Multi-Syndrome Analysis of Time Series: A new concept for outbreak investigation

Mojdeh Mohtashemi[1, 2], Ph.D., Katherine Yih[3], Ph.D., Ken Kleinman[3], Ph.D.

[1]*MITRE;* [2]*MIT CS and AI Lab;* [3]*Harvard Medical School and Harvard Pilgrim Health Care*

OBJECTIVE

The objectives of this study are to develop a mathematical multi-syndrome framework for early detection of temporal anomalies, to demonstrate improvement in detection sensitivity and timeliness of the multivariate technique compared with those of standard uni-syndrome analysis, and to put forward a new practical concept for timely outbreak investigation.

BACKGROUND

Temporal anomaly detection is a key component of real time surveillance. Today, despite the abundance of temporal information on multiple syndromes, multivariate investigation of temporal anomalies remains under-explored. Traditionally, an outbreak is thought of as disease localization in time. That is, for an event to qualify as an outbreak, a significant deviation from the observed distribution of the disease must occur. However, the underlying processes that govern the health seeking behavior of a population with respect to one disease can potentially impact multiple syndromes leading to observable correlation patterns in the daily rates of those syndromes. Thus, a deviation from the observed correlation pattern between different syndromes can be an early indicator of potential anomalies when the rise in the daily rates of one or more syndrome is not sufficiently discernable to be identified by standard univariate techniques.

METHODS

The data are provided by the National Bioterrorism Syndromic Surveillance Demonstration Program (NDP) and involve ambulatory care encounters of patients using a large medical practice in eastern MA under five syndromes [1]. By projecting the daily syndromic rates into the five dimensional space defined by the eigenvectors of the correlation matrix of the data, Principal Components Analysis (PCA) removes redundancy, or co-linearity, in the data while it preserves the correlation structure between the syndromes [2]. This is achieved by transforming the five original variables (syndromes) to a new set of predictor variables (scores) which are linear combinations of the original syndromes. Scores contain information on how the daily rates relate to each other, which can be used to locate outliers that are both localized in time, and in the five-dimensional space, namely within syndromes. Thus, this framework can detect outlying signals that are due to mean shifts, scale shifts, or both. Part of the data that was not used to construct the PCA model was infused with randomly generated stochastic outbreaks under one or more syndrome to test the model. The sample scores for the test set were estimated by being projected into the PCA model. Outlier detection statistics were developed based on the empirical distributions of the principal components' standardized scores, and thus were different for each component.

RESULTS

Figure 1 illustrates the detection sensitivity and timeliness, and the resulting ROC curve of the multi-syndrome PCA vs. those of the uni-syndrome PCA, where the syndrome influenza-like illness (ILI) was infused with synthetic data modeling exponentially growing outbreaks with Poisson generated additive noise. The univariate PCA is equivalent to a class of standard temporal detection techniques where the detection threshold is some function of the mean and standard deviation of the underlying distribution [3].

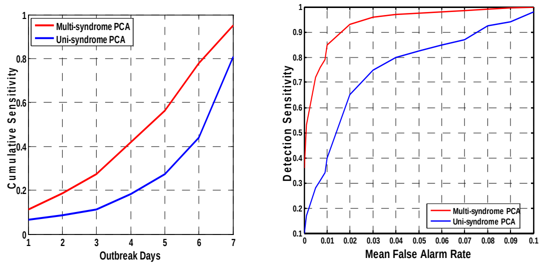

Figure 1 – Detection timeliness, sensitivity, and the ROC curve of multi-syndrome PCA vs. univariate under simulated stochastic exponential outbreaks in the ILI syndrome.

CONCLUSIONS

We proposed a non-parametric multivariate framework for detecting temporal anomalies in the daily rates of multiple syndromes. Our results indicate significant improvement in detection sensitivity and timeliness when compared with an analogous uni-syndrome scheme even when only one syndrome was injected with synthetic outbreaks (Figure 1). The proposed multi-syndrome framework lends itself to a new way of conceptualizing outbreaks and expanding on the conventional univariate mean-shift view of outbreaks. Thus, such a framework can be used for timely identification of temporal anomalies when the surge in the daily rates is not sufficiently large to be detected by standard detection techniques.

REFERENCES

[1] Yih WK, Caldwell B, Harmon R, et al. The National Bioterrorism Syndromic Surveillance Demonstration Program. In: Syndromic Surveillance: Reports from a National Conference, 2003. *Morbidity and Mortality Weekly Report* 2004;53 (suppl):43-46.

[2] Jobson JD (1992) Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods. Springer-Verlag, NY.

[3] Hutwagner L, Thompson W, Seeman GM, Treadwell T (2003) The bioterrorism preparedness and response: Early Aberration Reporting System (EARS). *Journal of Urban Health* 80(2), Supp 1,i89-i96.