McCallum, 2002). Following is the list of features used to train and test the CRF model:

- CRF Model Features

Features are inputs to the CRF model and the outputs are a sequence of "begin", "inside" and "out" tags. In the training set, we supply the CRF model with feature inputs (tokens, token categories, prefix and suffixes, POS, shallow parser) and known outputs (begin, inside and out) so that the CRF model could learn the pattern of data by adjusting

the weights of its feature function. The learned model was then used on an unseen data (test set) to predict its sequence of tags.

- Tokens and Prefixes and suffixes

Tokens and their prefix and suffix are used as features. Words such as family history, physical exam, etc. might be section headers and are important features to be considered as inputs. Also tokens' prefix and suffix were considered with the maximum length of 4 words.

- Token category

Token category can determine if a token is a section header or not. For instance tokens that starts with an uppercase letter are more probable of being a section header than tokens that contain only lowercase letters. Tokens are categorized based on their character. Table 4.2, obtained from LingPipe API documentation (Alias-i, 2008), shows the list of categories such as all upper case letters, all lower case letters, mixed of digits and letters, *n* number of digits, an uppercase letter followed by lowercase letters, punctuations, etc.

Table 4.2 List of Token Categories (Alias-i, 2008)

Category	Description
NULL-TOK	Zero-Length string.
1-DIG	A single digit.
2-DIG	A two-digit string.
3-DIG	A three-digit string.
4-DIG	A four-digit string.
5+-DIG	String of all digits five or more digits long.
DIG-LET	Contains digits and letters.
DIG--	Contains digits and hyphens.
DIG-/	Contains digits and slashes.
DIG-,	Contains digits and commas.
DIG-.	Contains digits and periods.
1-LET-UP	A single uppercase letter.
1-LET-LOW	One lowercase letter.
LET-UP	Uppercase letters only.
LET-LOW	Lowercase letters only.
LET-CAP	Uppercase letters followed by one or more lowercase letters.
LET-MIX	Letters only, containing both uppercase and lowercase.
PUNC-	A sequence of punctuation characters.
OTHER	Anything else.

- Part of speech

We used MedPost POS tagger for tagging the sentences (Smith, Rindflesch, & Wilbur, 2004). Section headers are more likely to appear as proper nouns and adjectives than verb or determinants. Table 4.3, shows the POS tags generated by the system for every token in the sentence "PROCEDURES DURING HOSPITALIZATION: The patient underwent…".

- Shallow parser

We used Apache OpenNLP (The Apache Software Foundation) phrase chunker to tag syntactical phrases in BIO (begin, inside and outside) sequence. Section headers are

normally consist of noun phrases rather than propositional or verb phrases. As it can be seen in Table 4.3, the section header (PROCEDURES DURING HOSPITALIZATION) is a noun phrase (NP) and tagged with BNP (begin NP) and INP (inside NP).

Table 4.3 Tokens of a Sentence with Their POS, Shallow Parser Generated by the System and Manually Tagged as BIO

Tokens	Procedures	During	Hospitalization	:	The	Patient	Underwent
POS	NN	NN	NN	:	DD	NN	VVD
Phrase Chunks	B-NP	I-NP	I-NP	O	B-NP	I-NP	B-VP
BIO tags	B	I	I	O	O	O	O

4.3.2 Family member and diagnosis identification

After family history section was identified, sentences reported under this section were detected using Ytex sentence detector (Garla, et al., 2011). A list of keywords indicating pancreatic cancer concepts, UMLS semantic type T099 for family group (Bodenreider, 2004) and manual review of clinical notes were used to assemble a list of family member and diagnosis keywords. This dictionary was then used to identify family member and pancreatic cancer concept within a sentence.

4.3.3 Relation between the family member and diagnosis

Associating family member with pancreatic cancer in a sentence with only one family member is trivial (i.e. Sentence A).

*A) "Notable for a **father** with what sounds like cirrhosis, colorectal cancer, as well as **pancreatic cancer**, and alcohol abuse."*

However for sentences with more than one family member, this task is challenging (i.e. Sentence B).

*B) "The only cancers in her **family** include a first **cousin** on her **mother's** side with breast cancer in her xxx, as well as a **paternal aunt** who had **pancreas cancer** in her xxx, and her **brother** who died of **pancreas cancer** at the age of xxx."*

We developed a set of rules that divides the sentence into sub-sentences based on tokens such as ",", ";" or ", and" and associate family member and disease in each sub-sentence.

For example in sentence "B" after dividing the sentence to three sub-sentences, we could link "paternal aunt" and "pancreas cancer" in the sub-sentence *" as well as a paternal aunt who had pancreas cancer in her xxx"* and "brother" and "pancreas cancer" in the sub-sentences "and her brother who died of pancreas cancer at the age of xxx".

If the pancreatic cancer concept were found with no family members in sentences under family history section, the general term "family history" was assigned to the concept.

In order to identify family history of pancreatic cancer that are mentioned in other section of the report other than family history, the family history section was removed from the report and the same algorithm was applied with the exception that at least one family member should be present in the sentence.

An NLP system using UIMA framework shown in Figure 4.1, was developed to accommodate the above steps. First two blocks in the UIMA pipeline are report separator and metadata annotator that extract each report's main body and its metadata information such as report name, ID, date and patient medical record number. Reports' main body was then used as an input to the next block of code where family history sections were detected. After the family history section was extracted, the section was split into sentences and family member and diagnosis were identified. We used our previously developed negation algorithm called DEpEndency ParsEr Negation (DEEPEN) to find out the negation status of diagnosis concepts in a sentence (Mehrabi, et al., 2015). DEEPEN improves the NegEx algorithm by double-checking the negation status of concepts using a nested chain of dependency relations between negation word and desired concepts within a sentence. And finally all the extracted information including patient medical number, report name, report date, the sentence containing the concept, the diagnosis concept and related family members found in the sentence, and their negation status were stored in a database.

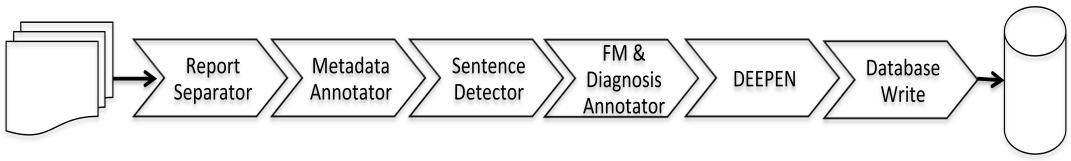


Figure 4.1 Analysis Engines Developed in the UIMA Pipeline to Identify Patients with FH of Pancreatic Cancer

4.4 Dataset

4.4.1 Indiana University (IU)

Clinical notes of patients who visited Sidney and Lois Eskenazi hospital in Indianapolis during March-December 2013 were used in this study. On average 7,270 patients visited the hospital each month with a range of 80 to 95 thousands of reports for all patients during that month. The dataset was randomly divided into 60% for training and 40% for testing.

4.4.2 Mayo Clinic

To investigate the portability of the family history extraction rules, we used Mayo cancer registry data to obtain a list of patients with pancreatic cancer. There were a total of 3573 patients in the registry, out of which 2923 had a family history section in their clinical notes. Clinical notes for those patients were extracted from Mayo clinic data repository and text from the family history section of those notes forms the data set.

4.5 Results

Table 4.4 shows the performance of the system on the IU training and testing sets. The system output consists of patient medical record number, sentence, diagnosis, family member and negation. The results were evaluated as correct or incorrect by two independent reviewers with inter annotator agreement of 95.9%. A result is correct if pancreatic cancer is associated with the correct family member and negation status of the diagnosis was identified accurately. Any errors in these finding were considered as an incorrect instance. We also considered hypothetical cases (i.e. a sister may have had pancreatic cancer.) as incorrect. If pancreatic cancer related to patient or his non-blood

relative (i.e. wife or husband) was mentioned in the family history section, it was considered as irrelevant.

Table 4.4 IU Dataset Evaluation

Train	Correct	Incorrect	Irrelevant	Precision
Affirmed	22	7	2	75.9
Test Set	Correct	InCorrect	Irrelevant	
Affirmed	14	2	2	88.9
Negated	2	0	0	100

We applied the same algorithm to the Mayo clinic dataset without any modifications (Table 4.5). Precision is defined as the number of correct instances over total of correct and incorrect instances. As it can be seen the performance of the system has been consistent across the two institutions.

Table 4.5 Mayo Clinic Dataset Evaluation

	Correct	InCorrect	Irrelevant	Precision
Affirmed	519	72	32	87.8
Negated	438	4	2	99.1

In order to make sure that we did not miss any patient with family history of pancreatic cancer, 100 reports were selected randomly and manually reviewed. Table 4.6 shows the result of our modified algorithm to incorporate missing patterns in these 100 reports.

Table 4.6 Results of Mayo Clinic Dataset Evaluation after System Customization

	Correct	Incorrect	Irrelevant	Precision
Affirmed	550	74	34	88.1
Negated	443	4	2	99.1

Another batch of 100 reports were randomly selected from Mayo dataset excluding the first 100 reports to manually review the family history of pancreatic cancer. There was no missing pattern in the second set of randomly selected reports.

In relation discovery evaluation, true positives were considered as instances where the pancreatic cancer concept was assigned to the correct family member in the sentence. False negatives were any family member relation that was missed by the system. A wrong family member assignment was considered as a false positive.

Table 4.7 Results of Family Member Identification

	True Positive	False Positive	False Negative
Family Member relation discovery	579	190	53
	Precision	Recall	F-Measure
	75.3	91.6	82.6

There were total of 268 patients with a family history of pancreatic cancer out of 3573 patients with pancreatic cancer in Mayo Clinic's data set. Table 4.7 and Figure 4.2 show the number of patients identified with first, second or third degree relative.

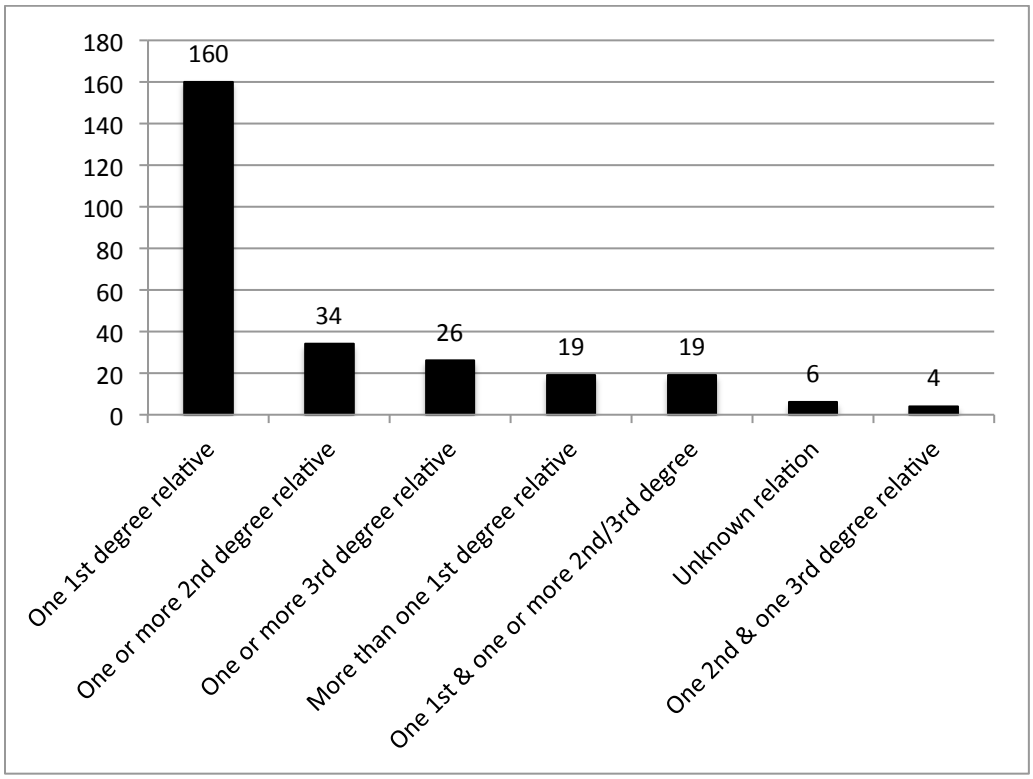


Figure 4.2 Number of Identified Patients with one or more 1st, 2ed or 3rd Degree Relative

4.6. Discussion

We have developed our system on IU dataset. IU dataset consisted of any patient who visited the Eskenazi hospital during 10 months for any reason. Due to low incidence of pancreatic cancer with a familial basis, we had a very few number of patients compared to Mayo Clinic dataset. Clinical notes at the Mayo Clinic are CDA 1.0 compliant; therefore section detection developed at IU dataset was not used for Mayo Clinic dataset. We also did not consider the family history mentions in other sections of clinical notes other than family history in Mayo Clinic dataset.

4.6.1 Error Analysis

We can classify the errors in our system based on the following reasons:

1) Complicated family relations

Sentences "*C*", "*D*", and "*E*" show examples of family relation were multiple family member terms were used to show the relation. As we did not have these complicated instances of relationship in our dictionary set, our system related each family member term to the pancreatic cancer separately. For instance in sentence "*C*" pancreatic cancer was related to mother, sister and granddaughter.

