Exploring Syndrome Definition by Applying Clustering Methods to Electronic Health Records Data

Samantha F. De Leon, PhD; Nicholas Soulakis; M.S., Farzad Mostashari, M.D.

New York City Department of Health & Mental Hygiene, New York, NY

OBJECTIVE

To investigate the utility of different sources of patient encounter information, particularly in the primary care setting, that can be used to characterize surveillance syndromes, such as respiratory or flu.

BACKGROUND

Electronic Health Record (EHR) data offers the researcher a potentially rich source of data for tracking disease syndromes. Procedures performed on the patient, medications prescribed (not necessarily filled by the patient), and reason for visit are just some characteristics of the patient encounter that are available through an EHR that can be used to define surveillance syndromes. Since procedures have not been used frequently in defining syndromes, encounter level procedures data, extracted from the EHR of a large local primary care practice with about 200,000 patient encounters per year was used to identify procedures associated with an established respiratory syndrome.

METHODS

Patients classified as having a respiratory syndrome as defined by the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) were included in the analyses. All ESSENCE ICD9-CM code sets can be found at: http://www.geis.fhp.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp#ICD9 . Binary indicator variables were automatically generated for each procedure performed during the patient encounter. Procedures data were analyzed using the partitioning clustering methods K-means[1] and the more robust Partitioning Around Medoids (PAM)[2] method(s). Thus far, PAM, which uses multidimensional medians as opposed to averages as the center of the cluster, has been useful in genomics, particularly in the area of tumor classification using gene expression profiles[3]. In this study, all clustering analyses were conducted in R. Corresponding diagnostic statistics were used to determine (k), the optimum number of clusters that best fit the data. Proportion of patients with fever, defined as having a temperature greater than 99.9 degrees Celsius, within each cluster was used to identify which clusters were most likely related to infectious respiratory illnesses.

RESULTS

For PAM, the most clearly defined clusters with the highest proportion of patients with fever included the following procedures: Albuterol administration (CPT= J7619); Total Vital Capacity (CPT= 94150); and Non-pressurized Inhalation procedure (CPT= 94640). Although these results have not been put to a more rigorous comparison, it is apparent that the "Asthma Procedures" syndrome identified by PAM closely follows the respiratory syndrome, particularly during the flu season when the largest peaks occur. However, the predictive value of these factors would have to be tested using predictive modeling techniques.

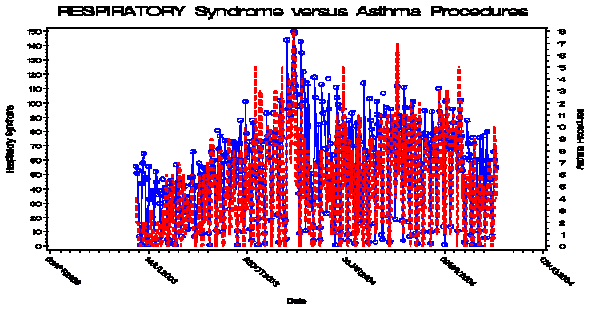

Figure 1 – Daily plot of Respiratory Syndrome (blue) and "Asthma Procedures" Syndrome (red) for July 2003 - July 2004.

CONCLUSIONS

These results indicate that simple, readily available clustering techniques can be used to identify patterns in patient encounter data that can provide useful information in further defining disease surveillance syndromes. Additional analyses that will be conducted include using medications prescribed, and comparing the performance of the K-means method versus PAM.

REFERENCES

[1] Hartigan and Wong. A K-Means Clustering Algorithm, *Applied Statistics* 1979; 28(1): 100-108.

[2] Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

[3] Dudoit and Fridlyand. Bagging to improve the accuracy of a clustering procedure. Bioinformatics 2003; 19(9): 1090-1099.computerized discharge diagnoses, J Urban Health. 2003 Jun;80(2 Suppl 1):i97-106.

Further Information:
Samantha F De Leon, sdeleon@health.nyc.gov