Emerging Health Threats Journal 2011, **4**:s55. **doi:** 10.3134/ehtj.10.055

www.eht-journal.org

ABSTRACT

Data quality in federated disease surveillance: using variability as an indicator of quality

N Um[1], S Visweswaran[1], J Espino[2], and M Wagner[1,2]

[1]*Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; and* [2]*Real-Time Outbreak and Disease Surveillance Laboratory, University of Pittsburgh, Pittsburgh, PA, USA*
E-mail: nau2@pitt.edu

Objective
We developed a novel method for monitoring the quality of data in a federated disease surveillance system, which we define as 'a surveillance system in which a set of organizations that are not owned or controlled by public health provide data.'

Introduction
Most, if not all, disease surveillance systems are federated in the sense that hospitals, doctors' offices, pharmacies are the source of most surveillance data. Although a health department may request or mandate that these organizations report data, we are not aware of any requirements about the method of data collection or audits or other measures of quality control.

Because of the heterogeneity and lack of control over the processes by which the data are generated, data sources in a federated disease surveillance system are black boxes the reliability, completeness, and accuracy of which are not fully understood by the recipient.[1]

In this paper, we use the variance-to-mean ratio (VMR) of daily counts of surveillance events as a metric of data quality. We use thermometer sales data as an example of data from a federated disease surveillance system. We test a hypothesis that removing stores with higher baseline variability from pooled surveillance data will improve the signal-to-noise ratio of thermometer sales for an influenza outbreak.

Methods
We computed the VMR for each of 178 drug stores in Allegheny County, PA, USA. In particular, we computed VMR for a non-influenza period, which we term the 'base period' (BP). We used the 81-day period from 1 June 2009 to 20 August 2009, inclusive.

Before computing VMR, we smoothed daily thermometer sales for each story by applying a 7-day moving average (MA7) to remove day-of-the-week effects.

To determine whether removal of stores with highest VMR improves the ability to detect an influenza outbreak, we systematically removed stores with the highest VMR from the total daily counts summed for the 178 stores in Allegheny County. To determine whether the timeliness of algorithmic detection was affected by the removal of high VMR stores, we used a detection threshold of three s.d. above

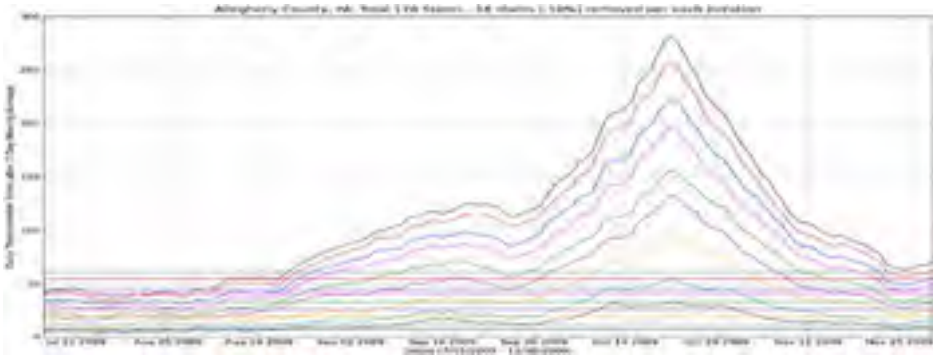

Figure 1 7-day moving average daily thermometer sales TS plot with progressive removal of stores. The uppermost line represents all 178 stores in Allegheny County, PA. Each subsequent TSs represents removal of 18 stores by VMR scores.

the mean of the BP. We also compared the date of peak thermometer sales and the signal-to-noise ratio at peak.

Results

The mean VMR for the 178 stores was 1.16 with maximum of 2.73 and minimum of 0.76. Figure 1 shows that the shape of the plot of county-wide sales of thermometers remained as high-variability stores were removed in 10% tranches. The detection date and peak date were unchanged (August 24 ± 1 day and October 21, 2009 ± 1 day, respectively) through the progressive removal process. The signal-to-noise ratio, measured as number of s.d.s above the BP mean on the 'peak' was 45 s.d. at 0% removal, 41 s.d. at −30%, 33 s.d. at −60%, and 21 s.d. at −80%.

Conclusions

There was significant difference in the VMR for sales of thermometers by different stores. However, removal of those stores from a surveillance system did not improve the ability of a typical surveillance algorithm to detect the 2009 influenza epidemic and the signal to noise ratio at the peak of the epidemic was not improved by the removal of the stores with more baseline variability.

Acknowledgements

This research was supported by a grant from the Lockheed-Martin Corporation. 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.

Reference

1 Wagner MM. Methods for evaluating surveillance data. *Handbook of Biosurveillance*, Chapter 21, p 313–20, 2006.