MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-604

Identification of Patients with Family History of Pancreatic Cancer - Investigation of an NLP System Portability

Saeed Mehrabi [a,b], Anand Krishnan[b], Alexandra M Roch[c], Heidi Schmidt[c], DingCheng Li[a], Joe Kesterson[d], Chris Beesley[d], Paul Dexter[d], Max Schmidt[c], Mathew Palakal[b], Hongfang Liu[a]

[a] *Department of Health Sciences Research, Mayo Clinic, Rochester, MN*
[b] *School of Informatics and Computing, Indiana University, Indianapolis, IN*
[c] *Department of Surgery, Indiana University, Indianapolis, IN*
[d] *Regenstrief Institute Inc., Indianapolis, IN*

Abstract

In this study we have developed a rule-based natural language processing (NLP) system to identify patients with family history of pancreatic cancer. The algorithm was developed in a Unstructured Information Management Architecture (UIMA) framework and consisted of section segmentation, relation discovery, and negation detection. The system was evaluated on data from two institutions. The family history identification precision was consistent across the institutions shifting from 88.9% on Indiana University (IU) dataset to 87.8% on Mayo Clinic dataset. Customizing the algorithm on the the Mayo Clinic data, increased its precision to 88.1%. The family member relation discovery achieved precision, recall, and F-measure of 75.3%, 91.6% and 82.6% respectively. Negation detection resulted in precision of 99.1%. The results show that rule-based NLP approaches for specific information extraction tasks are portable across institutions; however customization of the algorithm on the new dataset improves its performance.

Keywords:

Natural language processing, Unstructured Information Management Architecture, Family History, Pancreatic cancer.

Introduction

There has been a slow annual increase in the incidence of pancreatic cancer between 2000-2009 worldwide, in contrast to the decrease for most other major cancers. Pancreatic cancer is one of the deadliest cancers, with approximately 73% death rate among patients within the first year of their diagnosis [1]. It is estimated that 46,420 (23,530 men and 22,890 women) will be diagnosed with, and 39,590 (20,170 men and 19,420 women) will die of, cancer of the pancreas in 2014. Pancreatic cancer has several risk factors such as obesity, smoking, and alcohol intake, but its exact causes are not yet known. Screening the general population for early identification of pancreatic cancer is infeasible, and there is no reliable test for its early detection. Screening high-risk populations might be effective in reducing mortality. It is estimated that 10% of pancreatic cancers have a familial basis [2]. One first-degree relative (parents, siblings or children) with pancreatic cancer increases the risk 7-9 fold, and three or more first-degree relatives with pancreatic cancer increase the risk 17-32-fold [3]. Risk is also increased if a first-degree relative diagnosed with pancreatic cancer before age 50 [4].

Another group of patients that are at risk of having pancreatic cancer is patients with pancreatic cysts [5]. In our previous work, we developed natural language processing (NLP) techniques to identify patients with pancreatic cysts from clinical notes [6-8]. In this study, we are focusing on identifying patients with family history of pancreatic cancer using a rule-based algorithm.

Much of the information in clinical notes is in free text format, making it a challenge for secondary use of clinical data. 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 lack 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 [9]. Although, rule-based approaches are cumbersome to implement, they have been widely used in clinical NLP. In this study, we have developed a rule-based method to identify patients with family history of pancreatic cancer and assessed the generalizability of our algorithm on a different institutions than it was originally developed.

Related Work

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” (work weaknesses of others) and “basis” (authors’ work based on the work of others) [10]. Sequential tagging approaches such as Naïve Bayes (NB) and maximum entropy (MaxEnt) models have been used in solving this problem. MaxEnt model of Merity et al., achieved 96.88% F-Score [11]. Another closely related task is the classification of sentences, in abstracts of scientific articles, into separate sections such as introduction, methods, results, and conclusion. Machine learning algorithms such as SVM, Hidden Markov Models (HMM), and Conditional Random Fields (CRF) have achieved accuracies ranging from 90-94.3% [12-14].

In clinical domains, researchers have developed an algorithm called SecTag that uses a combination of NLP techniques, and rules-based and Naïve Bayesian scoring methods to identify section headers [15]. Section header terminology in this work was developed using the Quick Medical Reference (QMR) knowledge base, Logical Observation Identifiers Names and

Codes (LOINC), and various other resources with data models similar to UMLS [16]. 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% [17]. Tepper et al. used two methods: A one-step approach that segmented and classified sections in one step, and a two-step approach that used two different models for section segmentation and classification. In the one-step approach, they used the MaxEnt sequential tagging model to identify if a line was in the beginning, inside, or outside (BIO) of a section category. In the two-step approach, they used again MaxEnt sequential tagging to first label each line with BIO tags, and then used a separate classification algorithm 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 (%) respectively, on three different datasets [18].

