Application of Natural Language Processing to VA Electronic Health Records to Identify Phenotypic Characteristics for Clinical and Research Purposes

Adi V. Gundlapalli, MD, PhD, MS[1,2], Brett R. South, MS[1,3], Shobha Phansalkar, MS, RPh, PhD[2], Anita Y. Kinney, PhD, RN[1,3], Shuying Shen, MStat[1,3], Sylvain Delisle, MD, MBA[4], Trish Perl, MD, MSc[5], Matthew H. Samore, MD[1,2,3]

[1]Departments of Internal Medicine and [2]Biomedical Informatics, University of Utah School of Medicine, [3]Salt Lake VA Health Care System, Salt Lake City, Utah, [4]Maryland VA Health Care System and University of Maryland School of Medicine, [5]Johns Hopkins Medical Institutions and University, Baltimore, Maryland

Informatics tools to extract and analyze clinical information on patients have lagged behind data-mining developments in bioinformatics. While the analyses of an individual's partial or complete genotype is nearly a reality, the phenotypic characteristics that accompany the genotype are not well known and largely inaccessible in free-text patient health records. As the adoption of electronic medical records increases, there exists an urgent need to extract pertinent phenotypic information and make that available to clinicians and researchers. This usually requires the data to be in a structured format that is both searchable and amenable to computation. Using inflammatory bowel disease as an example, this study demonstrates the utility of a natural language processing system (MedLEE) in mining clinical notes in the paperless VA Health Care System. This adaptation of MedLEE is useful for identifying patients with specific clinical conditions, those at risk for or those with symptoms suggestive of those conditions.

Introduction

In the era of genome wide association studies and large bioinformatics databases, the limiting factor for the discovery of newer associations between diseases and genes seems only to be the availability of more comprehensive micro-arrays or SNPs, the computational tools to analyze the vast amounts of data and funding. On the other hand, phenotypic databases are in their infancy [1, 2] and detailed data on human phenotypes are locked largely in unstructured, free-text clinical notes which are difficult to extract and analyze. Over the past decade, there have been excellent attempts to data-mine selected domains of the electronic medical record, starting with a natural language processing (NLP) system called MedLEE [3-5] that looked at radiology and pathology notes. Others have adapted MedLEE and NLP systems to look at other clinical notes (reviewed by Lussier and Liu [2]) including the development of BioMedLEE to enable coding phenotypes from the scientific literature. Significant advances in mining clinical information in the future are expected especially with projects such as i2b2 [6].

To supplement the efforts to extract clinical information from electronic medical records for the benefit of clinicians, genetic and other biomedical researchers, we have taken a low-cost, high-yield approach to specifically answer the following questions: (1) Who among your patient base has the diagnosis of a specific clinical condition or disease as determined by a clinician? (2) Who is at risk to develop a specified condition? And (3) Who has symptoms that are compatible with or suggest a particular condition? Using inflammatory bowel disease (IBD) as an example, this study describes the adaptation of MedLEE to the ambulatory care notes from two large VA Health Care Systems and demonstrates its utility in identifying patients with IBD.

Setting

This study was carried out at the Baltimore VA Health Care System in Baltimore, Maryland and the Salt Lake VA Health Care System in Salt Lake City, Utah. Both sites serve as referral centers for a large patient base of veterans (nearly 90,000) in Maryland, Utah and surrounding states. The electronic medical

record in the VA Health Care System is one of the most comprehensive in the US and truly paperless [7].

Methods

This was a retrospective study that analyzed a random sample of patients presenting to the outpatient clinics at the two VA systems during the period October 1, 2003 to March 31, 2004. During this 6-month period, there were a total of 253,818 ambulatory care visits to the two sites. The random sample of 15,377 unique patient visits and associated note corpus of 76,500 clinical notes were representative of patient encounters from a variety of healthcare settings including primary care, specialty clinics and the emergency department. *For the purposes of this study, phenotypic characteristics were limited to the diagnoses of the patient as indicated by a clinician in their note, the symptoms elicited from the patient during their visit and associated elements in the history such as past medical history, family history, mention of colonoscopy in the note and genetic testing.*

