Syndromic Surveillance Case Definition Development Using Recursive Partitioning Techniques for Highly Dimensional Databases

Nicholas Soulakis M.S., Farzad Mostashari M.D., George Hripcsak M.D.

New York City Department of Health and Mental Hygiene

OBJECTIVE

The objective of this work was to evaluate the utility of classification tree methods for syndromic surveillance case definition development using an electronic medical record (EMR) system as a data source.

BACKGROUND

Seasonal influenza accounts for a high proportion of outpatient morbidity during the winter months. However, influenza case counts are greatly underestimated due to frequently undiagnosed influenza. EMR systems provide a very large, complex data source for influenza surveillance at both the patient and population level. It is important to identify influenza patients for specimen collection, respiratory isolation for school age children, prescription of an appropriate influenza drug, or to identify patients at risk for complications. At a population level, public health agencies monitor the tempo and spread of influenza season for resource management, as well as maintain situational awareness for avian influenza.

METHODS

Patients encounters assigned a diagnosis of influenza (ICD9 487.1) were identified from an outpatient EMR database for 2003-2004. Binary recursive partitioning algorithms [1] were used to discover candidate clinical findings among each of the following database tables: age, vital signs, reason for visit, procedures (CPT), and diagnoses (ICD-9). 10-fold cross validation was used for testing. Each tree was optimized to produce the most parsimonious variable set while maximize predictive success by selecting the smallest tree within one standard error of the minimum cost tree. Variables contributing to the model as a primary splitter or as a surrogate to primary splitter were included in the composite model.

RESULTS

301 (1.22%) influenza encounters were identified from a total of 24,691 patient encounters for the 2003-2004 influenza season. Temperature, age, respiratory rate, 51 CPT codes, 46 ICD-9 codes, and 41 ‘reasons for visit’ were identified for the final model analysis. The resulting model is illustrated in **Figure 1**. This model includes Temperature, Age, 8 reasons for visit (Sore throat, Fever, Cough, Cold symptoms, Throat problem, Fatigue, Vomiting, Perspiration), 4 ICD-9 codes (465 - Acute upper respiratory infection, 462 - Acute pharyngitis, 079 – Viral infection), 1 V-Code (V20 - Health supervision of child) and 3 CPT codes (87070 – Bacterial culture, 86403 - Particle agglutination, 87070 - Skin test; TB). All other variables fell out of the analysis.

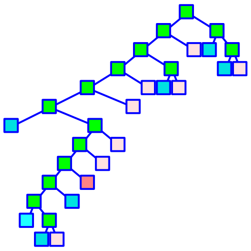

Figure 1 - Structure of classification tree.

The tree successfully predicted 213 of 301 (71%) influenza encounters and 21,676 of 24,390 (89%) non-influenza encounters. The tree successfully predicted 359 of 441 (81%) influenza encounters and 39,759 of 46,974 (84%) non-influenza encounters from the 2004-2005 influenza season in subsequent testing.

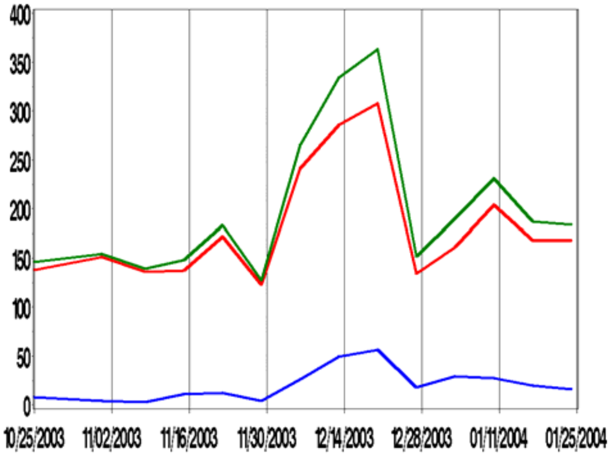

Figure 2. When plotted over time encounters classified by the tree (**red**) showed about a ten-fold increase over encounters diagnosed as 487.1. (**blue**). The two are added to form a third time series (**green**) illustrating the total estimated influenza cases for the 2003-2004 season.

CONCLUSIONS

As EMR system implementation progresses, a greater volume and variety of data sources will become available for syndromic surveillance. Very complex relational structures will require more sophisticated methods to take advantage of more diagnostic data per patient encounter. Data mining techniques such as classification trees allow discovery of discriminating clinical factors for syndromic surveillance of influenza based on a ‘gold standard.’

REFERENCES

[1] Breiman L., Friedman J.H., Olshen R.A., and Stone C.J. Classification and Regression Trees. Chapman and Hall. New York. 1998.

Further Information: Nicholas D. Soulakis (nsoulaki@health.nyc.gov)