Sentence "*F*" shows an example where semantic inference is needed to infer that pancreatic cancer is related to mother.

C) "recently she found out about her mother's sister's granddaughter who was diagnosed with pancreatic cancer at the age of xxx."

D) "he had an uncle that was actually a half-sibling to his mother that died of pancreatic cancer."

E) "he had one cousin on the patient's father's side of the family (the cousin was the son of the patient's father's brother) who had pancreas cancer at age xxx."

F) "her son (our patient) found her deceased about x p.m. a postmortem examination showed cause of death was due to multiple blood clots and she was found to have a widespread pancreatic cancer."

2) System Failure

As mentioned in the relation discovery section, a set of rules was developed to divide the sentence into sub-sentences. When there are multiple family relation terms in a sentence such as sentence "G". Each family relation term is then associated with the pancreatic cancer concept within the sub-sentence. In sentence "G", "*pancreatic cancer*" is associated with "*paternal grandfather*" but it failed to associate the "mother" and "father" with pancreatic cancer concept in the sub-sentence *"his mother, father,"* because there is no concept in the sub-sentence.

G) "his mother, father, and paternal grandfather died from pancreatic cancer."

There were few instances where co-referencing was needed to extract the right family relation (i.e. sentence "*H*"). Our current system does not handle co-referencing.

H) "She has one son living and one deceased. The one that is living has a recent diagnosis of pancreatic cancer, and three daughters."

4.6.2 Future Work

The future steps involves, refinement of the family relation discovery rules specially improvement of the sentence detection algorithm. A risk stratification method will also be developed based on the number and degree of family relations to assess patient's risk of having cancer and a surveillance strategy will be designed to follow up with patients according to their risk.

4.7 Conclusion

We have developed a rule-based algorithm to identify patients with family history of pancreatic cancer retrospectively from their clinical records (Mehrabi , et al., 2015). Development of clinical NLP system requires resources such as domain experts to develop guidelines, nurse abstractor to create gold standards and researchers/programmer to develop and analyze the system. Although rule-based method highly depend on the natural language that they have been developed on, this study shows that as long as the rules are kept simple and generalizable, we can transfer an algorithm developed in one institution to other institutions. Positive family history is the basis for the diagnosis of many familial conditions and with the increase of precision medicine practice, it can not only be a great source of information for diagnosis and screening but also can be utilized for targeted treatment. And therefore adequate family history information including detailed information such as affected family members and their social history, age of disease onset, and specific information regarding the disease in question are crucial.

So far, we described the various NLP methods to correctly identify concepts, their negation status and weather it is associated with patient or his family member. Although we extracted the encounter time along with the clinical entities from longitudinal records of patient, the temporal order of these events were not considered. Temporal analysis of clinical data of patients can help in discovering patterns that represent phenotypic path of disease progression. For instance, many pancreatic cysts demonstrate pre-malignant behavior, and some will ultimately progress to pancreatic cancer. Pancreatic lesions that may harbor invasive cancer depend on patient symptoms/signs/conditions, radiographic features, cytology, pancreatic fluid, and serum marker analysis. Identifying the common phenotypic features in patients with pancreatic cyst who developed pancreatic cancer

across their longitudinal health records is crucial in early pancreatic cancer detection. Unfortunately there were very few patients with pancreatic cyst in our dataset that developed pancreatic cancer; therefore we used a different dataset to test our temporal pattern discovery method that is described in the next chapter.

CHAPTER FIVE: TEMPORAL PATTERN DISCOVERY

5.1 Introductions and Background

Longitudinal patients' EHR consist of disease characteristics, its treatment and outcome. The analysis of such clinical information against temporal dimensions provides valuable information in clinical decision-making that includes phenotyping (Hripcsak & Albers, 2013) and early diagnosis (Jakkula, Crandall, & Cook, 2009). It also facilitates discovering novel patterns in the disease progression based on the knowledge acquired from similar patients (Jensen, et al., 2014). Temporal pattern discovery aims at finding temporal patterns among one or more groups of patients. Pattern discovery is an active research in many domains including image processing, signal processing, video content analysis, etc. (Wang, Zhao, & Yuan, 2013).

Signal processing methods have been extensively used in studying the ECG or EEG data to extract interesting patterns and classifying data into meaningful labels for clinicians (Gacek & Pedrycz, 2012). However, clinical events that are recorded in irregular time interval in patients' medical records are more challenging to process. Domain expert interpretation is needed to define a common temporal interval and normalize all irregular intervals into the common selected interval. One approach would be to transform the irregularly timed observations into time series format and analyse the time series data. Lasko et al used Gaussian process regression to transform the irregularly timed stamped uric acid measurements in patients with gout or acute leukemia, into a continuous longitudinal probability density before applying the deep learning algorithm (Lasko, Denny, & Levy, 2013). Perotte and Hirpcsak used kernel density to estimate distribution of ICD9 dignosis codes across cohort of patietns. The density estimate of positive mentions of a diagnosis was divided by all mentions of the diagnosis to estimate the probability of a specific condition at a given time after its first docuemntation (Perotte & Hripcsak, 2013). Also Lomb-Scargle periodgrams, a frequency spectrum estimation model that is based on a least square fit of sinusoid was used in ICD9 coded records of patients to discover seasonally linked diseases (Melamed, Khiabanian, & Rabadan, 2014).

Temporal abstraction is one of the most common approaches in studying the temporal pattern of unevenly timed-sampled clinical observation (Shahar & Musen, 1993) (Shahar & Musen, 1996) Temporal abstraction transfers time-stamped clinical events to interval-based representation so that temporal data mining methods can be applied (Batal, Fradkin, Harrison, Moerchen, & Hauskrecht, 2012) (Patnaik, Butler, Ramakrishnan, Parida, Keller, & David, 2011) (Sacchi, Larizza, Combi, & Bellazzi, 2007) (Batal, Valizadegan, Cooper, & Hauskrecht, 2013) KarmaLego is an algorithm for fast mining of temporal interval patterns. It is based on Allen's seven relations with addition of an epsilon value to all seven relations (Moskovitch & Shahar , 2009) (Moskovitch & Shahar, 2013) (Moskovitch, Walsh, Hripcsak, & Taton, 2014). ChronoMiner is an ontology driven temporal mining system that dynamically extracts temporal association at various hierarchical levels (Raj, O'Connor , & Das , 2007). It was applied to a data set of patients with HIV to find the association of new mutation corresponding to the related administered therapy. Temporal data mining techniques were also applied to administrative data in health care. Noren et al presented a pattern discovery method based on statistical and graphical approach to mine the association between medication prescription and clinical events stored in clinical administrative database (Norén, Hopstadius, Bate, Star, & Edwards, 2010). Hybrid of time stamped and interval-based representations were also used in mining the temporal association rules to find relationship between drug prescriptions and clinical conditions of diabetic patients (Concaro, Sacchi, Cerra, & Bellazzi, 2009) (Nabavizadeh, Greenleaf, Fatemi, & Urban, 2014).

Wang et al in a distinctive approach from the above represented each patients record as an image or event matrix where *y*-axis corresponds to clinical features such as symptoms, lab values and radiological features etc. and *x*-axis corresponds to the time they were recorded in longitudinal patient records. They proposed a convolutional nonnegative matrix factorization based framework to discover patterns of synthetic and real data in patients with diabetes (Wang, Lee, Hu, Sun, & Ebadollahi, 2012) (Wang, Lee, Hu, Ebadollahi, & Laine, 2013).

Various applications such as SAX (Lin, Keogh, Lonardi, & Chiu, 2003), PatientFinder (Plaisant, et al., 2008) and LifeLines (Plaisant, Mushlin, Snyder, Li, Heller, & Shneiderman, 1998) have been developed to visualize and cluster temporal patterns.

SAX introduced a symbolic representation of time series with dimensionality reduction where data mining algorithms can be efficiently applied to symbolic representation without any information loss comparing to the original dimension. LifeLine provides hierarchical timeline visualization organizing visits, lab tests and medications for a single patient and doesn't have any mechanism for temporal representation of facts across multiple patients' records. PatientFinder offers graphical visualization with ability of temporal query on Microsoft Amalga EHR. However, it requires the user to specify what patterns to look for in the data.

5.1.1 Deep Learning

Recent advances in GPU architecture and computer vision is one of the major reasons for the resurgence of deep learning neural networks. These architectures have been very popular in image and signal processing with significant improvement in error reduction. For instance Microsoft reported the error reduction of 23% in the GMM-HMM system to 13% in deep learning system on the switchboard automatic speech recognition task (Deng , et al., 2013). Back propagation invented in 1980's has been a well-known algorithm for learning weights of feed forward networks. Step Hochreiter has shown that back-propagation algorithm is too slow for practical use because of vanishing gradient problem (Hochreiter, 1998). Therefore simpler methods such as support vector machines dominated the field of machine learning during 1990s and 2000s. In 2006 Geoffrey Hinton introduced the idea of unsupervised pre-training of each layer of deep architectures (Hinton, Osindero, & Teh, 2006). This recent development of deep learning algorithms created a new wave of interest in unsupervised learning.

5.1.2 ICD9 and HCUP CSS Diagnosis Codes

ICD9-CM consists of more than 14,000 diagnostic codes with fine granularity and details. The Agency for Healthcare Research and Quality (AHRQ) developed a collection of databases and related software tools through Healthcare Cost and Utilization Project (HCUP) that enabled research on a broad range of topics including cost and quality of healthcare services, treatment outcome, medical practice patterns, etc. (Agency for Healthcare Research and Quality, 2014). The clinical classification software (CCS) in HCUP classifies ICD9-CM diagnosis codes and Current Procedural Terminology (CPT)

codes into more manageable and clinically meaningful categories (Elixhauser, Steiner, & Palmer, 2014) (Agency for Healthcare Research and Quality). The single-level HCUP codes clusters ICD9-CM codes into 280 groups and multi-level HCUP CSS codes further group the single-level codes into 18 main groups.

Previously we explored topic modeling for discovery of associations of diagnosis codes (Li, Thermeau, Chute, & Liu, 2014). However it did not take into account the temporal relationship between diagnosis codes. In this study we explored the deep learning technique to discover temporal pattern among diagnosis codes.

5.2 Rochester Epidemiology Project

The Rochester Epidemiology Project (REP) is a research infrastructure, linking together the medical records of the residents of Olmsted County, Minnesota and has supported various population based analytic studies of disease and outcome. The REP manages a dynamic cohort of 502,820 unique patients who lived in Olmsted County at some point during 1966 to 2010 and received healthcare from one of the 50 participating health care providers. The REP links together the longitudinal medical records of patients who contributed a total of 6,239,353 person-years of follow-up. The REP provides indexes to all the paper-based and electronic medical records for each patient, containing information such as demographic characteristics, medical diagnostic codes, surgical procedure codes and death information (including causes of death) (St Sauver, et al., 2012). Data collection is ongoing, and medical records are added either quarterly or twice a year. Dental clinics are recently being incorporated into the system as well (Rocca WA, Yawn BP, St Sauver, Grossardt, & Melton, 2012). Olmsted County has been one of the few places in the world where occurrence and natural history of almost any diseases or syndrome can be accurately described with a healthcare data of patients that spans over a half a century. In comparison to the entire US population, Olmsted County is less ethnically diverse (90.3% vs. 75.1% white), more educated (91.1% vs. 80.4% high school graduates) and wealthier ($51,316 vs. $41.944 median household income) (St. Sauver, Grossardt, Leibson, Yawn, Melton III, & Rocca, 2012).

The REP data utilized for this project consisted of patient ID, demographic information (sex, race, date of birth), ICD9-CM and HCUP CCS diagnostic codes, counts

of diagnosis codes in each visit, length of stay and visit dates. ICD9-CM consists of more than 14,000 diagnostic codes with fine granularity and details. The Agency for Healthcare Research and Quality (AHRQ) developed a collection of databases and related software tools through Healthcare Cost and Utilization Project (HCUP) that enabled research on a broad range of topics including cost and quality of healthcare services, treatment outcome, medical practice patterns, etc. (Agency for Healthcare Research and Quality, 2014). The clinical classification software (CCS) in HCUP classifies ICD9-CM diagnosis codes and Current Procedural Terminology (CPT) codes into more manageable and clinically meaningful categories (Elixhauser, Steiner, & Palmer, 2014) (Agency for Healthcare Research and Quality). The single-level HCUP codes clusters ICD9-CM codes into 280 groups and multi-level HCUP CSS codes further group the single-level codes into 18 main groups.

We selected patients in the REP that were 18 years old or younger at the time of their hospital visit by subtracting their visit data from their birth date. A cohort of 46,020 patients (23,128 female and 22,892 male) with 271 unique HCUP CCS codes and 6,902 unique ICD9 codes during 6 years from 2004 to 2009 was constructed. Figure 5.1 shows the patient population stratified by race/ethnicity with number of patients and their corresponding percentage separated by comma.