Inflammatory bowel disease includes the genetically complex diseases of Crohn's and ulcerative colitis. A patient was determined to have a reference standard diagnosis of Crohn's disease or ulcerative colitis if (a) At any time during this period the patient had a visit associated with an ICD-9 code for Crohn's disease (555.x) or ulcerative colitis (556.x); or (b) The keywords Crohn's, Ulcerative Colitis, inflammatory bowel disease or "IBD" appeared in the clinical note as detected by simple string searching coupled with a negation algorithm called NegEx [8] adapted for VA note types; and with either (a) or (b), the electronic record was reviewed by a physician to verify the diagnosis as IBD and the symptoms were not related to an acute infectious gastrointestinal illness of less than 7 days duration.

The free text clinical notes were then processed using MedLEE which is a natural language processing algorithm that employs a semantic lexicon and grammar to extract information from the text of electronic note documents (3). Words or search strings are mapped to the semantic lexicon containing concepts from the Unified Medical Language System (UMLS) and assigned a concept unique identifier (CUI), semantic category and concept modifier. Semantic categories include problems, procedures, medications, findings etc. Concepts modifiers include negations (certainty), temporality (status), change, degree, etc. Additionally, MedLEE is capable of detecting medical synonyms and abbreviations. MedLEE was originally developed to process radiographic and pathology reports, but has been used to process a diverse range of clinical texts [9, 10]. A wide range of output types are available from the MedLEE processor including plain HL7, markup, line, or XML.

XML output from the NLP system was analyzed to identify relevant semantic concepts mapped to the UMLS and CUI codes, and concept modifiers useful to elucidate specific phenotypic information, family history information, or patients showing clinical manifestation of inflammatory bowel disease with symptoms such as diarrhea and abdominal pain (Figure). We also report the same information identified using string matching and the adapted negation algorithm [8]. The accuracy of case detection of the different methods in terms of sensitivity (recall) and PPV (precision) in identifying inflammatory bowel disease from the electronic medical record was determined using standard statistical methods.

The study was reviewed and approved by the Institutional Review Boards of all participating institutions.

Results

The reference standard case finding identified 50 patients meeting ICD-9 code criteria for IBD (sample prevalence: 0.33%), and 202 patients identified by string matching and negation algorithm (sample prevalence: 1.31%). Final arbitrated chart review by a physician identified 91 patients (sample prevalence 0.6%) with IBD.

The MedLEE system identified a total of 183 patients with concepts that mapped to IBD. Sensitivity (recall) and specificity based on MedLEE identifying concepts for IBD were 86% with 95% confidence intervals (CI) of 77 – 92 and 99% (95% CI 99-99) respectively [11]. The precision (positive predictive value) was 43% and negative predictive value was 100% (Table 1). The area under the ROC curve was 0.9 for detection by MedLEE.

In analyzing the MedLEE XML output semantic categories (Table 2), specific symptoms suggestive of IBD included diarrhea in 29% of patients with a reference standard diagnosis and abdominal pain (21%). Other symptoms of vomiting and fever were less frequent. Family history information was documented among only 8% of patients with IBD, and mention of colonoscopy was noted for 17% of patients. Smoking history as a possible risk factor was identified in 57% of patients with a reference

standard diagnosis of IBD. Concepts denoting genetic testing were not identified by the MedLEE system (Table 2).

Though ICD-9 coding and the NegEx algorithm were also part of the case finding methodology, it is of interest to note the poor sensitivity of ICD-9 coding in detecting patients with a history of IBD (27%, 95% CI 19-38) and the high sensitivity of the NegEx method (85%, 95% CI 76-91). The specificities were 100 and 99, while their positive predictive values were 50 and 38 respectively. The area under the ROC curve was 0.64 for ICD-9 detection for IBD and 0.9 for NegEx. Additionally, the NegEx algorithm coupled with a list of terms identifying notation methods of family history and colonoscopy unique to VA notes was able to identify family history documentation among 26% of IBD patients, and colonoscopy documentation in 40% of patients with IBD.