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 [19]. Their method achieved higher precision and recall in comparison to a dependency parser based algorithm used in another study [20].

Nearly half of the sentences related to family history were negated. Negation detection has been an inevitable step in processing clinical notes that has attracted much attention [21]. NegEx is one of the most commonly used negation algorithms in clinical NLP [22]. Several other negation identification algorithms (such as NegExpander [23], NegFinder [24], ChartIndex [25], DepNeg [26] and DEEPEN [7]) have also been developed using context-free grammar and dependency parsing to improve negation detection accuracy.

To our knowledge, none of the previous work in family history identification consider all of the steps involved in this task. Friedlin et al. reported sensitivity of 93% and positive predictive value of 97% in extraction of family history, but their method only considers family histories reported under the family history section and not those buried under various other sections in clinical notes [27]. It also doesn't extract exact family members and classify family members as primary, secondary, and unknown relatives. Goryachev et al's work does not examine the negation status of diagnoses found under the family history section [19]. Lewis et al. reported that only 2% of sentences with a family member term were negated compared to 17% of sentences reported under the family history section being negated. Although they stated their interest to identify negation in their future work, they have not analyzed negation in their latest work [20].

PancPro is a Bayesian modeling framework used to assess the pancreatic cancer risk of patients with family history of pancreatic cancer [28]. However, it does not use NLP techniques to extract family history information from clinical notes, so information was collected using a questionnaire.

Materials and Methods

Data Source

The study was approved by the Institutional Review Board (IRB) protocol of each institution separately. Below are the descriptions of each institution dataset.

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 thousand reports for all patients during that month. A detailed description of the dataset has been previously published [8]. The dataset was randomly divided into 60% for training and 40% for testing.

Mayo Clinic

We used Mayo cancer registry data to obtain a list of patients with pancreatic cancer. There were a total of 3,573 patients in the registry, out of which 2,923 had a family history section in their clinical notes. Clinical notes for those patients were extracted from the Mayo Clinic data repository, and text from the family history section of those notes form the second data set.

Methods

Clinical reports are organized into sections with headers such as "Physical Examination," "Medication," "Family History." Usually a patient's family history is reported under the family history section of the narrative reports. However, this is not always the case. It is sometimes mentioned in the patient's history, diagnosis, or other sections. Based on this understanding, we divided family history identification into two parts: In the first part, the patient's family history, which is reported under the family history section, was identified. In the second part, the family history section was removed from the clinical note and any mention of a family history in other sections was identified. The first part consisted of three sub-parts: 1) section header detection, 2) family members and diagnoses identification, and 3) relation discovery between family members and diagnoses.

Section header detection

A rule-based algorithm based on the SecTag terminology was developed to identify the clinical notes sections for the IU data set. While at Mayo, clinical notes are CDA 1.0 compliant, wherein sections have been codified. Although SecTag terminology is a large-scale effort to assemble an exhaustive list of terminologies used as section headers of clinical notes, due to lack of standard and universal convention there are still terms used in other institutions that are not found in the SecTag terminology list. For instance, section headers such as "Past Medical, Social, Family History" and "Social and Family History" were used in IU clinical notes, but were not available in the SeTag terminology. We added these terms to our dictionary list to identify family history sections.

Family member and diagnosis identification

After a family history section was identified, sentences reported under this section were detected using Ytex sentence detector [29]. A list of keywords indicating pancreatic cancer concepts and family members were collected using UMLS metathesaurus [30] and manual review of a random set of clinical notes. This dictionary was then used to identify family member and pancreatic cancer concepts within a sentence.

Relation Discovery

Associating family member with pancreatic cancer in a sentence with only one family member is trivial, e.g.:

> 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 (e.g. 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 the terms "brother" and "pancreas cancer" in the sub-sentence *"and her brother who died of pancreas cancer at the age of xxx"*

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

In order to identify family history of pancreatic cancer mentioned other in sections of the report, the family history section was removed and the same algorithm was applied where at least one family member must be mentioned.

An NLP system using UIMA framework, shown in Figure 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 bodies were then used as an input for 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 [31]. DEEPEN improves the NegEx algorithm by double-checking the negation status of concepts using a nested chain of dependency relations between negation words and desired concepts within a sentence. 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) was found in the sentence, and their negation status was stored in a database.

Figure 1- Analysis engines developed in the UIMA pipeline to identify patients with family history of pancreatic cancer.

Results

Table 1 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 a non-blood relative (e.g., wife or husband) was mentioned, it was considered as irrelevant.