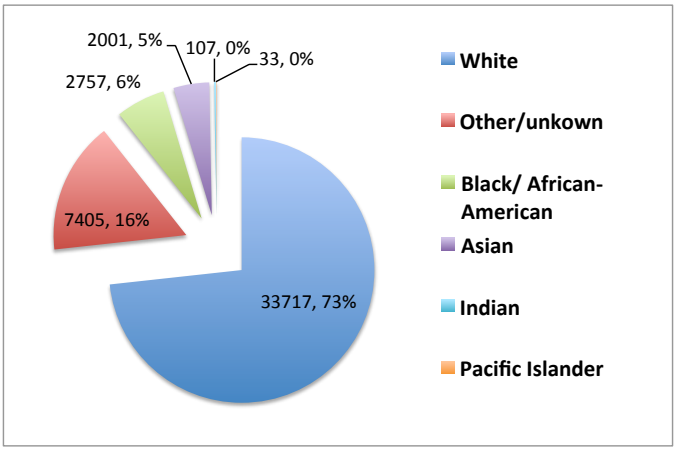


Figure 5.1 Patients Race Distribution

5.3 Methods

5.3.1 Representation Model

In order to model every patient clinical record as a diagnoses matrix, a patient class with attributes such as patient ID, race, gender, and two diagnoses (ICD9 and HCUP) matrices was constructed. For every patient, an instance of the patient object class was created by updating its attributes. In each diagnosis matrix, the row and column represent the diagnosis code and year of diagnosis respectively. In order to create a matrix of diagnosis codes and year of diagnosis with manageable size, we limited the granularities of diagnosis date to year of diagnosis to reduce the number of possible visit dates and correspondingly the number of columns in the matrix. To reduce the number of rows, 6,902 in ICD9 matrix and 271 in HCUP matrix, we selected the most frequent ICD9 and HCUP codes in our cohort. We selected unique ICD9 codes assigned to each patient in the cohort in order to find the number of patients in each ICD9 category. Figure 5.2 and Figure 5.3 show ICD9 and HCUP CSS codes assigned to each patient in the cohort with the number of patients on the logarithmic scale.

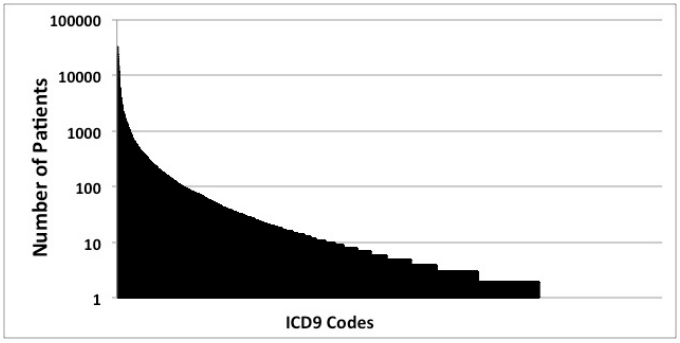


Figure 5.2 ICD9 Diagnosis Codes Histogram

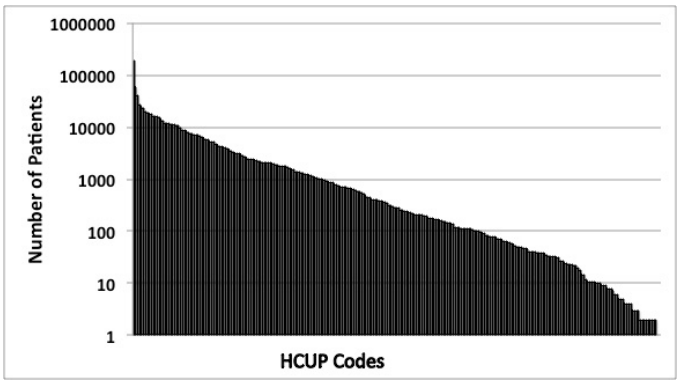


Figure 5.3 HCUP Diagnosis Codes Histogram

The most frequent diagnosis codes were selected by using 90 percentile of the HCUP code distribution and 99 percentile of the ICD9 codes distribution. There were 28 HCUP codes selected based on 90-percentile distribution and 70 ICD9 codes based on the 99-percentile distribution. Table 5.1 and Table 5.2 list the top 10 most frequent HCUP and ICD9 codes respectively in our data set.

Table 5.1 Ten Most Frequent ICD9 Codes in the Cohort

ICD9 Code	ICD9 Code Description	Number of Patients	Number of Occurrences
V20.2	Routine infant or child health check	32840	121672
V04.81	Need for prophylactic vaccination and inoculation against influenza	24487	67036
462	Acute pharyngitis	22595	61704
465.9	Acute upper respiratory infections of unspecified site	20707	53831
V06.1	Need for prophylactic vaccination with diphtheria-tetanus-pertussis combined [dtp] [dtap] vaccine	20706	33640
382.9	Unspecified otitis media	16939	55269
V05.3	Need for prophylactic vaccination and inoculation against viral hepatitis	16937	29212
V05.9	Need for prophylactic vaccination and inoculation against unspecified single disease	16224	32768
V04.0	Need for prophylactic vaccination and inoculation against poliomyelitis	16173	31646
V05.4	Need for prophylactic vaccination and inoculation against varicella	15535	17627

Table 5.2 Ten Most Frequent HCUP Codes in the Cohort

HCUP Code	HCUP Code Description	Number of Patients	Number of Occurrences
10	Immunizations and screening for infectious disease	39076	344545
126	Other upper respiratory infections	31676	136170
255	Administrative/social admission	34350	133889
92	Otitis media and related conditions	18277	73444
133	Other lower respiratory disease	17828	45628
94	Other ear and sense organ disorders	16385	36420
200	Other skin disorders	15333	37190
7	Viral infection	14680	28283
256	Medical examination/evaluation	14635	23233
259	Residual codes; unclassified	13695	25885

Not all patients have the most frequent selected codes therefore selecting the most frequent ICD9 codes reduced the number of patients from 46,020 to 45,066 while HCUP codes reduced this number of 45,627. The size of ICD9 event matrix is 70×6 representing 70 rows of most frequent ICD9 codes in the cohort and six diagnosis years of 2004 to 2009 as columns. Similarly the HCUP matrix has 28 rows of the most frequent HCUP codes and six diagnosis years of 2004 to 2009 as columns. To populate the matrix with the values corresponding to the patients' diagnosis code and year of visit, two default matrices of ICD9 codes with 70×6 dimensions and HCUP codes with 28×6 dimensions and default zero values were constructed. For every patient the matrix element corresponding to the diagnosis year and HCUP or ICD9 code was identified and its default value of zero was replaced with one. The matrix was updated by adding to the previous state for every new data entry related to that patient. And finally the matrix values were normalized by replacing any value higher than zero with one for simplicity and accommodating the bias in patients with frequent visits. The idea of representing

each patient records as a matrix is to find common patterns shared among all patient records similar to pattern discovery in image processing.

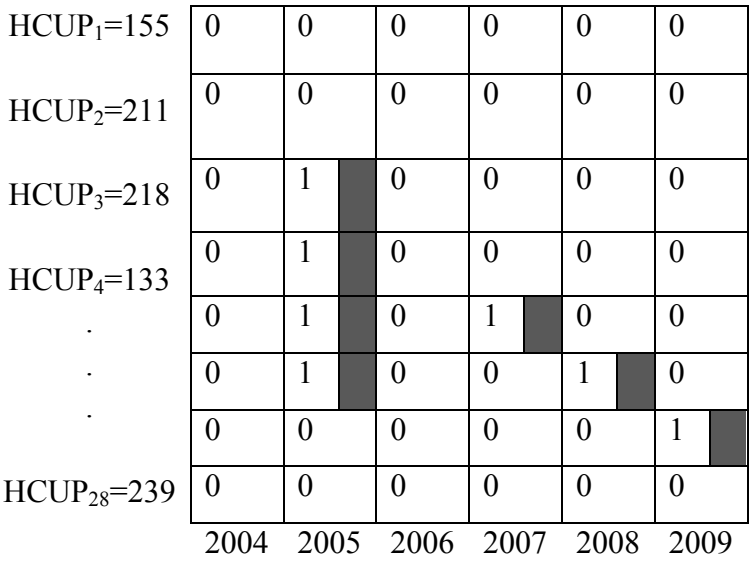


Figure 5.4 shows an example of HCUP matrix. Grayed out cells form two horizontal and diagonal lines showing an example of patterns that could be common among all HCUP matrices.

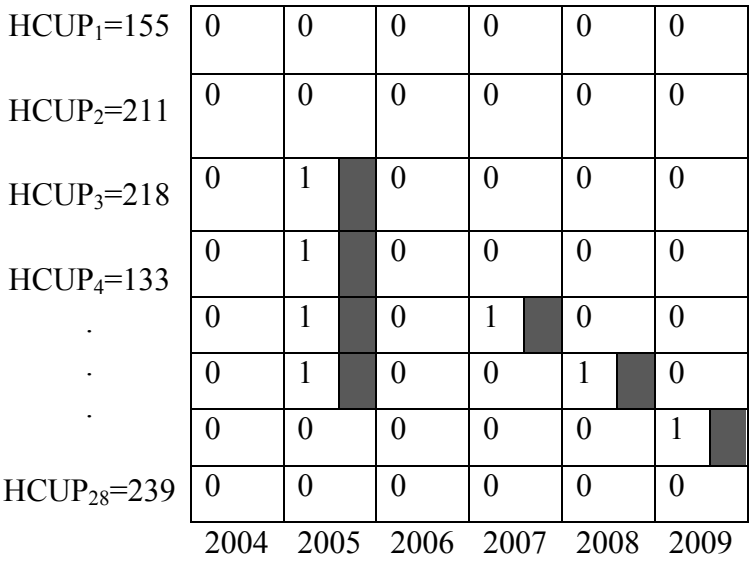


Figure 5.4 An Example of a Longitudinal Patient's Record Represented as a Matrix

5.3.2 Deep Learning Algorithm for Temporal Pattern Discovery

Restricted Boltzmann machine (RBM) is a two layer undirected graphical model that unlike Boltzmann machine algorithm (Hinton & Sejnowski., 1983) has no hidden to

hidden and visible to visible connections. It uses connecting weights *W* between visible units *v* and hidden units *h* as shown in Figure 5.5 to define the joint probability of these two layers *P(v, h; W)* with an energy function E (Hinton G. , 2002).

$$p(v,h) = \frac{\exp(-E(v,h))}{z} \quad (5\text{-}1)$$

Where $Z == \sum_v \sum_h \exp(-E(v,h))$ is called the partition function. The energy function can be described as either equation (5-2) or (5-3) depending on the visible units having real or binary values correspondingly:

$$E(v, h; W) = \frac{1}{2} \sum_i v_i^2 - \sum_{i,j} v_i W_{i,j} h_j - \sum_j b_j h_j - \sum_i c_i v_i \quad (5\text{-}2)$$

$$E(v, h; W) = -\sum_{i,j} v_i W_{i,j} h_j - \sum_j b_j h_j - \sum_i c_i v_i \quad (5\text{-}3)$$

Where b_j are hidden units biases and c_i are visible unit biases.

The network assigns a probability to every pair of visible and hidden vector using energy function defined in (5-1). The weights can be learned using stochastic gradient descent on the log likelihood of training data. However computing the exact gradient of log-likelihood is intractable. A common alternative is contrastive divergence (CD) approximation (Arnold, Rebecchi, Chevallier, & Paugam-Moisy, 2011), which still has some limitations such as optimal choice of the number of Markov chain transitions. A new learning algorithm called Persistent Contrastive Divergence (PCD) remove this limitation by persisting the Markov chain states from the previous iteration of the gradient calculation rather than from the training data (Tieleman, 2008) (Tieleman & Hinton, 2009).

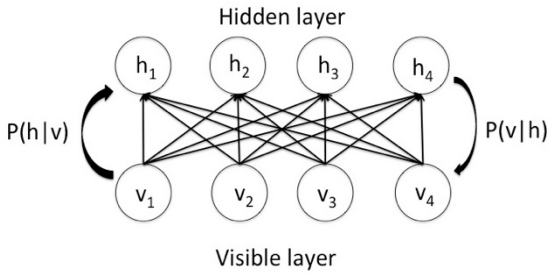


Figure 5.5 The RBM Architecture with a Visible (v) and Hidden (h) Layer

Multiple RBMs can be trained by using each hidden layer as training data for the next higher-level layer. This stack of RBMs can be viewed as a single probabilistic model

called deep belief network (DBN). Salakhutdinov and Hinton introduced Deep Boltzmann Machines (DBM) that also composed of multiple layers of RBM with a small modification to the DBN algorithm (Salakhutdinov & Hinton, 2009). They used variational approximation to estimate data dependent expectations and persistent Markov chain to estimate data independent expectation. These two estimation techniques make it practical to learn Boltzmann machines with multiple hidden layers and millions of parameters (Salakhutdinov & Larochelle, 2010).

Pylearn2 was used to implement DBM with three hidden layers and PCD learning algorithm (Goodfellow, et al., 2013). Pylearn2 is a machine library built on top of Theano (Bastien, et al., 2012) and written in python with an emphasis on flexibility and extensibility.

There are number of parameters that are required to be tuned to optimize the DBM learning algorithm such as learning rate, number of epochs, and initial momentum. Epoch is the number of iteration over the input data to learn the patterns. Learning rate is the parameter that controls the weight and bias size changes during the learning, the lower it is, the slower the learning will be and if it is too high the weights and objective function will diverge and there would be no learning. We used a manual search to find the hyper parameters by monitoring the network error rates.

5.4 Results

5.4.1 CCS-HCUP Diagnosis Codes

A set of 45,627 HCUP matrices with 28×6 dimension were used as inputs to a DBM network. Figure 5.6, Figure 5.7 and Figure 5.8 show the heatmaps of hidden layer weights of the DBM network.

