An Open Source Web Services Toolkit For Event Detection Algorithms

William B Lober[1], Dan Drozd[1], Thomas Lumley[2], Krisztian Sebestyen[2], Ian Painter[3]

[1]*Biomedical and Health Informatics,* [2]*Department of Biostatistics, University of Washington*

[3] *Foundation for Healthcare Quality, Seattle, Washington.*

Abstract: *We describe the design and initial implementation steps for a web-services based toolkit for evaluating outbreak detection methods. The toolkit will include components for combining simulated and historical data to create artificial outbreaks and components that implement various outbreak detection algorithms. The first algorithms implemented are: a) the three Cumulative Sums (cusum) methods described in the CDC Early Aberration Reporting System[1], b) ARIMA, and c) Smart Scores. The web services interface allows for the simple design of language-neutral user interfaces targeted to disparate user populations within the public health community and minimizes the necessity to develop complex statistical algorithms at the local public health agency.*

Introduction: As major disease outbreaks are rare, empirical evaluation of statistical methods for outbreak detection requires the use of modified or completely simulated health event data in addition to real data. Comparisons of different techniques will be more reliable when they are evaluated on the same sets of artificial and real data. To this end, we are developing a toolkit for implementing and evaluating outbreak detection methods and exposing this framework via a web services interface.

Methods: The toolkit consists of two main components: a statistical analysis and graphics package written in R[2] and a web services tier that exposes the analysis package. R is an open-source statistical programming environment based on a dialect of Bell Labs' S language, which includes many of the components needed for anomaly detection and analysis, such as ARIMA time series models, statistical process control charts, change-point regression models. It is perhaps the most popular system among research statisticians for implementing and evaluating new statistical methods.

R is a functional programming language, which we use to implement a pipeline structure for event streams, allowing simple data processing tasks to be composed in powerful ways, either for simulation or operational use. For example, a simulated outbreak of food poisoning might use an existing surveillance data stream and add a fixed or random set of extra events. The anomaly detection methods take information from an event stream and return an object summarizing the results. This object can have methods for printing a text summary and for displaying an appropriate graphical summary (such as a CUSUM chart)

Results: The Apache Axis web services framework running within the Apache Jakarta Tomcat application server is used to process all web service requests. All web services interfaces are defined using WSDL documents allowing for easy connectivity between diverse systems. We have developed custom data management and service code in Java which manages service requests and passes them to the R environment via Rserve, a TCP/IP server. We have currently implemented the three CUSUM-based EARS algorithms for outbreak detection[1] and verified them on test data (http://www.bt.cdc.gov/surveillance/ears/datasets.asp), and have implemented simulated data streams from Poisson and negative binomial distributions, with and without seasonal variation. We have also implemented the ARIMA methods of Reis and Mandl, and the Smart Scores method of Kleinman. Each of these algorithms is exposed via a web services interface, as are related data management and utility functions.

Conclusion: Syndromic surveillance efforts are generally focused at the level of the LPHA. Often LPHAs would like to know the parameters of performance of particular algorithms and data sources in their own setting. To determine this however requires implementing the various algorithms and simulating outbreaks, which currently requires expertise in programming to generate simulations. By developing a toolkit of tested and understood surveillance algorithms and exposing them via a web services interface, our platform will enable LHJs to easily evaluate syndromic surveillance algorithms in their own setting, while providing a powerful and flexible environment for research into the performance of algorithms and data sources.

Acknowledgments: Foundation for Healthcare Quality, US Army Medical Research Acquisition Activity W23RYX-3263-N612.

References

[1]Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *J Urban Health*. 2003 Jun;80(2 Suppl 1):i89-96

[2] R Development Core Team (2004) *R: a language and environment for statistical computing.* Version 2.0.1. R Foundation for Statistical Computing: Vienna, Austria.