OJPHI

ISDS 2013 Conference Abstracts

Identifying Clusters of Rare and Novel Words in Emergency Department Chief Complaints

Andrew Walsh*[1], Teresa Hamby[2] and Tonya Lowery St. John[3]

[1]Health Monitoring Systems, Inc, Pittsburgh, PA, USA; [2]New Jersey Department of Health, Trenton, NJ, USA; [3]Hawaii State Department of Health, Honolulu, HI, USA

Objective

Develop a method for detecting groups of related healthcare encounters without having to specify details of the reasons for those encounters in advance.

Introduction

A goal of biosurveillance is to identify incidents that require a public health response. The challenge is creating specific definitions of such incidents so they can be detected. In syndromic surveillance, this is accomplished by classifying emergency department chief complaints, nurse triage calls, and other prediagnostic data into categories, and then looking for increases in visits related to those categories.

This approach can only find incidents that match those predefined categories. It is well-suited to handle common diseases; data from prior years provides information not only on which symptoms correlate with the disease, but also on how patients report them and how they appear in prediagnostic data streams[1]. For unique or rare events, it is hard to know in advance how they will be described or recorded.

Another approach is to look for similarities in the time of the healthcare encounters alone[2]. This method can detect events which are missed by syndrome-oriented surveillance, but healthcare encounters that only have time of occurrence aren't necessarily related. To address this limitation, we propose a set of similarity criteria which incorporates both timing and reason.

Methods

Emergency department chief complaints for the period between July 1, 2008 and June 30, 2013 were assembled from 95 hospitals. Each complaint was split into individual words; punctuation, numbers, single-letter words and common stopwords were removed.

A word frequency table was created for a single day and compared to the historical frequency table from preceding days; fixed and moving historical windows were compared. Novel words from the current day's table were identified. A Poisson distribution was used to identify rare words which occurred significantly more in the current table than the historical one.

The chief complaints containing each rare and novel word were identified. The number of chief complaints containing each word was compared to a threshold to determine if an alert should be generated. The range of time over which those chief complaints were entered was also compared to a threshold.

Results

Comparison of word tables between facilities revealed that the vocabularies of different facilities were too different to be combined; analysis would need to be done on a per-facility basis.

Analysis of only novel words yielded a large number of alerts from unique abbreviations or misspellings. Since a single interaction is unlikely to be relevant to public health, a threshold of 10 encounters containing the same word was employed to improve specificity.

Analysis of only novel words did not provide adequate sensitivity; if a second event related to the same word occurred at that facility, it would not be detected again. Thus, additional consideration is needed for rare words that occur much more frequently than expected. These rare words were determined based on the probability from a Poisson distribution defined by the historical frequency of the words.

Examination of the resulting alerts showed that some hospitals changed registration processes, resulting in a significant shift in the overall vocabulary, leading to large numbers of alerts. Using a moving window to define the word table reduced, but did not eliminate, these process alerts; until the window absorbed the data post-change, those alerts still occurred. Setting a threshold on the total time spanned by interactions involving a given word further reduced the number of alerts associated with these process changes.

Among the events detected by this analysis were chemical exposure incidents at schools in New Jersey that had gone undetected by the existing syndromic surveillance.

Conclusions

Using a surveillance method based on word frequency, rather than predefined classifications, can detect certain kinds of rare events. This could be a useful supplement to standard syndromic surveillance. It might also be useful to facilities as a quality control measure.

Keywords

novel word; unknown unknown; syndromic

References

1 Fleischauer, Aaron T., et al. The Validity of Chief Complaint and Discharge Diagnosis in Emergency Department–based Syndromic Surveillance. Academic Emergency Medicine. 2004; 11(12).

2 Burkom, H., et al. A collaboration to enhance detection of disease outbreaks clustered by time of patient arrival. International Society for Disease Surveillance Conference 2010 Track 2: Public Health Surveillance. 2011.

*Andrew Walsh
E-mail: andy.walsh@hmsinc.com