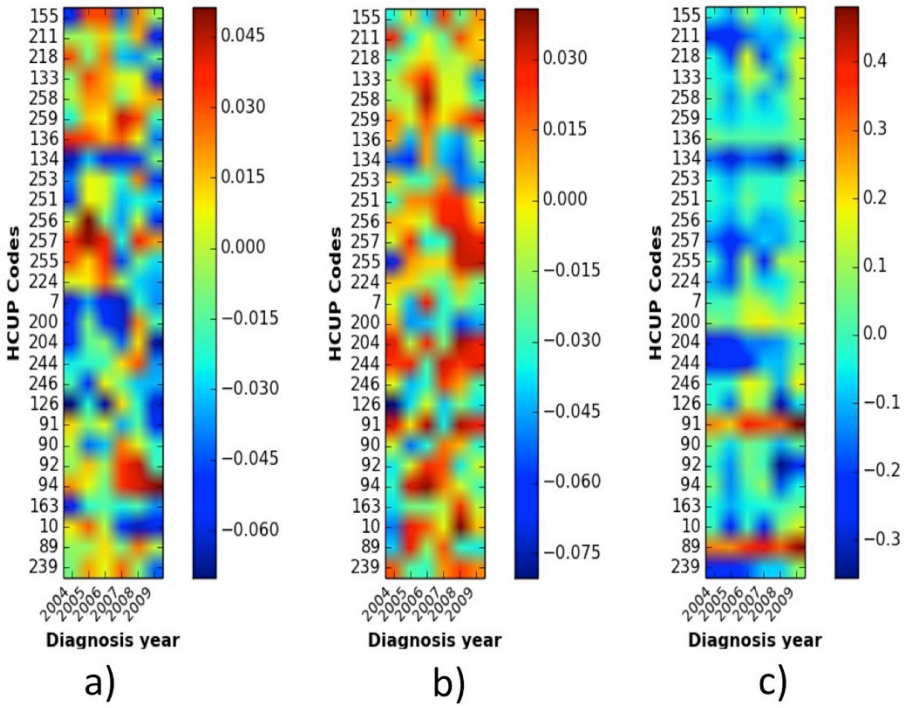


Figure 5.6 Heatmaps of First Hidden Layer Weights

We have randomly chosen 3 hidden nodes with the most distinctive patterns among the other nodes in the network. Figure 5.6 shows the value of weights connecting the visible units to the first hidden layer in the network. The value of these weights shows the strength or importance of the visible nodes contribution to the hidden layer nodes. Higher values of weights are shown by red color versus the blue color that represents lower values in the heatmaps figures. Figure 5.7 and 5.8 show the weights of second and third hidden layer of the DBM network.

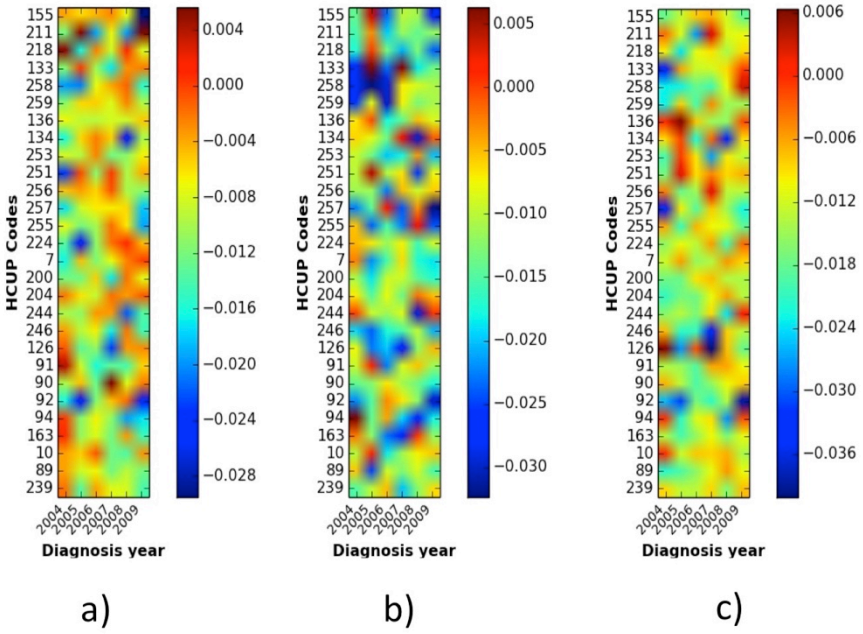


Figure 5.7 Heatmaps of Second Hidden Layer Weights

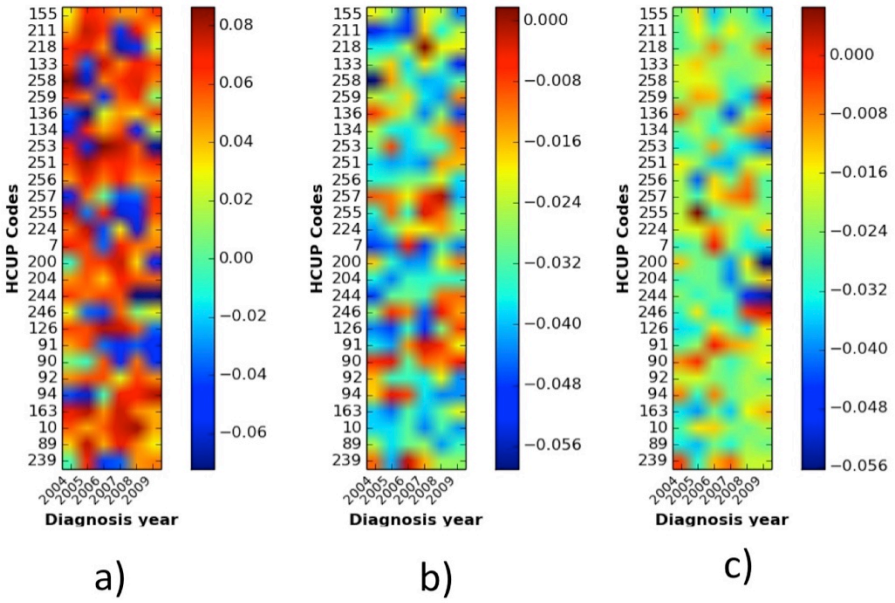


Figure 5.8 Heatmaps of Third Hidden Layer Weights

5.4.2 ICD9 Diagnosis Codes

Similarly, 45,066 ICD9 matrices were constructed with ICD9 codes as rows and years of diagnosis as columns. Figure 5.9, Figure 5.10 and Figure 5.11 show heatmaps of first, second and third hidden layers weights correspondingly.

