ISDS 2014 Conference Abstracts

A Term-based Approach to Asyndromic Determination of Significant Case Clusters

Howard Burkom*, Yevgeniy Elbert, Christine Piatko and Clay Fink

Johns Hopkins Applied Physics Laboratory, Laurel, MD, USA

Objective

Explain and demonstrate the performance of a statistical method for detection of anomalous terms in pooled, contiguous blocks of free-text chief complaints from a health facility with emergent or urgent care capability.

Introduction

Biosurveillance systems commonly depend on free-text chief complaints (CC)s for timely situational awareness. However, diagnosis codes may not be available soon enough and may have uncertain value because they are assigned for billing purposes rather than for population monitoring. Existing systems use syndrome categories to classify records based on these free-text fields. A syndromic cluster determination method (TOA) based on patient arrival times has been implemented in versions of ESSENCE and in NCDETECT [1]. While effective for finding case clusters whose CC terms are classifiable into syndromes, TOA implementations do not find clusters whose CC terms share only uncategorized terms.

Natural language processing literature is rich with topic-detection methods, many involving medical ontologies and language models. However, identifying clusters of CCs poses several challenges. Free-text CCs often contain only a few words, include abbreviations that vary among institutions, and evolve with usage and presentation trends. Our approach pools CCs into contiguous time blocks and uses a statistical hypothesis test to seek current terms that are anomalous relative to their occurrence in a large sliding baseline. Sets of anomalous terms (if any) are then presented for further investigation.

Methods

Our development used a collection of emergency department (ED) patient records with CCs spanning 7 years from 15 hospitals in the National Capital Region. For pre-processing, spelling correction was applied followed by removal of English stop-words and of the common words “pain”, “left”, and “right”. For anomalous word detection, we applied Fisher’s exact test to the count of each word in the tested 8hour block of CCs, based on a 30-day sliding baseline of pooled CCs ending exactly 7 days before the current test block. Using Fisher’s exact test, anomalous words from the current 8-hr block may be presented to human health monitors up to 3 times each day.

We tested this method to detect two anomaly types: a small number of instances of rare words with little or no representation in the sliding baseline, and disproportionately large numbers of common words. Multiple alerting thresholds, block sizes, and hypothesis tests were tested to achieve both types of detection with at most a handful of anomalous words expected from each block. For scenario-based detection, sets of records were injected using stochastic draws from distributions of demographic fields, of words associated with chosen outbreak types, and of date/time distributions covering from 6 hours to 3 days and kept consistent with diurnal ED visit patterns.

Results

The methods were tested on CCs from individual hospitals and on pooled data from 15-20 hospitals. Individual anomalous word tests were conducted to test the ability to detect excess clusters of rare, moderately common, and very common terms, respectively. These terms were injected into each 8-hour block over 7 years using sliding baselines with no injects.

As an alert burden measure for 15 single hospitals, the method returned no anomalous words in 60-90% (depending on the facility) of the tested (>7,600 consecutive) 8-hr blocks for a pvalue threshold of 0.01. In 95-100% of these blocks, fewer than 5 anomalous words were found. Multiple ROC-like curves showed practical sensitivity. For example, using data from the busiest single ED tested, 3 additional occurrences of “exposure” were detected in >90% of the tested blocks with less than 2 anomalous words identified per block.

Conclusions

The presented method is a practical, understandable way to monitor single care facilities for CC clusters of concern based on unusually high occurrence of rare or common terms that need not be related to syndromes. Routine implementation requires a human monitor to inspect CCs containing the anomalous terms and make follow-up decisions. A prototyped combined visualization shows recent blocks of these anomalous CC terms with syndromic time-of-arrival alerts and unusual groupings of demographic strata.

Keywords

chief complaint; Fisher’s Exact Test; asyndromic; cluster

References

Deyneka L, Xu Z., Burkom H., Hicks P., Benoit S., Vaughan-Batten H., Ising A. Finding time-of-arrival clusters of exposure-related visits to emergency departments in contiguous hospital groups. Emerging Health Threats Journal. 2011; 4:11702. doi: 10.3402/ehtj.v4i0.11702.

*Howard Burkom
E-mail: howard.burkom@jhuapl.edu