ISDS

*Emerging Health Threats Journal* 2011, **4**:s77. **doi:** 10.3134/ehtj.10.077

www.eht-journal.org

ABSTRACT

# Applying classification and anomaly detection techniques to real-world data

W Edwards[1], A Vaid[2], and I Brooks[1]

[1]*National Center for Supercomputing Applications, Urbana, IL, USA; and* [2]*Champaign-Urbana Public Health District, Champaign, IL, USA*
E-mail: wedwards@ncsa.uiuc.edu

## Objective
This paper compares different approaches with classification and anomaly detection of data from an emergency department.

## Introduction
Real-world public health data often provide numerous challenges. There may be a limited amount of background data, data dropouts, noise, and human error. The data from an emergency department (ED) in Urbana, IL includes a diagnosis field with multiple terms and notes separated by semicolons. There are over 7000 distinct terms, excluding the notes. Because it begins in April 2009, there is not yet adequate background data to use some of the regression-based alerting algorithms. Values for some days are missing, so we also needed an algorithm that would tolerate data dropouts.

INDICATOR[1] is a workflow-based biosurveillance system developed at the National Center for Supercomputing Applications (NCSA). One of the fundamental concepts of INDICATOR is that the burden of cleaning and processing incoming data should be on the software, rather than on the health care providers.

## Methods
There were two major challenges to processing the ED data. First, we needed a way to reduce the vocabulary to a more manageable size. Second, we needed an algorithm that could tolerate a limited amount of baseline data and some data dropouts.

We grouped the terms into six syndromic groups: 'GI-Sensitive', 'GI-Specific', 'Respiratory-Sensitive', 'Respiratory-Specific', 'Flu-Like Illness', and 'Constitutional.'[2] The 'sensitive' groups include a larger set of symptoms than the 'specific' groups. Generated graphs suggested elevated activity in the Respiratory and Flu-like Illness groups around the time of the H1N1 flu outbreak in Fall 2009.

To generate alerts, we used a modified version of CDC's Early Aberration and Reporting System (EARS).[3] EARS uses an Estimated Weight Moving Algorithm (EMWA) to generate alerts. The modified approaches expand the baseline period to 28 days, separate data into weekdays and weekends, and also adjust for the total number of ED visits on a particular day.

For comparison, we grouped the data into Flu-like Illness, Respiratory-Sensitive, Respiratory-Specific, GI-Sensitive, and GI-Specific, and ran the EARS algorithm on the raw data, segregated into weekends and weekdays, and adjusted for total number of visits.

## Results
For the Flu-Like Illness Syndrome, the algorithm that separated weekend from weekday data and adjusted for the total number of ED visits was the most sensitive, generating alerts on 19 days (~6% of the total). This pattern also held for the respiratory-sensitive and respiratory-specific groups Figure 1.

## Conclusions
This approach has yielded promising results, and in the future, we plan to expand the number of syndromic groups to explore rates of ED activity related to substance abuse and

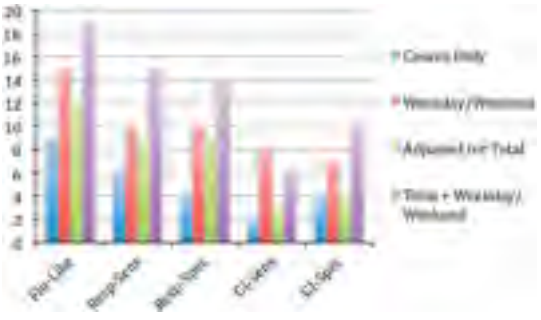

**Figure 1** Number of alerts generated for each syndrome group using different modifications of the EARS algorithm. The blue color analyzes only the daily counts for a particular syndrome. The red color separates the counts into weekdays and weekends. The green color adjusts for the total number of ED visits that day. The purple color both separates weekday/weekend data and adjusts for ED totals.

West Nile virus. The modified EARS algorithms also worked well for us, and we plan to apply them to the school absence data, where enrollment, as well as absence figures have fluctuated.

## Acknowledgements

This paper was presented as an oral presentation at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA, on 1–2 December 2010.

## References

1 Brooks I, Edwards W. INDICATOR: A Cyberenvironment for biosurveillance and response. *Syndromic* 2009, 8th Annual Conference of the International Society for Disease Surveillance, 2009.
2 Chapman W. Developing Syndrome Definitions based on Consensus and Current Use. *J Am Med Inform Assoc* 2010;**17:** 595–601.
3 Tokars J, Burkom H, Xing J, English R, Bloom S, Cox K, *et al*. Enhancing Time-Series Detection Algorithms for Automated Biosurveillance. *Emerging Infectious Diseases* 2009; **15**:4.