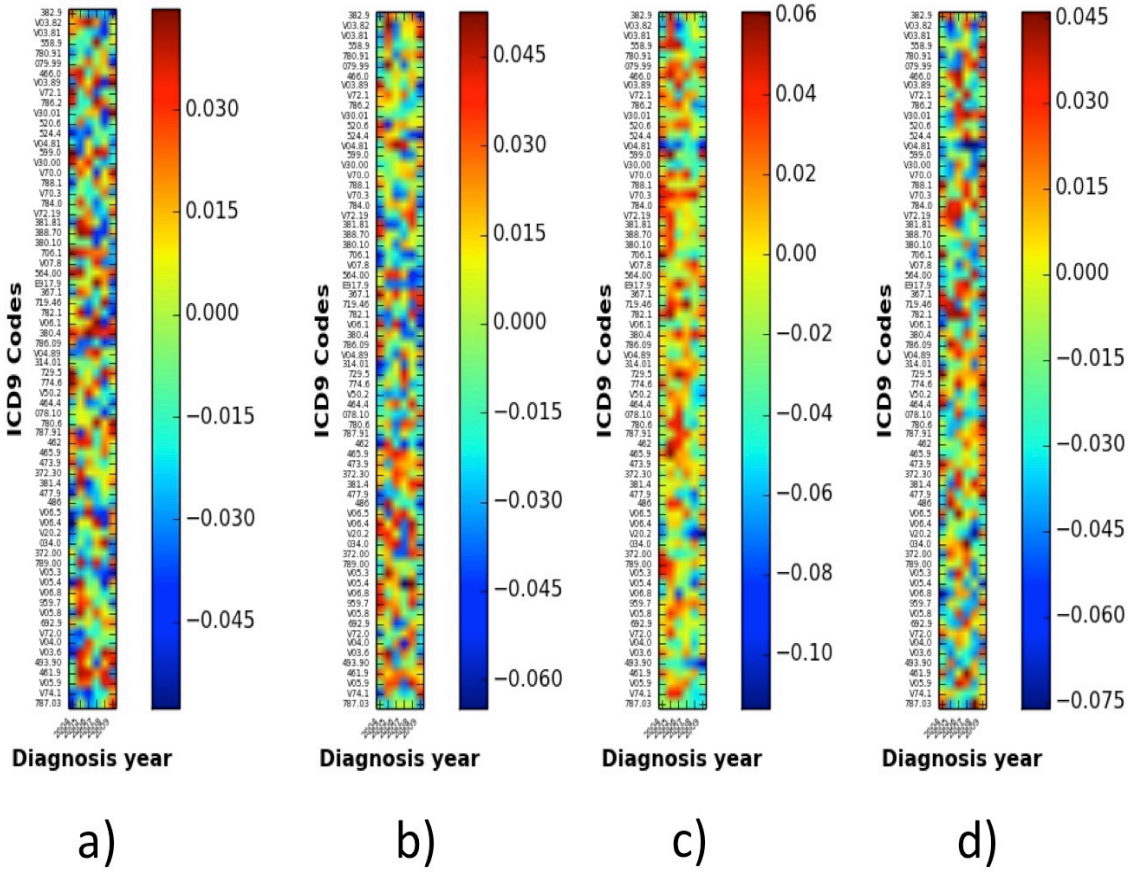


Figure 5.9 Heatmaps of First Hidden Layer Network with ICD9 Matrices as Inputs

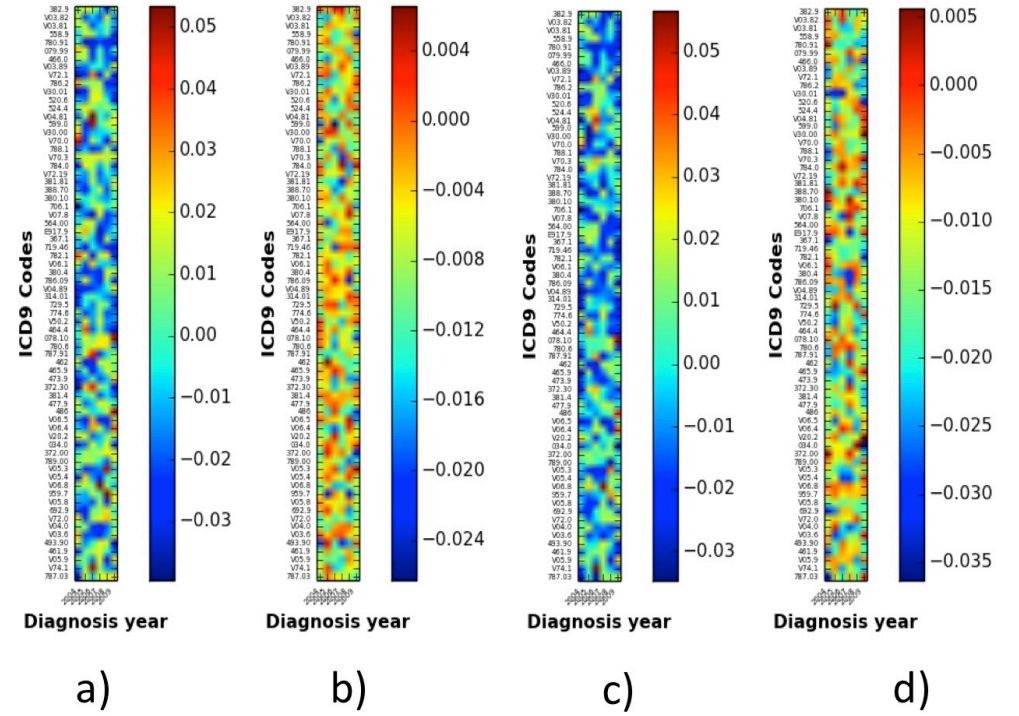


Figure 5.10 Heatmaps of Second Hidden Layer Weights

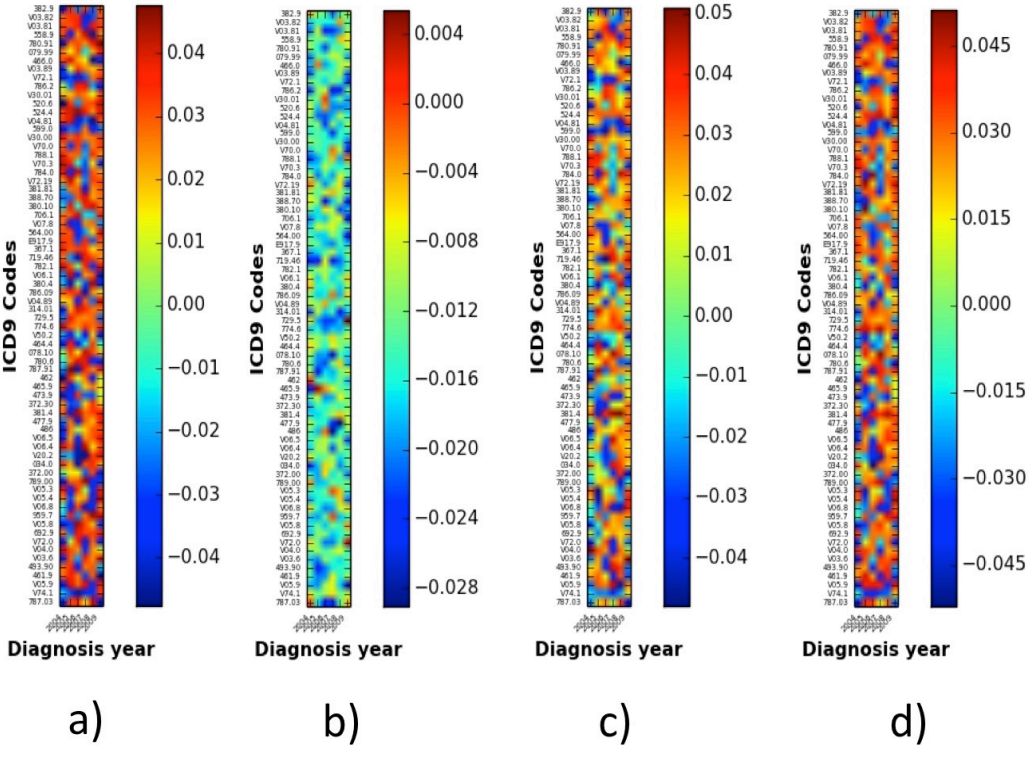


Figure 5.11 Heatmaps of Third Hidden Layer Weights

5.5 Discussion

Figures 5.6 to 5.8 visualize the final learned weights of three layers of DBM network after training on 45,627 matrices of patients with HCUP diagnosis. Figures 5.7 and 5.8 show the value of weights connecting the first hidden layer to the second hidden layer and second hidden layer to the third hidden layer correspondingly. The value of nodes in these hidden layers are an aggregate of visible node values multiplied by their weights values. Therefore, the discovered patterns in higher-level layers of DBM are abstracted representation of inputs.

If we associate the red dots in each column of Figures 5.6*a* to 5.6*c* we can find an association between HCUP codes in each year. For instance in Figure 5.6b an association between following codes can be seen: 211 (Other connective tissue disease), 204 (Other non-traumatic joint disorders), 244 (Other injuries and conditions due to external causes) and 239 (Superficial injury; contusion).

Horizontal lines on Figure 5.6c show the chronic patterns of HCUP codes 91 (Other eye disorders) and 89 (Blindness and vision defects). Also from Figure 5.6a, it can be inferred that disorders of teeth and jaw (HCUP code 136) was significant in first years of patients from 2004 to 2006 with a red line that changed to blue in later years showing how the disorder evolved through time.

Also the clustering in the years 2008 and 2009 of Figure 5.6a between codes 92 (Otitis media and related conditions) and 94 (Other ear and sense organ disorders) shows the strong relationship between the two codes, the same clustering with a reverse steep can be seen in Figure 5.6b. There are other temporal clusters in Figure 5.6b such as cluster between 256 (Medical examination/evaluation), 257 (Other aftercare) and 255 (Administrative/social admission) and a cluster between 204 (Other non-traumatic joint disorders) and 244 (Other injuries and conditions due to external causes).

Similarly Figure 5.9a-d show the weights of DBM network learned over 45,066 ICD9 matrices of patients' record.

An association between various diagnosis codes related to respiratory disease can be seen such as association between 473.9 (Unspecified Sinusitis), 786.2(Cough), 466.0 (Acute Bronchitis), and 380.4(Impacted Cerumen) in year 2004 of Figure 5.9a.

Association between 466.0 (Acute Bronchitis), 461.9 (Acute Sinusitis Unspecified), 462 (Acute Pharyngitis), 381.81 (Dysfunction Of Eustachian Tube), 388.7 (Otalgia Unspecified), 381.4 (Nonsuppurative Otitis Media Not Specified As Acute Or Chronic), and 493.9 (Asthma Unspecified) in the same figure along the 2005 columns and finally on the year 2006 of the same figure association between 462 (Acute Pharyngitis), 461.9 (Acute Sinusitis Unspecified), and 388.70 (Otalgia Unspecified). In figure 9b there is association between 520.6 (Disturbances In Tooth Eruption) and 524.4 (Malocclusion Unspecified) which both are mouth disorders.

We can also see horizontal line across the columns representing chronic disease such as 706.1 (Other Acne) and 380.4 (Impacted Cerumen) in Figure 5.9a. A diagonal line between v50.2 (Routine Or Ritual Circumcision), 774.6 (Unspecified Fetal And Neonatal Jaundice) and 729.5 (Pain In Limb) can be observed in Figure 5.9a. In a study in 2008, researchers studied the effects of circumcision on jaundice in newly born babies (Eroğlu, Balci, Ozkan, Yörükalp, & Göksel, 2008) (Shandiz, MacKenzie, Hunt, & Anglin, 2014). They selected 60 male patients, of whom 30 were circumcised. Babies were tracked for 35 to 40 gestational weeks and no statistically significant result were found between the two groups of patients. Literature has shown the decrease rate of urinary tract infection in circumcised newborn babies (Singh-Grewal, Macdessi, & Craig, 2005). Comparing the diagnosis codes v50.2 (Routine Or Ritual Circumcision) and 599 (Urinary Tract Infection Site Not Specified) in figure 9a shows that the higher values (red color) of ICD9 code 599 are associated with lower values (blue color) of ICD9 code V50.2 on the same year.

5.5.1 Limitations

Convolutional restricted Boltzmann machines (CRBM) are similar to RBM but the weights between hidden and visible layers are shared among all locations in the hidden layer. It has been shown that CRBM has better performance on several pattern recognition tasks (Lee, Grosse, Ranganath, & Ng, 2009) (Norouzi, Ranjbar, & Mori, 2009). However CRBM creates new features that are nonlinear combinations of the input variables and therefore it is not possible to identify the original input variables that derived the final detected pattern.

Although we constructed a DBM network in this study, we could only analyze the first hidden layer patterns due to the same reason as above and basically utilized a RBM rather than a DBM in our temporal pattern discovery.

5.5.2 Future Work

There are various optimization methods such as grid search, random search, gradient-based optimization etc. to select the best hyper parameters of a network. Hyperopt[2] is a state of the art hyper parameter optimization package that uses sequential model-based optimization techniques to automatically select the optimum hyper parameters. Hyperopt will be used in pylearn2 to select the best parameters. Also comparison of our methodology with other pattern discovery methods such as temporal topic modeling will be studied.

The Kids' Inpatient Database (KID) is a nationwide database of pediatric inpatient care from community hospitals participating in HCUP (Chu, Houchens, Elixhauser, & Ross, 2007) (Agency for Healthcare Research and Quality Healthcare Cost and Utilization Project (HCUP), 2015). Expanding our work to the KID's nationwide database and comparing the results with REP cohort will be explored.

Another factor of this investigation may be to select various different cohorts with different characteristics to compare results. The initial cohort of 0-18 year olds will have significantly different results than a cohort of 60-80 year olds. Also, incorporating longer temporal trends of information (10 or more years) may also show different types of trends over time.

On the interpretation side, further investigation and exploration needs to be done in order to summarize and interpret the results in a manner that will allow medical experts to start analyzing the patterns that have emerged in this analysis. A pre-screening of data may be required, filtering out any extraneous billing coded information and allowing the algorithm to focus on the most relevant and impactful diagnostic codes.

---

[2] http://hyperopt.github.io/hyperopt/

5.6 Conclusion

In this study we selected patients 18 or younger in REP cohort and modeled their medical record as diagnosis matrices with the diagnosis code as the rows and year of diagnosis as the column. We used DBM network to find common temporal patterns among the diagnosis matrices (Mehrabi, et al., 2015).

The deep learning results showed relationships, which would be expected, such as diagnosis codes for blindness correlated with codes for eye disorders. This face validation shows that the underlying technique of developing these patterns via deep learning can find expected results. Further exploration of additional patterns will need to involve more work with medical subject matter experts, and exploration of patterns that may not have preconceived significance but may indicate underlying patterns that need further exploration.

The significance of this research is the ability to generate new hypotheses based on data, and identify relationships in data that are not readily apparent. By utilizing deep learning and other advanced analytical techniques, new potential correlations can be found which can be the basis of new research questions. Those potential questions will have to be filtered by researchers with medical domain knowledge and the ability to discern and interpret the patterns as they relate to the medical diagnostic events.

In the following we discuss some of the limitation to this approach and our future works.

CHAPTER SIX: CLOSING REMARKS

6.1 Summary of Contributions

There are enormous amount of clinical information embedded in free text format that pose challenges in its secondary use. NLP has offered opportunities to tap into clinical text to extract information needed for various clinical applications. We used sublanguage analysis to identify pancreatic cyst concepts from patient longitudinal records. However the meaning of concept can be affected by its surrounding contextual text, for instance the presence of a clinical concept in patient's report does not imply that the patient has the finding. In this work, we studied the effects of two such contextual modifiers negation and family history. We developed a negation algorithm called DEEPEN that is built on top of NegEx and uses a chain of dependency relations to find the contextual relationship between negation words and clinical concepts. The detection of family history in clinical notes consists of various preprocessing steps such as section segmentation, negation detection, relation discovery, etc. We used a conditional random field algorithm to identify family history section and a set of rules to identify the relationship between family members and diagnosis concepts.

When the concepts of interest are extracted using NLP from longitudinal records of a patient, their temporal order is not considered. In order to represent the temporal dimension of healthcare data, we modeled each patient's records as a matrix with clinical events as rows and time of encounter as columns. Once a cohort of patients is represented by diagnosis/event matrices, common pattern discovery methods such as deep learning algorithm can be applied. We analyzed the heatmap representation of weights of a deep Boltzmann machine to find common features among patients' records.

APPENDIX

Following are the detailed explanation of DEEPEN rules with example sentences.

a) **"Conjunction And" *(conj_and)* Rule**

The concept "*pseudocyst*" in Figure 1 is negated by NegEx because of negation verb "does not". In DEEPEN however if the dependency parser contains the dependency relation "conj_and", the sentence is split into sub sentences and negation is checked for each sub sentence. Because there is no negation term in the second sub sentence containing the concept "*pseudocyst*", it is affirmed by DEEPEN.

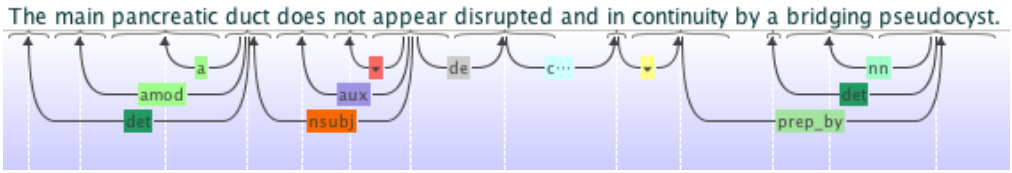


Fig. 1: Dependency relation for a sentence with "conj_and" relation

b) **"Preposition Without" *(prep_without)* Rule**

If there is "*prep_without*" dependency in the SDP chain, its governor is added to the first level token list. Therefore, the dependency chain for this sentence would be *(size)(pancreas is normal inflammation) (The dilatation)* where "*size*" is the first level token, "*pancreas", "is", "normal", and "inflammation*" are second level tokens and "the", "*peripancreatic", and "dilatation" are third level tokens.* For concepts that are noun phrase such as "*pancreatic duct dilatation*", even if part of the noun phrase is in the dependency chain (dilatation), the concept is negated.

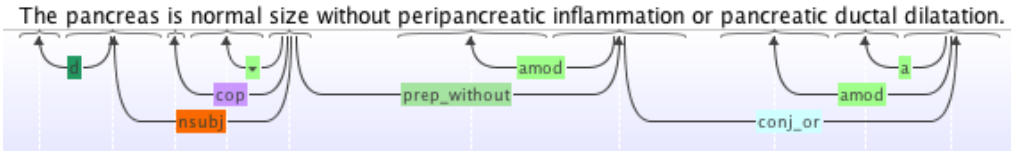


Fig. 2: Dependency relation for a sentence with "prep_without" relation

c) **Preposition (*prep_in*, *prep_with*, *prep_within*) Rule**

The sentence in Figure 3 contains the dependency relation (*conj_and*), therefore based on the rule "a) conj-and" it is split into two sentences and dependency relations is generated for each sub-sentence as shown in Fig. 3a and Fig. 3b. If the SDP contains one

of the dependencies: "*prep_in*", "*prep_with*" or "*prep_within*" and either the governor or dependent term is the concept, then the dependency chain is generated otherwise the concept is affirmed. In the sub-sentence "*gallbladder with no dilated ducts*" because the dependent of relation *prep_with (gallbladder-1, ducts-5*) is part of the concept "*dilated duct*", the dependency chain is generated, which is "*ducts (first level token) dilated (second level tokens)"*. The concept (*dilated ducts*) is in dependency chain and therefore negated.