Table 1- 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 2). Precision is defined as the number of correct instances over the total of correct and incorrect instances. As shown, the performance of the system has been consistent across the two institutions.

Table 2- Mayo dataset evaluation.

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

In order to ensure that we did not miss any patient with family history of pancreatic cancer, 100 reports were selected randomly and manually reviewed. Table 3 shows the result of our algorithm to incorporate missing patterns in these reports:

Table 3- Result of Mayo 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 the 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 a false positive.

There were total of 268 patients with a family history of pancreatic cancer out of 3,573 patients with pancreatic cancer in the Mayo Clinic's data set. Figure 2 shows the number of patients identified with first, second, or third degree relative.

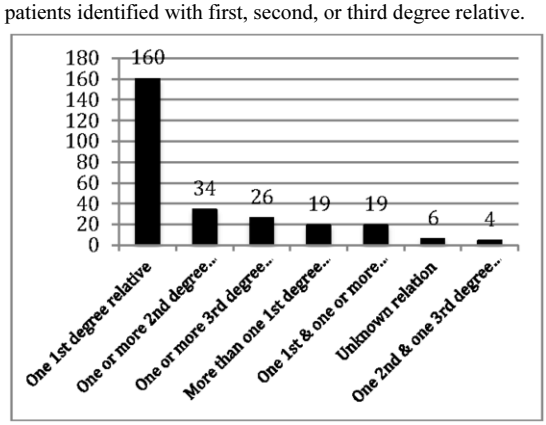

Figure 2- Number of identified patients with one or more 1st, 2nd, or 3rd degree relative.

Table 4- Results of family member identification.

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

Discussion

In this work, we have developed an NLP system in a UIMA framework with multiple analysis engines, including section segmentation, negation detection, and relation discovery to identify patients with a family history of pancreatic cancer from clinical notes.

We have developed our system on an IU dataset. The 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 low number of patients at IU compared to the Mayo Clinic dataset. Clinical notes at the Mayo Clinic are CDA 1.0 compliant; therefore, section detection developed at IU was not used for the 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.

We can classify the errors in our system as follows:

1. Sentence Detector

Sentences "C" and "D" show two examples of instances where sentence detector fails to identify the correct boundary of a sentence and therefore the pancreatic cancer concept was improperly assigned to multiple family members.

C) "Father deceased xx-xx
colon polyp pancreatic cancer heart disease alcohol abuse depression
mother
mother alive
heart disease asthma stroke/tia high cholesterol arthritis depression
brothers
2 brothers alive 1 brother deceased
colon polyp asthma
sisters
2 sisters alive
osteoporosis
famhx updates (cvi)
high blood pressure – yes"

D) "His brother died of pancreatic cancer at the age of xx. A sister has a history of breast cancer."

2. Complicated Family Relations

Sentences "*E*", "*F*", and "*G*" show examples of family relation where 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 "E" pancreatic cancer was related to mother, sister, and granddaughter. Sentence "*H*" shows an example where semantic inference is needed to infer that pancreatic cancer is related to the mother.

E) "Recently she found out about her mother's sister's granddaughter who was diagnosed with pancreatic cancer at the age of xx."

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

G) "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 xx."

H) "Her son (our patient) found her deceased about xx 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."

3) 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 "I". Each family relation term was then associated with the pancreatic cancer concept within the sub-sentence. In sentence "I," "*pancreatic cancer*" is associated with "*paternal grandfather,*" but it failed to associate the "mother" and "father" with the pancreatic cancer concept in the sub-sentence *"his mother, father,"* because there is no pancreatic cancer concept in the sub-sentence.

I) "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 (see sentence "J"). Our current system does not handle co-referencing.

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

Conclusions

Pancreatic cancer is referred to as silent killer due to its few sign and symptoms until it is in well-advanced cancer stages. Screening the general population for pancreatic cancer is not feasible because of its low incidence and the lack of effective screening tests. Pancreatic cyst and family history of pancreatic cancer represent two windows of opportunity for early detection of pancreatic cancer. We have developed a rule-based algorithm to identify patients with a family history of pancreatic cancer retrospectively from their clinical records. Development of clinical NLP system requires resources, such as domain experts to develop guidelines, nurse abstractors to create gold standards, and researchers/programmers to develop and analyse the system. Although rule-based methods 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.

Future steps involve refinement of the family relation discovery rules, especially regarding the sentence detection algorithm. A risk stratification method will also be developed based on the number and degree of family relations to assess patients' risk of having cancer and a surveillance strategy will be designed to follow up with patients according to their risk.

