ISDS 2012 Conference Abstracts

Extracting Surveillance Data from Templated Sections of an Electronic Medical Note: Challenges and Opportunities

Adi Gundlapalli*[1, 2], Guy Divita[1, 2], Marjorie Carter[1, 2], Shuying Shen[1, 2], Miland Palmer[1], Tyler Forbush[1, 2], Brett South[1, 2], Andrew Redd[1, 2], Brian Sauer[1, 2] and Matthew Samore[1, 2]

[1]VA Salt Lake City Health Care System, Salt Lake City, UT, USA; [2]Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA

Objective

To highlight the importance of templates in extracting surveillance data from the free text of electronic medical records using natural language processing (NLP) techniques.

Introduction

The main stay of recording patient data is the free text of electronic medical records (EMR). While stating the chief complaint and history of presenting illness in the patients 'own words', the rest of the electronic note is written by the provider in their words. Providers often use boiler-plate templates from EMR pull-downs to document information on the patient in the form of checklists, check boxes, yes/no and free text responses to questions. When these templates are used for recording symptoms, demographic information or medical, social or travel history, they represent an important source of surveillance data [1]. There is a dearth of literature on the use of natural language processing in extracting data from templates in the EMR.

Methods

A corpus of 1000 free text medical notes from the VA integrated electronic medical record (CPRS) was reviewed to identify commonly used templates. Of these, 500 were enriched for the surveillance domain of interest for this project (homelessness). The other 500 were randomly sampled from a large corpus of electronic notes. An NLP algorithm was developed to extract concepts related to our target surveillance domain. A manual review of the notes was performed by three human reviewers to generate a document-level reference standard that classified this set of documents as either demonstrating evidence of homelessness (H) or not (NH). A rule-based NLP algorithm was developed that used a combination of key word searches and negation based on an extensive lexicon of terms developed for this purpose. A random sample of 50 documents each of H and NH documents were reviewed after each iteration of the NLP algorithm to determine the false positive rate of the extracted concepts.

Results

The corpus consisted of 48% H and 52% NH documents as determined by human review. The NLP algorithm successfully extracted concepts from these documents. The H set had an average of 8 concepts related to homelessness per document (median 8, range 1 to 34). The NH set had an average 2 concepts (median 1, range 1 to 13)". Thirteen template patterns were identified in this set of documents. The three most common were check boxes with square brackets, Yes/No and free text answer after a question. Several positively and negatively asserted concepts were noted to be in the responses to templated questions such as "Are you currently homeless: Yes or No"; "How many times have you been homeless in the past 3 years: (free text response)"; "Have you ever been in jail? [Y] or [N]"; Are you in need of substance abuse services? Yes or No". Human review of a random sample of documents at the concept level indicated that the NLP algorithm generated 28% false positives in extracting concepts related to homelessness when templates were ignored among the H documents. When the algorithm was refined to include templates, the false positive rate declined to 22%. For the NH documents, the corresponding false positive rates were 56% and 21%.

Conclusions

To our knowledge, this is one of the first attempts to address the problem of information extraction from templates or templated sections of the EMR. A key challenge of templates is that they will most likely lead to poor performance of NLP algorithms and cause bottlenecks in processing if they are not considered. Acknowledging the presence of templates and refining NLP algorithms to handle them improves information extraction from free text medical notes, thus creating an opportunity for improved surveillance using the EMR. Algorithms will likely need to be customized to the electronic medical record and the surveillance domain of interest. A more detailed analysis of the templated sections is underway.

Keywords

natural language processing; surveillance; templates; VA

Acknowledgments

Funding from the US Department of Veterans Affairs (HSR&D); resources from Veterans Informatics Computing Infrastructure and VA Salt Lake City Health Care System and all our research team members who have worked on this project.

References

1. DeLisle, S., et al., Combining free text and structured electronic medical record entries to detect acute respiratory infections. PloS one, 2010. 5(10): p. e13377.

*Adi Gundlapalli
E-mail: adi.gundlapalli@hsc.utah.edu