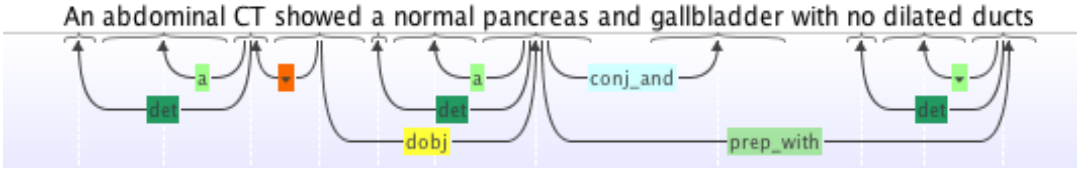


Fig. 3. a) Dependency relation for a sentence with "prep_with" relation

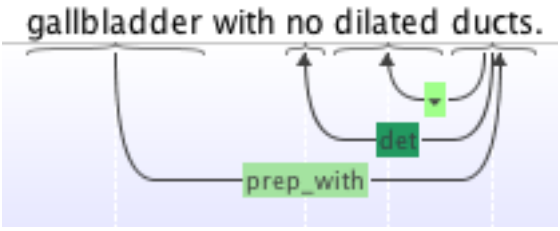


Fig. 3. b) SDP after splitting the sentence into two sentences

d) **Nominal Subject (*nsubj*) Rule**

If the SDP contains the relation "nsubj" and its dependent term is in the dependency chain, then its governor term is added to the dependency chain. In the sentence "*No abnormally dilated pancreatic duct*", shown in Fig. 5, "*abnormally*" is the dependent term in the relation nsubj (dilated-3 abnormally-2). It is also in dependency chain as the first level token, therefore its governor "*dilated*" is added to the dependency chain. The final dependency chain is (*dilated pancreatic duct*) and the concept is negated.

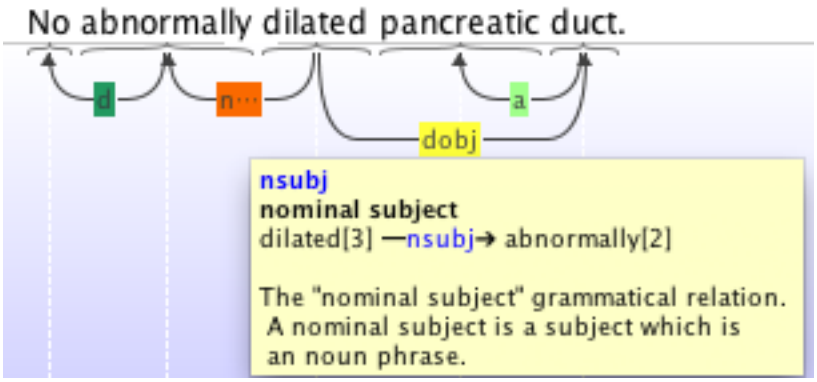


Fig. 4. Dependency relation for a sentence with "nsubj" relation

e) **Suggest Rule**

If a sentence contains "*suggest*" and its dependent is a first level token then "*suggest*" is added to the first level tokens. In the sentence "*No associated fluid collection to suggest* ***pseudocyst*** *or abscess.*" Shown in Fig. 5, "*suggest*" is the governor in the following dependency relations: nsubj (suggest-6, collection-4*)*, and its dependent terms "*collection*" is a first level token det (collection-4, No-1) therefore the dependency chain is "*(collection) (associated fluid) (pseudocyst)"*. The concept "*pseudocyst*" is in the dependency chain and therefore it is negated.

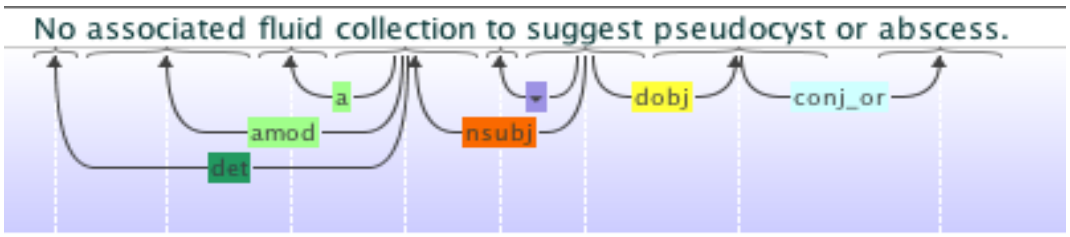


Fig. 5. Dependency relation for a sentence with "suggest" as the dependent term

REFERENCES

Afzal, Z., Pons, E., Kang, N., Sturkenboom, M., & Schuemie, M. (2014). ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics, 15*, 373.

Agency for Healthcare Research and Quality Healthcare Cost and Utilization Project (HCUP). (2015, May 21). Retrieved from Introduction to the HCUP KIDS' Inpatient Database (KID) 2012 : https://www.hcup-us.ahrq.gov/db/nation/kid/kid_2012_introduction.jsp

Agency for Healthcare Research and Quality. (2014, August). Healthcare Cost and Utilization Project (HCUP). Retrieved from http://www.ahrq.gov/research/data/hcup/index.html

Agency for Healthcare Research and Quality. (n.d.). (2015, March). HCUP Clinical Classifications Software (CCS) for ICD-9-CM. Healthcare Cost and Utilization Project (HCUP). Retrieved from www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

Al-Haddad , M. A., Friedlin , J., Kesterson , J., Waters , J. A., Aguilar-Saavedra, J. R., & Schmidt , C. M. (2010). Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *International Hepato-Pancreato-Biliary Association, 12* (10), 688-95.

Alias-i (2008). Indo European Token Categorizer. Retrieved July 10, 2013, from Indo European Token Categorizer: http://alias-i.com/lingpipe/docs/api/com/aliasi/tokenizer/IndoEuropeanTokenCategorizer.html

Arnold, L., Rebecchi, S., Chevallier, S., & Paugam-Moisy, H. (2011). An Introduction to Deep Learning. *European Symposium on Artificial Neural Networks (ESANN)*.

Aronow, D. B., Feng, F., & Croft , W. B. (1999). Ad Hoc Classification of Radiology Reports . *Journal of the American Medical Informatics Association, 6* (5), 393-411.

Ballesteros, M., Francisco, V., Díaz, A., Herrera, J., & Gervás, P. (2012). Inferring the scope of negation in biomedical documents. *CICLING*. New Delhi.

Batal, I., Valizadegan, H., Cooper, G. F., & Hauskrecht, M. (2013). A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology, 4* (4).

Batal, I., Fradkin, D., Harrison, J., Moerchen, F., & Hauskrecht, M. (2012). Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. *KDD*, (pp. 280-288). Beijing, China.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., … Bengio, Y. (2012). Theano: new features and speed improvements. *NIPS deep learning workshop*.

Behnam, E., & Smith, A. D. (2014). The Amordad database engine for metagenomics. *Bioinformatics, 30* (20), 2949-2955.

Behnam, E., Waterman, M. S., & Smith, A. D. (2013). A geometric Interpretation for local alignment-free sequence comparison. *Journal of Computational Biology, 20* (7), 471-485.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res, 32*, 267-70.

Brune, K. A., Lau , B., Palmisano, E., Canto, M., Goggins, M. G., Hruban, R. H., & Klein, A. P. (2010). Importance of age of onset in pancreatic cancer kindreds. *Journal of the National Cancer Institute, 102* (2), 119-26.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., Savova, G. K. (2011). The MiPACQ clinical question answering system. *American Medical Informatics Association Annual Symposium proceedings*, (pp. 171–180).

Cancer Research UK. (2012). Cancer Worldwide - the global picture. Retrieved Feb 1, 2013, from Cancer Research UK: http://www.cancerresearchuk.org/cancer-info/cancerstats/world/the-global-picture/cancer-overall-world#source2

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22* (2).

Carbonell, J. G., & Hayes, P. J. (1992). Natural Language Understanding. *Encyclopedia of Artificial Intelligence* (pp. 997-1016). John Wiley & Sons.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for Identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics, 34* (5), 301–310.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., W D'Avolio, L., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association, 18* (5), 540-543.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). Evaluation of negation phrases in narrative clinical reports. *American Medical Informatics Association Annual Symposium proceedings.* (pp. 105–9).

Choi, J. D., & Palmer, M. (2011). Getting the most out of transition-based dependency parsing. *ACL: HLT*, (pp. 687–92). Portland, Oregon.

Chu, B., Houchens, R., Elixhauser, A., & Ross, D. (2007). Using the KIDS' Inpatient Database (KID) to Estimate Trends. HCUP Methods Series, U.S. Agency for Healthcare Research and Quality.

Cleary, S. P., Gryfe , R., Guindi , M., Greig , P., Smith , L., Mackenzie , R., … Gallinger, S. (2004). Prognostic factors in resected pancreatic : analysis of actual 5-year survivors. *Journal of the American College of Surgeons 198*, 722-31.

Cohen, R., & Elhadad, M. (2012). Syntactic dependency parsers for biomedical-NLP. *American Medical Informatics Association Annual Symposium proceedings*, (pp. 121–8).

Concaro, S., Sacchi, L., Cerra, C., & Bellazzi, R. (2009). Mining administrative and clinical diabetes data with temporal association rules. Stud Health Technol Inform , 150, 574–578.

Cruz Díaz, N. P., Maña López, M. J., Vázquez, j. M., & Álvarez V, V. P. (2012). A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the Association for Information Science and Technology, 63* (7), 1398–1410.

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., & Zhu, X. (2011). Machine- learned solutions for three stages of clinical information extraction: the state of the art at

i2b2 2010. *Journal of the American Medical Informatics Association*, *18*, 557-562.

de Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. *The International Conference on Language Resources and Evaluation*.

de Marneffe, M. C., & Manning, C. D. (2008, September). Stanford typed dependencies manual. Retrieved September 7, 2014, from The Stanford Natural Language Processing Group: http://nlp.stanford.edu/software/dependencies_manual.pdf

Deng , L., Li, J., Huang, J-T., Yao, K., Yu, D., Seide, F., et al. (2013). Recent Advances in Deep Learning for Speech Research at Microsoft. Acoustics, Speech, and Signal Processing (ICASSP).

Denny , J. C., Spickard 3rd, A., Johnson , K. B., Peterson , N. B., Peterson, J. F., & Miller, R. A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association, 16* (6), 806-15.

Denny, J. C., Miller, R. A., Johnson, K. B., & Spickard 3rd, A. (2008). Development and evaluation of a clinical note section header terminology. *American Medical Informatics Association Annual Symposium proceedings*, Nov 6, pp. 156-60.

Elixhauser, A., Steiner, C., & Palmer, L. (2014). Clinical Classifications Software (CCS). Agency for Healthcare Research and Quality.

Elkin, P. L., Brown, S. H., Bauer, B. A., Husse, C. S., Carruth, W., Bergstrom, L. R., & Eahner-Roedler, D. L. (2005). A controlled trial of auto- mated classication of negation from clinical notes. *BMC Medical Informatics and Decision Making, 5* (13).

Eriksson, R., Jensen, P. B., Frankild, S., Jensen, L. J., & Brunak, S. (2013). Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association, 20* (5), 947–953.

Eroğlu, E., Balci, S., Ozkan, H., Yörükalp, O., & Göksel, A. (2008). Does circumcision increase neonatal jaundice? *Acta Paediatrica, 97* (9), 1192-3.

Farkas, R., Vincze, V., Mora, G., Csirik, J., & Szarvas, G. (2010). The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. *Computational Natural Language Learning: Shared Task* (pp. 1-12). Uppsala, Sweden: Association for Computational Linguistics.

Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, 10* (3-4), 327 - 348 .

Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., & Sinclair, G. (2004). Exploiting context for biomedical entity recognition: from syntax to the web. *Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, (pp. 88–91). Geneva, Switzerland.

Firpo, M. A., Gay, D. Z., Granger, S. R., Scaife, C. L., DiSario , J. A., Boucher, K. M., & Mulvihill, S. J. (2009). Improved diagnosis of pancreatic adenocarcinoma using haptoglobin and serum amyloid A in a panel screen. *World Journal of Surgery , 33*, 716–722.

Friedlin, J., Overhage, M., Al-Haddad, , M. A., Waters, J. A., Aguilar-Saavedra, J. R., Kesterson, J., & Schmidt, M. (2010). Comparing methods for Identifying pancreatic cancer patients using electronic data sources. American Medical Informatics Association Annual Symposium proceedings (pp. 237–241).

Friedman, C., Hripcsak , G., DuMouchel, W., Johnson , S. B., & Clayton. (1995). Natural language processing in an operational clinical inforamtion system. *Journal of Natural Language Engineering, 1* (1), 83-108.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association, 1* (2), 161-74.

Friedman, C., Hripcsak, G., Shagina, L., & Liu H, H. (1999). Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, 6 (1), 76-87.

Fuchun, P., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *International Journal of Information Processing and Management, 42* (4), 963-979 .

Fundel, K., Küffner, R., & Zimmer, R. (2007). RelEx—Relation extraction using dependency parse trees. *Bioinformatics, 23* (3), 365–71.

Fu, J., Liu, Z., Liu , W., & Guo, Q. (2011). Using dual-layer CRFs for event causal relation extraction. *IEICE Electronics Express, 8* (5), 306-310)

Gacek, A., & Pedrycz, W. (Eds.). (2012). ECG Signal Processing, Classification and Interpretation A Comprehensive Framework of Computational Intelligence. Springer.

