ISDS

Emerging Health Threats Journal 2011, **4**:s69. **doi:** 10.3134/ehtj.10.069

www.eht-journal.org

ABSTRACT

Automated detection of data entry errors in a real time surveillance system

L Chen[1], A Dubrawski[1], N Waidyanatha[2], and C Weerasinghe[2]

[1]*Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA; and* [2]*LIRNEAsia, Colombo, Sri Lanka*
E-mail: lujiec@andrew.cmu.edu

Objective
We present a method for automated detection of systematic data entry errors in real time biosurveillance.

Introduction
Real-Time Biosurveillance Program (RTBP) introduces modern surveillance technology to health departments in Sri Lanka and Tamil Nadu, India.[1] Triage data from each patient visit (basic demographics, signs, symptoms, preliminary diagnoses) is recorded on paper at health facilities. Case records are transmitted daily to a central database using the RTBP mobile phone application. It is done by medical professionals in India, but in Sri Lanka, due to staffing constraints, the same duty is performed by lower cost personnel with limited domain knowledge. That results in noticeable differences in data entry error rates between the two locations. Most of such issues are due to systematic and subjective misinterpretations of the handwritten doctor notes by the data entry personnel. If not identified and remedied quickly, these errors can adversely affect accuracy and timeliness of health events detection. There is a need to support system managers in their efforts to maintain high reliability of data used for public health surveillance.

Methods
To address the need, we develop algorithms for automated detection of systematic data quality issues. They are used in automated, on-line screening of incoming data for potential discrepancies. Lists of potential issues are presented to human operators for evaluation. The operators may then correct the contents of the database, re-execute analyses, which might have been affected by errors, and follow-up with data entry personnel to prevent similar issues from happening in the future.

One of such algorithms relies on the assumption that most disease outbreaks do not affect population served by just one health facility. We expect neighboring facilities to show at least some level of correlated activity. We use entropy as a measure of uniformity of the geographic distribution of disease. Spatially isolated outbreaks will be characterized with low values of entropy, whereas widespread events will show high entropies. The data used in our experiments contains 62,000 disease cases from 13 locations in Sri Lanka covering 158 disease categories. Each data entry clerk is assigned to a specific health facility. Therefore, systematic subjective data entry errors tend to show as localized patterns ('spikes' or 'dips') with low spatial entropy. Our algorithm exhaustively searches data for such instances.

Results and conclusion
The Figure 1 below presents the results of analysis limited to notifiable diseases observed in Sri Lanka during a specific period of time. Horizontal axis denotes normalized entropy of disease geospatial distribution, vertical axis depicts relative frequency of diseases, and each point corresponds to one notifiable disease. Points in the upper left quadrant of the plot are likely the results of miscoding. These diseases show relatively high numbers of cases combined with a very limited spatial dispersion. Indeed, upon investigation, it was revealed that the spike of *Tetanus* resulted from mistakenly entering immunization records as if they were disease cases. The phantom outbreak of measles resulted from mistakenly recording lexically similar diagnoses of muscle pain. However, the unexpected spike of 150 cases of *fever greater than 7 days* observed at just one location in February and March of 2010, an ailment not typically seen in dry season, was a legitimate health event.

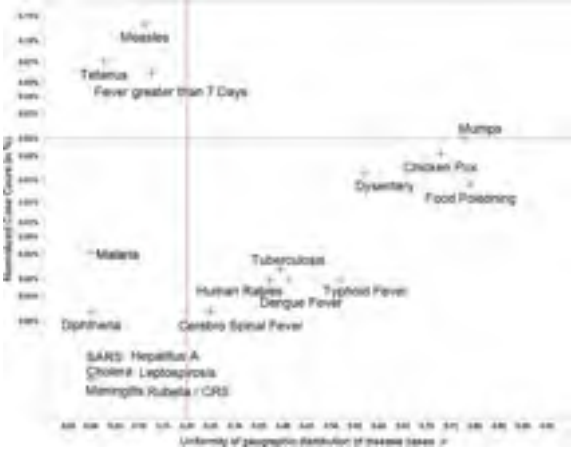

Figure 1 Entropy map of notifiable disease in Sri Lanka.

Low quality of data can easily invalidate analyses. Implementers of analytic systems must be prepared to face such challenges, especially in environments with limited resources. We have shown how to automatically detect certain types of avoidable systematic data entry errors in support of real time biosurveillance.

Acknowledgements

This work was supported in part by the International Development Research Centre of Canada (Award 105130) and National Science Foundation under grant number 0911032. 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 Sabhnani M, Dubrawski A, Waidyanatha N. T-Cube Web Interface for Real-time Biosurveillance in Sri Lanka. Proceedings of the 2009 ISDS Annual Conference, Miami, Florida.