An Open Source Environment For The Statistical Evaluation Of Outbreak Detection Methods

Thomas Lumley[1], PhD, Krisztian Sebestyen[1], BS,William B Lober[2] MD, Ian Painter[3] PhD

[1]*Department of Biostatistics,* [2]*Biomedical and Health Informatics, Seattle, Washington,*
[3] *Foundation for Healthcare Quality, Seattle, Washington.*

Abstract

We describe the design and initial steps to implementation of a computational framework for evaluating outbreak detection methods. The framework will include components for combining simulated and historical data to create artificial outbreaks and components that implement various outbreak detection algorithms. The first algorithms to be implemented are the three Cumulative Sums (cusum) methods described in the CDC Early Aberration Reporting System[1].

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 framework for implementing and evaluating outbreak detection methods.

Methods

The two main components of this framework are a pipeline system for simulating and modifying data streams, and a standardized interface for anomaly detection algorithms that read data from a supplied input stream and return objects summarizing their results.

The statistical analysis and graphics is programmed using R[2]. R is an open-source statistical programming environment based on a dialect of Bell Labs' S language. R already 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, and is perhaps the most popular system among research statisticians for implementing and evaluating new statistical methods.

The R language per se is a functional programming language allowing for computation on the language. This feature is used to implement a pipeline structure for event streams. This structure allows simple data processing tasks to be composed in powerful ways. For example, a simulated outbreak of food poisoning would take an existing input stream and add a fixed or random set of extra events. A filter for severity of cases could thin out an input stream to allow for a lower probability of hospital visits for less severe cases. A location-of-residence filter could allow for some cases being diverted to hospitals in other counties or regions. A miscoding filter could take two input streams corresponding to different event types and allow for misclassification between them.

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

We have currently implemented the three CUSUM-based EARS algorithms for outbreak detection[1] and verified them on test data from the Centers for Disease Control (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.

Conclusion

Syndromic surveillance efforts are generally focused at the level of the local health jurisdiction (LHJ). Our experience working with several LHJ's is that they would like to know the parameters of performance of particular algorithms and data sources in their own local setting. To determine this however requires implementing the various algorithms and simulating outbreaks. Methods for the statistical detection of outbreaks have been implemented in a range of programming languages, and currently expertise in programming is required to generate simulations. This makes comparisons at the local level difficult for LHJ's with limited resources. Our environment will enable LHJ's 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.