Garla , V., Lo Re III , V., Dorey-Stein , Z., Kidwai , F., Scotch , M., Womack , J., . . . Brandt, C. (2011). The Yale cTAKES extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association , 18* (5), 614-620.

Gindl, S., Kaiser, K., & Mik, S. (2008). Syntactical negation detection in clinical practice guidelines. *Studies in Health Technology and Informatics, 136*, 187–192.

Grishman , R., & Kittredg, R. (Eds.). (1986). Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu , R., . . . Bengio, Y. (2013). Pylearn2: a machine learning research library. *arXiv*1308.4214.

Goonetilleke, K. S., & Siriwardena, A. K. (2007). Systematic review of carbohydrate antigen (CA19-9) as a biochemical marker in the diagnosis of pancreatic cancer. *European Journal of Surgical Oncology, 33*, 266–270.

Goryachev, S., Kim, H., & Zeng-Treitler, Q. (2008). Identification and extraction of family history Information from clinical reports. AMIA Annual Symposium Proceedings, (pp. 247–251).

Goldin, M., & Chapman, W. W. (2003). Learning to detect negation with 'Not' in medical texts. *ACM- Special Interest Group on Information Retrieval*.

Grouin, C., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., … Zweigenbaum, P. (2010). CARAMBA: Concept, Assertion, and Relation Annotation using Machine-learning Based Approaches. *i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston.

Harkema, H., Dowling, J. N., Thornblade, T., & Chapman, W. W. (2009). ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics, 42*, 839–851.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18* (7), 1527–1554.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. Neural Computation, 14 (8), 1711–1800.

Hinton, G., & Sejnowski., T. (1983). Optimal perceptual inference. *IEEE conference on Computer Vision and Pattern Recognition*, (pp. 448-453). Washington, D.C.C.

Hirohata, K., Okazaki, N., Ananiadou, S., & Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. *International Joint Conference on Natural Language Processing*, (pp. 381–388).

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge based Systems, 6* (2), 107-116.

Howlader, N., Noone , A. M., Krapcho, M., Neyman , N., Aminou, R., Altekruse, S. F., . . . Cronin, K., A. (Eds.) (2012, November 1). SEER Cancer Statistics Review, 1975-2009 (Vintage 2009 Populations). Retrieved April 1, 2013, from SEER web site: http://seer.cancer.gov/csr/1975_2009_pops09/

Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association , 20*, 117–121.

Hu, H., Swaminathan, V. V., Zamani Farahani, M. R., Mensing, G., Yeom, J., Shannon, M. A. (2013). Hierarchical and re-entrant Micro/Nano-structures for superhydrophobic surfaces with extremely low hysteresis. *American Chemical Society*.

Huang, Y., & Lowe, H. J. (2007). A novel hybrid approach to auto- mated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association , 14*, 304-311.

Jakkula, V. R., Crandall, A. S., & Cook, D. J. (2009). Enhancing anomaly detection using temporal pattern discovery. *Advanced Intelligent Environments ,* 175-194.

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., . . . Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications, 5*, 4022.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association, 18* (5), 601–606.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., . . . Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science, 321* (5897), 1801–1806.

Klein, A. P., Brune , K. A., Petersen, G. M., Goggins, M., Tersmette, A. C., Offerhaus, G. J., et al. (2004). Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. *Cancer Research, 64* (7), 2634-8.

Laffan, T. A., Horton , K. M., Klein, A. P., Berlanstein, B., Siegelman, S. S., Kawamoto, S., . . . Hruban, E. H. (2008). Prevalence of unsuspected pancreatic cysts on MDCT. *American Journal of Roentgenology, 191* (3), 802-7.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282-289). San Francisco: Morgan Kaufmann Publishers Inc.

Lasko, T. A., Denny, J. C., & Levy, M. A. (2013). Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLOS ONE, 8* (6).

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. *Advances in Neural Information Processing Systems (NIPS).*

Lee , H. J., Kim, M. J., Choi, J. Y., Hong , H. S., & Kim , K. A. (2011). Relative accuracy of CT and MRI in the differentiation of benign from malignant pancreatic cystic lesions. *Clinical Radiology, 66* (4), 315-21.

Leaman , R., & Gonzalez , G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, (pp. 652-63).

Lennon, A. M., & Wolfgang, C. (2013). Cystic neoplasms of the pancreas. *Journal of Gastrointestinal Surgery, 17* (4), 645-53.

Lewis, N., Gruhl , D., & Yang, H. (2011). Dependency parsing for extracting family history. *International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB),* (pp. 237-242). IEEE.

Li , Z., Liu , F., Antieau, L., Cao , Y., & Yu, H. (2010). Lancet: a high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association, 17* (5), 563-7.

Li, D., Thermeau, t., Chute, C., & Liu, H. (2014). Discovering Associations Among Diagnosis Groups Using Topic Modeling. *AMIA Summits on Translational Science Proceeding*

Li, Y., Gorman, S. L., & Elhadad, N. (2010). Section classification in clinical notes using supervised hidden markov model. *ACM International Health Informatics Symposium*, , (pp. 744–750).

Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. *HLT-NAACL BioNLP Workshop*, (pp. 65–72).

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (pp. 2-11). ACM.

Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., et al. (2013). An Information Extraction Framework for Cohort Identification Using Electronic Health Records. *AMIA Summits on Translational Science Proceeding*, (pp. 149–153).

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved 9 27, 2014, from http://mallet.cs.umass.edu

McKnight, L., & Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. AMIA Annual Symposium Proceedings, (pp. 440– 444).

Mehrabi, S., Sohn, S., Li, D., Pankratz, J. J., Therneau, T., St. Sauver, J. L., . . . Palakal, M. (2015). Temporal Pattern and Association Discovery of Diagnosis Codes using Deep Learning. IEEE International Conference on Healthcare Informatics. Dallas, TX.

Mehrabi, S., Krishnan, A., Roch, A. M., Schmidt, H., Li, D., Kesterson, J., . . . Liu, H. (2015). Identification of Patients with Family History of Pancreatic Cancer-Investigation of an NLP System Portability. *Studies in health technology and informatics,* (pp. 604-608). São Paulo.

Mehrabi, S., Krishnan, A., Sohn, S., Roch, A. M., Schmidt, H., Kesterson, J., . . . Palakal, M. (2015). DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics, 54*, 213–219.

Melamed, R. D., Khiabanian, H., & Rabadan, R. (2014). Data-driven discovery of seasonally linked diseases from an Electronic Health Records system. *BMC Bioinformatics, 15* (Suppl 6), S3.

Merity, S., Murphy, T., & Curran, J. R. (2009). Accurate argumentative zoning with maximum entropy models. ACL Workshop on Text and Citation Analysis for Scholarly Digital Libraries, (pp. 19–26).

Morante, R., & Daelemans, W. (2011). Annotating modality and negation for a machine reading evaluation. CLEF.

Morante, R., & Sporleder, C. (2010). A special issue of the computational linguistics journal on modality and negation. *Computational Linguistics, 38* (2).

Morante, R. (2010). descriptive analysis of negation cues in biomedical texts. *The International Conference on Language Resources and Evaluation*. Valletta, Malta.

Morante, R., & Blanco, E. (2012). SEM 2012 Shared Task: Resolving the scope and focus of negation. *Lexical and Computational Semantics (SEM)* (pp. 265-274). Montreal, Canada: Association for Computational Linguistics.

Moskovitch, R., & Shahar, Y. (2013). Fast time intervals mining using the transitivity of temporal relations. *Knowledge and Information Systems, 42* (1), 21-48.

Moskovitch, R., & Shahar , Y. (2009). Medical temporal knowledge discovery via temporal abstraction. *American Medical Informatics Association Annual Symposium proceedings*, (pp. 452–456).

Moskovitch, R., Walsh, C., Hripcsak, G., & Taton, N. (2014). Prediction of biomedical events via time Intervals mining. *ACM KDD Workshop on Connected Health in Big Data Era.*

Mutalik, P. G., Deshpande, A., & Nadkarni, P. (2001). Use of general purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls. *Journal of the American Medical Informatics Association, 8*, 589-609.

Nabavizadeh, A., Greenleaf, J. F., Fatemi, M., & Urban, M. W. (2014). Optimized shear wave generation using hybrid beamforming methods. *Ultrasound in Medicine and Biology, 40* (1), 188-199.

Nabavizadeh, A., Greenleaf, J. F., Fatemi, M., & Urban, M. W. (2012). Shear wave generation using hybrid beamforming methods. *The Journal of the Acoustical Society of America, 132* (3), 1982-1982.

Norén, G. N., Hopstadius, J., Bate, A., Star, K., & Edwards, I. R. (2010). Temporal pattern discovery in longitudinal electronic patient records. *Data Mining Knowledge Discovery, 20* (3), 361–387.

Norouzi, M., Ranjbar, M., & Mori, G. (2009). Stacks of Convolutional Restricted Boltzmann Machines for Shift-Invariant Feature Learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2735 - 2742).

OpenNLP. Retrieved Nov 24, 2012, from OpenNLP: http://opennlp.sourceforge.net/index.html.

Ogren, P., Savova, G., & Chute, C. (2008). Constructing evaluation corpora for automated clinical named entity recognition. *International Conference on Language Resources and Evaluation Conference,* (p. 3143e50). Marrakesh, Morocco.

Ogren, P. V. (2006). Knowtator: A Protégé plug-in for annotated corpus construction. *North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*. New York.

Pandol, S., Gukovskaya, A., Edderkaoui, M., Dawson, D., Eibl, G., & Lugea, A. (2012). Epidemiology, risk factors, and the promotion of pancreatic cancer: role of the stellate cell. *Journal of Gastroenterology and Hepatology Research , 27* Suppl 2, 127–134.

Patnaik, D., Butler, P., Ramakrishnan, N., Parida, L., Keller, B. J., & David, A. H. (2011). Experiences with Mining Temporal Event Sequences from Electronic Medical Records: Initial Successes and Some Challenges. KDD. San Diego, California.

Payne, T. D. (1997). Describing Morphosyntax: A Guide for Field Linguists. Cambridge, UK: Cambridge University Press.

Permuth-Wey, J., & Egan, K. M. (2009). Family history is a significant risk factor for pancreatic cancer: results from a systematic review and meta-analysis. Fam Cancer , 8 (2), 109-17.

Permuth-Wey, J., & Egan, K. M. (2009). Family history is a significant risk factor for pancreatic cancer: results from a systematic review and meta-analysis. Familial cancer , 8 (2), 109-17.

Perotte, A., & Hripcsak, G. (2013). Temporal properties of diagnosis code time series in aggregate. IEEE Journal of Biomedical and Health Informatics , 17 (2), 477-483.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. B., et al. (2007). A shared task involving multi-label classification of clinical free text, *Workshop on BioNLP: Biological, Translational, and Clinical Language Processing*, (pp. 97-104).

Plaisant, C., Lam, S., Shneiderman, B., Smith, M. S., Roseman, D., Marchand, G., . . . Rappaport, H. (2008). Searching electronic health records for temporal patterns in patient histories: A case study with Microsoft Amalga. *AMIA Annual Symposium Proceedings*, (pp. 601-605).

Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., & Shneiderman, B. (1998). LifeLines: Using visualization to enhance navigation and analysis of patient records. *AMIA Annual Symposium Proceedings*, (pp. 76-80).

Raj, R., O'Connor , M. J., & Das , A. K. (2007). An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research. *AMIA Annual Symposium,* (pp. 614–619). Chicago.

Rastegar-Mojarad, M. (2013, December). Extraction and classification of drug-drug Interaction from biomedical text using a two-stage classifier. Wisconsin, Milwaukee, USA.

Rastegar-Mojarad, M., Boyce, R. D., & Prasad, R. (2013). UWM-TRIADS: classifying drug-drug Interactions with two-stage SVM and post-processing. *Joint Conference on Lexical and Computational Semantics,* (p. 667).

Rich, E. C., Burke, W., Heaton, C. J., Haga, S., Pinsky, L., Short, M. P., Acheson, L. (2004) Reconsidering the Family History in Primary Care. *Journal of General Internal Medicine, 19* (3), 273–280.

Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberge, L., Beebe, C. E., . . . Chute, C. G. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn Project. *Journal of Biomedical Informatics, 45* (4), 763-71.

Redekop, W. K., & Mladsi, D (2013). The Faces of Personalized Medicine: A Framework for Understanding its Meaning and Scope. *Value in Health, (16)* S4-S9

Rocca W. A., Yawn , B. P., St Sauver, J. L., Grossardt, B. R., & Melton, L. J. (2012). History of the Rochester Epidemiology Project: Half a Century of Medical Records Linkage in a US Population. *Mayo Clinic Proceedings, 87* (12), 1202–1213.

Roch, A. M., Mehrabi, S., Krishnan, A., Schmidt, H. E., Kesterson, J., Beesley, C., . . . Schmidt M. (2015). Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer. *International Hepato-Pancreato-Biliary Association, 17* (5), 447-53.

Sacchi, L., Larizza, C., Combi, C., & Bellazzi, R. (2007). Data mining with Temporal Abstractions: learning rules from time series. *Data Mining and Knowledge Discovery, 15* (2), 217-247.