Acknowledgments

This work was made possible by funding from NIH R01GM102282, R01LM11369, R01LM11829, and R01 LM011934.

References

[1] American Cancer Society. Cancer Facts & Figures American Cancer Society, Atlanta, 2014.

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

[3] Shi C , Hruban RH, and Klein AP. Familial pancreatic cancer. Arch Pathol Lab Med 2009: 133(3): 365-74.

[4] Klein AP et al. Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. Cancer Res 2004: 64(7): 2634-8.

[5] Pandol S et al. Epidemiology, risk factors, and the promotion of pancreatic cancer: role of the stellate cell. J Gastroenterol Hepatol 2012: 27: 127–134.

[6] Al-Haddad MA et al. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. International Hepato-Pancreato-Biliary Association 2010: 12(10): 688-95.

[7] Mehrabi S et al. An efficient pancreatic cyst identification methodology using natural language processing. Stud Health Technol Inform 2013: 192: 822-6.

[8] Roch AM et al. Automated pancreatic cyst screening using natural language processing: a new tool in pancreatic cancer early detection. HPB doi: 10.1111/hpb.12375.

[9] Liu H et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013:149–153.

[10] Teufel S and Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics 2002: 28(4): 409-445.

[11] Merity S, Murphy T, and Curran JR. Accurate Argumentative Zoning with Maximum Entropy models. Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL), 2009.

[12] Lin J, Karakos D, Demner-Fushman D, and Khudanpur S. Generative content models for structural analysis of medical abstracts. HLT-NAACL BioNLP Workshop. 2006:65–72.

[13] Hirohata KK, Okazaki N, Ananiadou S, and Ishizuka M. Identifying sections in scientific abstracts using conditional random fields. International Joint Conference on Natural Language Processing 2008: 381–388.

[14] McKnight L and Srinivasan P. Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc. 2003:440–444.

[15] Denny JC et al. Evaluation of a method to identify and categorize section headers in clinical documents. JAMIA 2009;16(6):806-15.

[16] Denny JC, Miller RA, Johnson KB, and Spickard A. Development and evaluation of a clinical note section header terminology. AMIA Annu Symp Proc. 2008:156-60.

[17] Li Y, Gorman SL, Elhadad N. Section classification in clinical notes using supervised hidden markov model. ACM International Health Informatics Symposium. 2010: VOL???: 744–750.

[18] Tepper M, Capurro D, Xia F, Vanderwende L, and Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. Eight International Conference on Language Resources and Evaluation. 2012:2001-2008.

[19] Sergey Goryachev, Hyeoneui Kim, and Qing Zeng-Treitler. Identification and Extraction of Family History Information from Clinical Reports. AMIA Annu Symp Proc. 2008:247–251.

[20] Lewis N, Gruhl D, and Yang H. Dependency Parsing for Extracting Family History. IEEE HISB 2011: VOL????: 237-242.

[21] Wu S et al. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. PLoS ONE 2014: 9(11): pages?????.

[22] Chapman WW, Bridewell W, Hanbury P, Cooper GF, and Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 2001: 34(5): 301–310.

[23] Aronow DB, Fangfang F, and Croft WB. Ad Hoc Classification of Radiology Reports. J Am Med Inform Assoc 1999: 6(5) Sep-Oct: 393–411.

[24] Mutalik PG, Deshpande A, and Nadkarni P. Use of general purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls. J Am Med Inform Assoc 2001: 8: 589-609.

[25] Huang Y and Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc 2007: 14: 304-311.

[26] Sunghwan S, Stephen W, and Chute CG. Dependency Parser-based Negation Detection in Clinical Narratives. AMIA Summits Transl Sci Proc. 2012:1–8.

[27] Friedlin J and McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925.

[28] Wang W et al. PancPRO: risk assessment for individuals with a family history of pancreatic cancer. J Clin Oncol 2007: 25(11): 1417-22.

[29] Garla V et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011: 18(5) Sep-Oct: 614-20.

[30] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004: 32 Jan: 267-70.

[31] Mehrabi S et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx JBI 2015 doi:10.1016/j.jbi.2015.02.010.

[32] M C de Marneffe , B MacCartney , and C D Manning , "Generating Typed Dependency Parses from Phrase Structure Parses," in LREC, 2006.

[33] Robert J Carroll et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012: 19: e162-169.

Address for correspondence

Mayo Clinic, 200 1st SW, Rochester MN 55905, Tel: 507-773-0057
Email address: liu.hongfang@mayo.edu (Dr. Hongfang Liu).

535 W. Michigan St, Indianapolis, IN, 46202, USA

Tel: 317-278-7689, Fax: 317-278-7669 (Dr. Mathew Palakal)