

Detection of Outbreak Signals Using R

Steve E. Rigdon*¹, George Turabelidze², Richard DeClue¹, Sarah Patrick¹, Ehsan Jahanpour², Rong He² and Jennifer Lloyd²

¹Saint Louis University, Saint Louis, MO, USA; ²Missouri Department of Health and Senior Services, Saint Louis, MO, USA

Objective

Develop a statistically rigorous automated process for weekly communicable disease report analysis to improve the speed and accuracy of outbreak detection in Missouri.

Introduction

The goal of adequate biosurveillance is to signal that an outbreak may be occurring and through subsequent work is confirmed or refuted. Such a system should be equally able to detect outbreaks of diseases of extremely low reporting frequency, or those with high seasonality.

Methods of detecting increases in notifiable communicable diseases reported to the Missouri Department of Health and Senior Services (MDHSS) were based on quartile comparisons to 5-year historical disease reports for the report week and resulted in frequent detection of statistically significant increases that were, in fact, not indicative of disease outbreaks. Frequently generated alerts led to “alarm fatigue” in epidemiologists.

Methods

MDHSS’s 10-year communicable diseases database (2002-2011) was used to establish the background frequency of selected diseases: E. coli O157:H7 (a predominantly foodborne disease with strong outbreak potential, EC), ehrlichiosis (a highly seasonal tickborne infection, ERH), Q fever (low incidence disease, Q), and pertussis (high incidence disease, P).

Seasonality was modeled by fitting a first- or second-order trigonometric model¹ to the average reported cases in each week of the year. Assuming a Poisson distribution for the number of cases per week, we can get an estimate of the mean (or expected) count for each week. Since the mean and variance of the Poisson are equal, we are able to estimate the variance, and therefore the standard deviation, of the count from this trigonometric model. With this information, we can get upper and lower control limits for the weekly observed disease counts. These limits vary due to the seasonality of the baseline.

A user interface was created in the open source software R² using the Shiny³ package. The interface allows the user to select through drop-down boxes many of the important parameters of the procedure, including the disease to be monitored, the degree of the trigonometric regression, the upper control limit (measured in standard deviations from the mean) and whether the control limits should be recomputed excluding points outside the control limits. In this pilot test, the Disease Electronic Surveillance with Trigonometric Models (DESTEM) was applied to Missouri’s weekly counts of selected diseases in weeks 1 through 35 of 2013 and comparisons made to the earlier system.

Results

MDHSS’s weekly report system generated significantly more alerts during the pilot for EC and ERH compared to the new DESTEM analysis. Alerts generated by DESTEM correctly identified two outbreaks and one cluster of EC. No alerts were generated for P and Q by either technique. There were no outbreaks of P, ERH, or Q during the study period. Alerts generated by the existing system were based on underlying “descriptive” statistics of the historical data, and

produced multiple alerts which required confirmatory work by epidemiologists before starting investigation.

Conclusions

DESTEM provides a signal of sufficient probability to trigger further public health response without the burden of false alarms from the earlier system. Generated signals appropriately indicated when further investigation was warranted. A user friendly interface allowed users easy and “on-demand” access to the surveillance data analysis. More accurately generated signals enabled epidemiologists at all levels to interpret surveillance data more accurately. Further study is needed to validate DESTEM using the historical infectious disease outbreak data of additional organisms and to quantify the new system’s timeliness and sensitivity.

Keywords

Poisson distribution; trigonometric regression; disease surveillance; outbreak

References

1. Cryer, J. D. and Chan, K.-S., Time Series Analysis: With Applications in R, Springer, New York, 2008.
2. R Core Development Team, R: A language and environment for statistical computing, reference index version 2.2.1. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>, 2005.
3. RStudio, Shiny, <http://www.rstudio.com/shiny/>, 2013.

*Steve E. Rigdon
E-mail: srigdon@slu.edu

