Using NLP on VA Electronic Medical Records to Facilitate Epidemiologic Case Investigations

Adi V. Gundlapalli, MD, PhD, MS[1,2,6], Brett R. South, MS[1,2], Wendy W. Chapman, PhD[3], Shobha Phansalkar, MS, RPh, PhD[1,2], Shuying Shen, MStat[1,2], Sylvain Delisle, MD, MBA[4], Trish Perl, MD, MSc[5],Matthew H. Samore, MD[1,2,6]

[1]VA Salt Lake City Health Care System, [2] Internal Medicine, University of Utah School of Medicine, [3]Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, [4]VA Maryland Health Care System and University of Maryland School of Medicine, [5]Johns Hopkins Medical Institutions and University, [6] Biomedical Informatics, University of Utah School of Medicine

Objective

To identify epidemiologically important factors such as infectious disease exposure history, travel or specific variables from unstructured data using natural language processing (NLP) methods.

Background

A major goal of biosurveillance is the timely detection of an infectious disease outbreak. Once a disease has been identified, another very important goal is to find all known cases of the disease to assist public health investigators. NLP systems may be able to assist in identifying epidemiological variables and decrease time-consuming manual review of records.

Methods

The setting was the Baltimore and Salt Lake City VA Health Care Systems that together care for 90,000 patients every year. We set out to respond to a hypothetical communication from the local health department asking us to assist them in identifying patients that may have been ill from respiratory illnesses circulating in the community. Risk factors for these illnesses were alcohol or drug abuse, smoking, homelessness and travel to SE Asia. This study follows a two-stage surveillance approach to identify patients in this hypothetical targeted surveillance. In the first stage, we used structured data and BioSense and ESSENCE ICD 9 codes on records during the study period October 2003 – March 2004. The cohort was further refined by using a text-based classifier using string matching for ILI concepts from the case definition that were mapped to the Unified Medical Language System (UMLS) coupled with a negation detection algorithm called NegEx[1] which was used as such with no modifications. In the second stage, we used an NLP system called MedLEE[2] to identify patients with specific clinical and epidemiologic features of interest, Results were compared to a reference standard determined by manual review of the records.

Results

Case finding based on structured data identified 1,394 patients (sample prevalence: 9%), The NegEx negation algorithm coupled with string matching identified 1,064 patients (sample prevalence: 7%). Final physician arbitrated chart review identified 280 patients (sample prevalence 1.8%) with ILI. For cases flagged by the text classifier, and processed by MedLEE concepts for pneumonia, fever and respiratory infection were identified in unstructured sources in 12%, 44% and 79% of cases. Epidemiologic factors for alcohol abuse, drug abuse, and smoking status were identified in 4%, 29%, and 45% of cases processed by MedLEE. A reference to homelessness was found in 11% of cases. MedLEE was not designed to identify travel history. In comparison, clinical factors based on structured BioSense ICD9 codes for pneumonia, fever and respiratory infections were present in <1%, 1%, and 32% of cases. Epidemiologic factors based on ICD9 codes for alcohol abuse, drug abuse, and smoking status were present in 3%, 4%, and 1% of cases.

Conclusions

Our pilot study shows that NLP-based methods are useful for identifying important patient-related features from free-text electronic medical records to assist in public health investigations as compared to structured data sources.

References

1. 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 Oct;34(5):301-10.
2. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270-4.

Further information: adi.gundlapalli@hsc.utah.edu