OJPHI

ISDS 2016 Conference Abstracts

Automated Classification of Alcohol Use by Text Mining of Electronic Medical Records

Lisa Lix*[1, 2], Sree Nihit Munakala[3] and Alexander Singer[1]

[1]Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada; [2]George and Fay Yee Centre for Healthcare Innovation, Winnipeg, MB, Canada; [3]Birla Institute of Technology and Science, Pilani - Hyderabad Campus, Hyderabad, India

Objective

The research objective was to develop and validate an automated system to extract and classify patient alcohol use based on unstructured (i.e., free) text in primary care electronic medical records (EMRs).

Introduction

EMRs are a potentially valuable source of information about a patient's history of health risk behaviors, such as excessive alcohol consumption or smoking. This information is often found in the unstructured (i.e., free) text of physician notes. It may be difficult to classify and analyze health risk behaviors because there are no standardized formats for this type of information[1]. As well, the completeness of the data may vary across clinics and physicians. The application of automated classification tools for this type of information could be useful for describing patterns within the population and developing disease risk prediction models.

Natural Language Processing (NLP) tools are currently used to process EMR free text in an automated and systematic way. However, these tools have primarily been applied to classify information about the presence or absence of disease diagnoses[2]. The application of NLP tools to health risk behaviors, particularly alcohol use information from primary care EMRs, has thus far received limited attention.

Methods

Study data were from the Manitoba regional network of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) for the period from 1998 to 2016. CPCSSN is a national primary care surveillance network for chronic diseases comprised of 11 regional networks with publicly funded healthcare systems. Currently, a total of 53 clinics and more than 260 physicians provide data to CPCSSN in Manitoba. We classified each record based on unstructured text from physician notes into the following mutually exclusive categories: current drinker, not a current drinker, and unknown[1]. A standardized de-identification process was applied to each record prior to applying an NLP tool to the data.

Text classification used a support vector machine (SVM) applied to both unigrams (i.e., single words) and mixed grams (i.e., unigrams, and pairs of words known as bigrams) from a bag-of-words model in which each record is quantified by the relative frequency of occurrence of each word in the record[3]. The training dataset for the SVM was comprised of 2000 records classified by two primary care physicians. These physicians were initially trained using an independent sample of 200 EMR text strings containing specific reference to alcohol use.

Cohen's kappa statistic, a chance-adjusted measure, was used to estimate agreement. Internal validation of the SVM was conducted using 10-fold cross-validation techniques. Model performance was assessed using recall and precision statistics, as well as the F-measure statistic, which is a function of their average. All analyses were conducted using the R open-source software package.

Results

A total of 57,663 unique records were included in the study. The estimate of the kappa statistic for the physician training phase was 0.98, indicating excellent agreement. Subsequent classification of the training dataset by the physicians resulted in 1.7% of records assigned as not a current drinker, 16.8% as current drinker, and 81.5% as unknown. Average estimates of recall for the 10 validation folds using unigrams were 0.62 for not current drinkers, 0.86 for current drinkers, and 0.98 for the unknown category. Average estimates of recall using mixed grams were 0.48, 0.84, and 0.97, respectively. Estimates of precision were higher with mixed grams than unigrams for the not currently drinking category, but the opposite was true for the current drinker category. There was no difference in precision between the two methods for the unknown category. The F-measure statistic was higher for classification of current drinkers using unigrams (0.89) than mixed grams (0.83), although the differences for the unknown category were negligible (0.98 versus 0.97). Application of the SVM with unigrams to the entire dataset resulted in 15.3% of records classified as current drinkers, 2.0% classified as not current drinkers, and 82.7% as unknown.

Conclusions

This study developed an automated system to classify unstructured text about alcohol consumption into mutually-exclusive alcohol use categories. However, we found that only a small percentage of primary care records contained documentation about alcohol consumption, which limits the utility of the automated tool and the data source for disease risk prediction or alcohol use prevalence estimation[1]. While our automated approach is useful for processing existing EMR data, systematic documentation of alcohol consumption will benefit from standardized entry fields and terms to produce clinically meaningful information that will improve the understanding of health risk behaviors in primary care populations.

Keywords

automated analytics; unstructured text; support vector machine; health risk behaviors; primary care surveillance

Acknowledgments

We acknowledge the Mitacs Globalink program for providing project funding support for SNM and the Manitoba regional network of the Canadian Primary Care Sentinel Surveillance Network for providing data access.

References

1. Torti J, Duerksen K, Forst B, Salvalaggio G, Jackson D, Manca D. Documenting alcohol use in primary care in Alberta. Can Fam Physician 2013 Oct;59(10):1128.
2. Wang Y, Chen ES, Pakhomov S, Arsoniadis E, Carter EW, Lindemann E, Sarkar IN, Melton GB. Automated extraction of substance use information from clinical texts. AMIA Annu Symp Proc. 2015 2015:2121–2130.
3. Figueroa RL, Flores CA. Extracting information from electronic medical records to identify obesity status of a patient based on comorbidities and bodyweight measures J Med Syst. 2016 Aug;40(8):191.

***Lisa Lix**
E-mail: lisa.lix@umanitoba.ca