Limitations

As the number of patients in this study was large (15,377) and chart review is labor-intensive, we identified the reference standard cases of IBD using a combination of case finding using ICD-9 coding, string searches and manual review of records. The recall (sensitivity) calculated in this case is the *maximum recall* rather than the true recall as the statistic is calculated from an enriched sample as opposed to a random sample. Though the prevalence of IBD is this sample is comparable to population estimates, there is a possibility that we did not capture all the cases. Review of potential reference standard cases was performed by an internist and it is possible that a specialist's review would have revealed other cases. Finally, we have applied these methods to one disease condition and validation studies must be conducted across a range of diagnoses and conditions using different data sets.

Conclusions

Extraction of pertinent information from free text clinical notes presents a challenge in terms of unstructured writing with variability between authors and health care settings. Clinical notes associated with routine health care encounters are often unstructured and in free-text format. Nevertheless, these notes contain detailed information on patients that goes beyond the ICD-9 diagnosis and attempts to reliably extract phenotypic data from these records must continue.

We have demonstrated that a relatively simple case finding method based on string matching for specific keywords coupled with an adapted negation algorithm and information extracted by a more complex NLP system can offer insights into the electronic clinical note. We have used this method to identify patients with a particular procedure, history, symptom, risk factor or condition. Though this study focused on inflammatory bowel disease, the MedLEE system can be easily adapted to accommodate other diseases and conditions. While large scale efforts are underway to provide structure to phenotypic databases and attempt integration with genotypic data, there is a place for NLP-based methods to mine the wealth of clinical information for both clinicians and researchers that is generalizable and adaptable to other sites and situations.

As noted above and by others, case finding by ICD-9 coding alone is not sufficient to reliably identify patients with a particular disease or risk factors [12]. Coding that is meant specifically for billing purposes does not usually capture the nuances of phenotypic characteristics such as past medical history, family history, genetic testing or known risk factors for a disease.

Future Directions

It is envisioned that these methods will be further validated using other conditions of interest and a more comprehensive case finding algorithm in conjunction with subject matter experts. With appropriate ethical and legal safeguards, these results will be offered to investigators to identify potential patients for genetic and other biomedical research to bolster traditional recruiting efforts.

Further refinements to the MedLEE lexicon are planned to identify genetic testing, past medical history and other risk factors for disease using the methods described by Rindflesch and colleagues [13]. Further modifications to allow us to differentiate a patient with a history of colonoscopy versus those for whom it has been recommended for screening for IBD will also be considered. NLP methods can also be used to identify control patients without characteristics of a particular disease.

A second area that would benefit from such text mining would be quality improvement where detailed clinical information may provide improved measurements for quality indicators. Finally, these methods have a place in surveillance activities including patient safety, adverse events and bio-surveillance for existing and emerging infections.

Acknowledgements

The authors gratefully acknowledge funding from the CDC for the BioSense Evaluation Grant (Johns Hopkins Medical Institutions and University) and the Center of Excellence in Public Health Informatics (University of Utah School of Medicine).

References

1. Freimer N, Sabatti C. The human phenome project. Nat Genet 2003; 34:15-21.
2. Lussier YA, Liu Y. Computational approaches to phenotyping: High-throughput phenomics. Proc Am Thorac Soc 2007; 4:18-25.
3. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000:270-4.
4. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004; 11:392-402.
5. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994; 1:161-74.
6. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the i2b2 project. AMIA Annu Symp Proc 2006:931.
7. Department of Veterans Affairs. Vista monograph. Washington DC: Department of Veterans Affairs, 2005:146.
8. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001; 34:301-10.
9. Sager N, Lyman M, Nhan NT, Tick LJ. Medical language processing: Applications to patient data representation and automatic encoding. Methods Inf Med 1995; 34:140-6.
10. Mendonca EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005; 38:314-21.
11. Matrix of reference standard cases versus those detected by MedLEE

	Reference Standard		
	(+)	(-)	
MedLEE (+)	78	105	**183**
MedLEE (-)	13	15181	**15194**
	91	**15286**	15377

12. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors. Med Care 2005; 43:480-5.
13. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003; 36:462-77.

