How good is your data?

Ian Painter[1]*, Julie Eaton[1], Don Olson[2], Debra Revere[1] and William Lober[1]

[1]University of Washington, Seattle, WA, USA; [2]International Society for Disease Surveillance, Boston, MA, USA

Objective

The goal of this session will be to briefly present two methods for comparing aggregate data quality and invite continued discussion on data quality from other surveillance practitioners and to present the range of data quality results across participating Distribute sites.

Introduction

Distribute is a national emergency department syndromic surveillance project developed by the International Society for Disease Surveillance (ISDS) for influenza-like illness (ILI) that integrates data from existing state and local public health department surveillance systems. The Distribute project provides graphic comparisons of both ILI-related clinical visits across jurisdictions and a national picture of ILI.

Unlike other surveillance systems, Distribute is designed to work solely with summarized (aggregated) data, which cannot be traced back to the unaggregated 'raw' data. This and the distributed, voluntary nature of the project create some unique data quality issues, with considerable site to site variability. Together with the ISDS, the University of Washington has developed processes and tools to address these challenges, mirroring work done by others in the Distribute community.

Methods

University of Washington together with the ISDS has undertaken a comprehensive analysis of the quality of the data being received by Distribute, primarily using visual methods, examining data quality characteristics within and between sites. This process included basic exploratory analysis of data quality problems and analytical analysis of specific aspects of data quality, including the relationship between timeliness, completion and accuracy.

Results

Considerable variability was seen between sites in terms of timeliness and completion, and completion rates did not necessarily correlate with accuracy. In our talk, we will present results comparing the quality of data between sites (sites will be unidentified), in particular comparisons between timeliness, completion and accuracy. We will also examine the types of observed relationships between timeliness, completeness and accuracy exhibited across the sites.

The purpose of this talk is to facilitate discussion between Distribute participants around data quality and the role that the ISDS can play in ensuring data quality. We will show prototypes of two features that could be hosted on the Distribute restricted site. The first feature would allow each site to compare the quality of their data (identified only to them, with site linked to the id of the user) with the remaining sites (each unidentified). The second feature would allow each site to see time series of their data together with prediction intervals for the accuracy of the ILI ratio for recent dates where the data are incomplete (see Fig. 1).

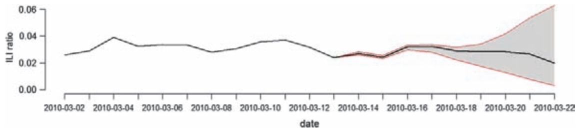

Fig. 1. ILI ratio timeseries calculated from incomplete data with superimposed 95% prediction interval for the complete data value for a representative site.

Conclusions

Our goal is to spark discussion on data quality with respect to syndromic surveillance data and, in particular, how the Distribute project can be leveraged to improve the quality of aggregate data produced by participating sites.

Keywords

Data quality; surveillance; public health practice; data quality

***Ian Painter**
E-mail: ipainter@uw.edu

Citation: Emerging Health Threats Journal 2011, **4**: 11142 - DOI: 10.3402/ehtj.v4i0.11142