Sahani, D. V., Shah, Z. K., Catalano, O. A., Boland, G. W., & Brugge, W. R. (2008). Radiology of pancreatic adenocarcinoma: current status of imaging. *Journal of Gastroenterology and Hepatology, 23*, 23–33.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann Machines. *International Conference on Artificial Intelligence and Statistics,* Clearwater Beach, Florida.

Salakhutdinov, R., & Larochelle, H. (2010). Efficient Learning of Deep Boltzmann Machines. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, (pp. 693-700). Chia Laguna Resort, Sardinia, Italy.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng , J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17* (5), 507-513.

Schmidt, C. M., White, P. B., Waters, J. A., Yiannoutsos, C. T., Cummings, O. W., Baker M, M., . . . Lillemoe, K. D. (2007). Intraductal papillary mucinous neoplasms: predictors of malignant and invasive pathology. *Annals of Surgery, 246* (4), 644-51.

Shahar, Y., & Musen, M. A. (1996). Knowledge-based temporal abstraction in clinical domains. *Artificial Intelligence in Medicine, 8* (3), 267-298.

Shahar, Y., & Musen, M. A. (1993). RÉSUMÉ: A temporal-abstraction system for patient monitoring. *Computers and biomedical research, 26*, 255-255.

Shandiz, M. A., MacKenzie, J. R., Hunt, S., & Anglin, C. (2014). Accuracy of an adjustable patient-specific guide for acetabular alignment in hip replacement surgery (Optihip). *Proceedings of the Institution of Mechanical Engineers, 228(9),* 876-89.

Shi, C., Hruban, R. H., & Klein, A. P. (2009). Familial pancreatic cancer. *Archives of Pathology & Laboratory Medicine, 133* (3), 365-74.

Singh-Grewal, D., Macdessi, J., & Craig, J. (2005). Circumcision for the prevention of urinary tract infection in boys: a systematic review of randomised trials and observational studies. *Archives of Disease in Childhood, 90*, 853-858.

Smith, L., Rindflesch, T., & Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics, 20* (14), 2320-2321.

Sutton, C., & McCallum, A. (2011). An Introduction to conditional random fields. *Foundations and Trends in Machine Learning* .

Sun, Y. K., & Nguyen, A. (2010). Rule-based approach for Identifying assertions in clinical free-text data. *Fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data*. Washington, DC.

Skeppstedt, M. (2010). Negation detection in swedish clinical text. *NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, (pp. 15-21). Los Angeles.

Sohn, S., Wu, S., & Chute, C. G. (2012). Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceeding*, (pp. 1–8).

St. Sauver, J. L., Grossardt, B. R., Leibson, C. L., Yawn, B. P., Melton III, L. J., & Rocca, W. A. (2012). Generalizability of Epidemiological Findings and Public Health Decisions: An Illustration From the Rochester Epidemiology Project. *Mayo Clinic Proceedings , 87* (2), 151-160.

St Sauver, J. L., Grossardt, B. R., Yawn, B. P., Melton, L. J., Pankratz, J. J., Brue, S. M., & Rocca, W. A. (2012). Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. *International Journal of Epidemiology, 41* (6), 1614-24.

Suna, W., Rumshisky, A., & Ozlem Uzuner. (2013). Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics , 46*, S5–S12.

Tanaka, M., Chari, S., Adsay, V., Fernandez-del Castillo, C., Falconi, M., Shimizu , M., . . . Matsuno, S. (2006). International consensus guidelines for management of intraductal papillary mucinous neoplasms and mucinous cystic neoplasms of the pancreas. *Pancreatology , 6* (1-2), 17-32.

Tanaka, M., Fernandez del Castillo, C., Adsay, V., Chari, S., Falconi, M., Jang, J. Y., . . . Yamao, K. (2012). International consensus guidelines 2012 for the management of IPMN and MCN of the pancreas. *Pancreatology , 12* (3), 183-97.

Teufel, S., & Moens, M. (2002). Summarising scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics, 28* (4), 409–445.

Tepper, M., Capurro, D., Xia, F., Vanderwende, L., & Yetisgen-Yildiz, M. (2012). Statistical section segmentation in free-text clinical records. *Proceedings of the Eight International Conference on Language Resources and Evaluation*, (pp. 2001-2008). Istanbul.

The Apache Software Foundation. (n.d.). Apache OpenNLP. Retrieved July 10, 2013, from Apache OpenNLP: http://incubator.apache.org/opennlp/

Tieleman, T., & Hinton, G. (2009). Using fast weights to improve persistent contrastive divergence. *International conference on Machine Learning*, (pp. 1033-1040).

Tieleman, T. (2008). Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. *International conference on Machine Learning (ICML)* (pp. 1064-1071). ACM.

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association, 15* (1), 14-24.

Uzuner, Ö. (2009). Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association , 16*, 561-570.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication Information from clinical text. *Journal of the American Medical Informatics Association , 17,* 514-518.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association , 18* (5), 552-6.

Uzuner, O., Bodnari, A., Shen , S., Forbush, T., Pestian , J., & South , B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. BMC Bioinformatics , 9 (Suppl 11:S9).

Wang, H., Zhao, G., & Yuan, J. (2013). Visual pattern discovery in image and video data: a brief survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery: Data Mining and Knowledge Discovery , 4* (1), 24-37.

Wang, F., Lee, N., Hu, J., Sun, I., & Ebadollahi, S. (2012). Towards Heterogeneous Temporal Clinical Event Pattern Discovery:A Convolutional Approach. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 453-461). Beijing.

Wang, F., Lee, N., Hu, J., Ebadollahi, S., & Laine, A. F. (2013). A Framework for Mining Signatures from Event Sequences and Its Applications in Healthcare Data. *IEEE Transactions to pattern analysis and machine intelligence , 35* (2), 272-285.

Wang, W., Chen, S., Brune, K. A., Hruban, R. H., Parmigiani, G., & Klein, A. P. (2007). PancPRO: Risk assessment for Individuals with a family history of pancreatic cancer. *Journal of Clinical Oncology , 25* (11), 1417-1422.

Weinberg, B. M., Spiegel, B. M., Tomlinson, J. S., & Farrell, J. J. (2010). Asymptomatic Pancreatic Cystic Neoplasms: Maximizing Survival and Quality of Life Using Markov-Based Clinical Nomograms. *Gastroenterology, 138*, 531–540.

Wilson, B. J., Qureshi, N., Santaguida, P., Little, J., Carroll, J. C., Allanson , J., & Raina, P. (2009). Systematic Review: Family History in Risk Assessment for Common Diseases. Ann Intern Med, 151 (12), 878–885.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *HLT/EMNLP*, (pp. 347–54). Vancouver, British Columbia, Canada.

Wu, S., Miller, T., Masanz, J., Coarr, M., Halgrim, S., Carrell, D., & Clark, C. (2014). Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS ONE , 9* (11).

Xu, Y., Hong, K., Tsujii, J., & Chang, E. I.-C. (2012). Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association , 19* (5), 824–832.

Ye, N., Sun, L., Chieu, H. L., & Wu, D. (2009). Conditional random fields with high-order features for sequence labeling. Neural Information Processing System Foundation.

Yoon, P. W., Scheuner, M. T., Peterson-Oehlke, K. L., Gwinn , M., Faucett , A., & Khoury, M. J. (2002). Can family history be used as a tool for public health and preventive medicine? *Genetics in Medicine, 4*, 304 –310.

Zhang, X. M., Mitchell, D. G., Dohke, M., Holland, G. A., & Parker, L. (2002). Pancreatic cysts: depiction on single-shot fast spin-echo MR images. *Radiology, 223* (2), 547-53.

Zhao, D., & Weng, C. (2011). Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. Journal of Biomedical Informatics, 44, 859–868.

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *Journal of biomedical informatics, 46* (6).

CURRICULUM VITAE

Saeed Mehrabi

Education

PhD in Health Informatics May 2009 - Feb 2016

Indiana University, Indianapolis, IN.

MSc in Biomedical Engineering Sep 2004 - May 2007

Tehran University of Medical Sciences, Tehran, Iran

BSc in Biomedical Engineering Sep 1999 - May 2004

Azad University, Tehran, Iran

Work Experience

Informatics Specialist II, Mayo Clinic Sep 2014 - Present

- Member of Clinical Natural Language Processing Team Led by Dr. Hongfang Liu

Data Scientist, AstraZeneca June 2014 - Sept 2014

- Generalizability of RCT to RWE using graph theory

Academic Experience

- Research Assistant, Indiana University May 2009 - May 2014
- Teaching Assistant, Indiana University
  - INFO B581 Health Info Standards And Terms Spring 2014
  - INFO-I 529 Machine Learning Bio-Informatics Spring 2013
  - INFO-I 643 NLP and Text Mining Fall 2012
  - NEWM-N510 Web-Database concepts Fall 2010

Publications

Journals

1. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt HE, Kesterson J, Beesley C, Dexter PR, Schmidt CM, Palakal MJ, Liu H. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. JBI 2015

2. Safarova MS, Mehrabi S, Liu H, Kullo I. An Electronic Phenotyping Algorithm to Rapidly Ascertain Familial Hypercholesterolemia in Healthcare Systems. Journal of Clinical Lipidology 2015; 9(3), 426-427
3. Roch, AM, Mehrabi S, Krishnan, A, Schmidt, HE, Kesterson J, Beesley C, Dexter PR, Palakal, M. and Schmidt CM. Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer. HPB 2014
4. Mehrabi S, Maghsoudloo M, Arabalibeik H, Noormand R, Nozari Y. Application of multi-layer perceptron and radial basis function neural networks in differentiating between chronic obstructive pulmonary and congestive heart failure diseases. Expert Systems with Applications 2009; 36(3): 6956-6959.

Conference Proceedings

1. Mehrabi S, Wang Y, Ihrke D, Liu H. Exploring Gaps of Family History Documentation in EHR for Precision Medicine - A Case Study of Familial Hypercholesterolemia Ascertainment. Accepted for publication AMIA 2016 Joint Summits on Translational Science.
2. Mehrabi S, Sohn S, Li D, Pankratz JJ, Therneau T, St. Sauver JL, Liu H, Palakal M. Temporal Association Discovery of Diagnosis Codes Using Deep Learning. IEEE International Conference on Healthcare Informatics 2015. 408-416
3. Wang Y, Mehrabi S, Mojarad MR, Liu H. Retrieval of Semantically Similar Healthcare Questions in Healthcare Forums. IEEE International Conference on Healthcare Informatics 2015 pp. 517 - 518.
4. Li D, Rastegar M, Elayavilli RK, Wang Y, Mehrabi S, Yu Y, Sohn S, Li Y, Afzal N, Liu H. A Frequency-filtering Strategy of Obtaining PHI-free Sentences from Clinical Data Repository. ACM BIB 2015 Atlanta, GA
5. Mehrabi S, Krishnan A, Roch AM, Schmidt H, Li D, Kesterson J, Beesley C, Dexter P, Schmidt CM, Palakal M, Liu H. Identification of Patients with Family History of Pancreatic Cancer - Investigation of an NLP system Portability. Studies in Health Technology and Informatics Vol.216, Proceedings of the 15th World Congress on Medical and Health Informatics pp 604-608 (Nominated for best student paper award)

6. Mehrabi s, Schmidt CM, Waters JA, Beesley C, Krishnan A, Kesterson J, Dexter P, Al-Haddad MA, Tierney WM, Palakal M. An efficient pancreatic cyst identification methodology using natural language processing. Studies in Health Technology and Informatics Vol.19 MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics pp 822 - 826.
7. Mehrabi S, Krishnan A, Tinsley E, Sligh J, Crohn N, Bush H, DePasquale J, Bandos J, Palakal M. Event Causality Identification Using Conditional Random Field in Geriatric Care Domain. IEEE ICMLA 2013 pp339-343

Posters

1. D Li, Mojarad M Rastegar, Y Li, S Sohn, S Mehrabi, Elayavilli R Komandur, Y Yu, H Liu. A Frequency-based Strategy of Obtaining Sentences from Clinical Data Repository for Crowdsourcing. Studies in health technology and informatics MEDINFO 2015 - Proceedings of the 15th World Congress on Medical and Health Informatics pp.604-608
2. Mehrabi S, Mohammadi I, Kunjan K, Kharrazi H. Effects of data transformation methods on classification of patients diagnosed with Myocardial Infarction. Studies in Health Technology and Informatics Vol. 192 MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics pp 1203-1203
3. Mehrabi S, Kharrazi H, Meaningful use of PHR. The first AMA-IEEE Medical Technology Conference on Individualized Healthcare, March 2010, Washington DC.

Honors and Awards

- Finalist, Healthcare Data Analytics Challenge ICHI 2015
- ACM TAPIA scholarship recipient 2014
- Second place, American Medical Informatics Association Natural Language Processing Working
- Group Doctoral Symposium, Washington DC 2013.
- IEEE International Conference on Health Informatics, NSF Travelling Support for Doctoral Students 2013
- Hayes Research Travel Fund for presentation at AMA-IEEE, 2010

- Second place, Azad University master's degree entrance exam in 2004.
- Third place, Tehran University of medical sciences master's degree entrance exam in 2004

Professional Services

- Student Editorial Board of Journal of Methods of Information in Medicine 2013-2015
- Assistant Editor MedInfo 2015
- Conference Reviewer
    - Medinfo (2013, 2015).
    - AMIA (2012-2015).