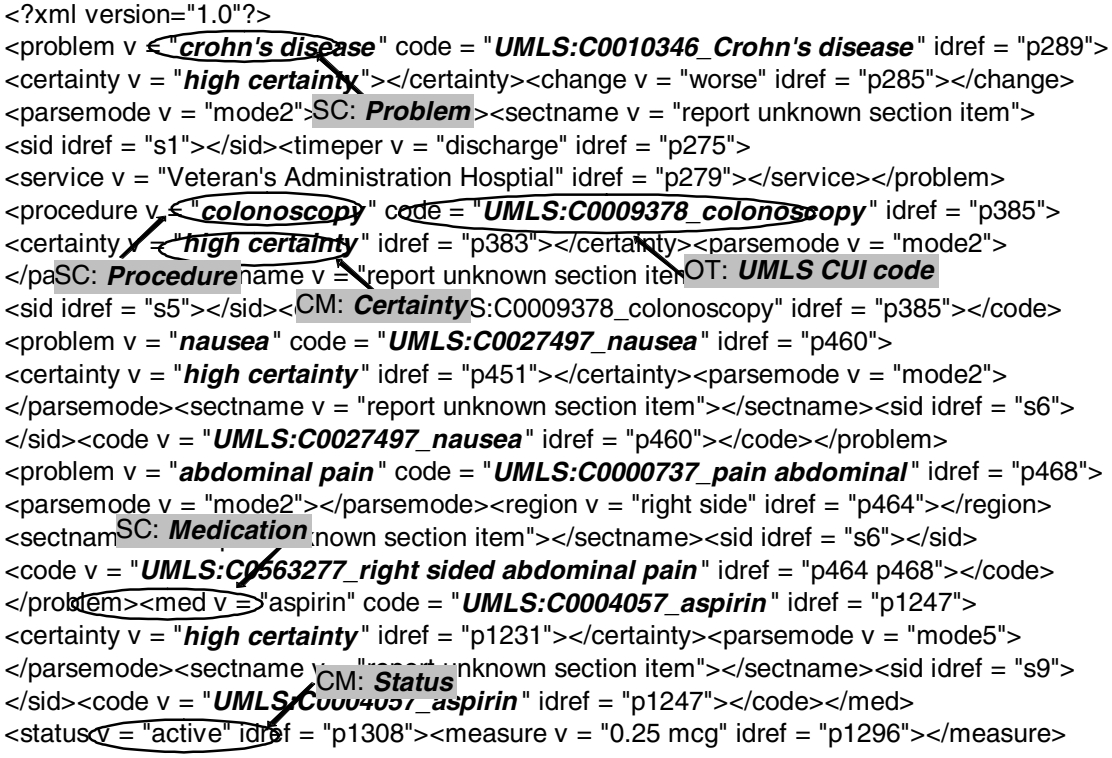
Table 1. Test characteristics of various algorithms for the detection of inflammatory bowel disease

Case detection model	Sensitivity (Recall) (95% CI)	Specificity (95% CI)	Positive Predictive Value (Precision) (95% CI)	Area Under the Receiver Operating Characteristic Curve (ROC) (95% CI)
ICD-9 Alone	27 (19, 38)	100 (100, 100)	50 (36, 65)	0.64 (0.6, 0.7)
NegEx	85 (76, 91)	99 (99, 99)	38 (31, 45)	0.9 (0.88, 0.96)
ICD-9 OR NegEx	100 (96, 100)	99 (99, 99)	40 (33, 46)	0.99 (0.99, 0.99)
MedLEE	86 (77, 92)	99 (99, 99)	43 (35, 50)	0.9 (0.89, 0.96)

Table 2. MedLEE output semantic category analyses for the detection of inflammatory bowel disease

Semantic category	Number (%) Total = 91 reference standard cases of IBD
Procedures	
Colonoscopy	15 (16)
Any endoscopy	1 (1)
Findings	
Family history	7 (8)
Genetic testing	0 (0)
Symptoms	
Abdominal pain	19 (21)
Diarrhea	26 (29)
Vomiting	9 (10)
Fever	14 (15)
Risk factors	
Smoking	52 (57)

Figure MedLEE XML Output: Semantic categories and Concept Modifiers



Legend
SC = Semantic category
CM = Concept modifier
OT = Output type