Propagating the effects of missing data sources in a surveillance system

I. S. Painter, PhD[1], A. Harvey[2] BS, K. Peterson[2], MS, P. Tran[2], BS and Kevin Taylor[2]
[1]Foundation for Health Care Quality, Seattle, WA; [2]Paladin Data Corp., Poulsbo, WA

Abstract

Sporadically missing data sources in syndromic surveillance systems result in inaccurate counts and detection algorithm results. We examined how data quality issues relating to missing data sources propagate through a surveillance system and devised a method to track and visualize the resulting data quality issues.

Introduction

A common problem in syndromic surveillance using ED department data is temporary gaps in the data received from individual ED departments caused by delays in receiving the data.[2]

Currently most syndromic surveillance systems provide information about the status of the data sources feeding into the system, for example on the home page of the system, but do not show the effects of any missing data sources on individual derived data elements (except in that graphs may show obvious drops in counts on days when data sources are missing).

Methods

We traced the effects of missing data sources on the integrity of derived syndrome counts within a syndromic surveillance system.

Missing data sources propagate data errors up to aggregations (roll-ups) of counts over data sources, and to drill-downs within demographic dimensions (for example, the number of ED visits coded to the respiratory syndrome in 0-2 year olds in all ED departments in Acme County). Data errors in analytical results, which typically consist of single or multidimensional time series analyses, can also be caused by missing data sources occurring on any of the dates used for baseline data by the analysis method.

We characterized derived counts into three categories: complete, meaning all data sources involved in the derivation of the count provided data; incomplete, meaning at least one of the data sources did not provide data; or unavailable, meaning that no data sources from which the count derives from reported data. Analytical results can similarly be categorized into complete, incomplete or unavailable, though multiple causes can result in the same end point (for example, an unavailable count for today results in an unavailable analytic result for today, or an available count for today and unavailable counts for all baseline days also results in an unavailable analytic result for today). In addition, some detection algorithms produce meaningful results with certain level of unavailable baseline data, so that incomplete baseline data can result in either incomplete analytical results or unavailable analytical results depending on the unavailability level of the baseline data.

We designed two methods of incorporating the completeness of derived data elements in a syndromic system: directly incorporating completeness information into meta-data associated with the derived data elements (a 'push' method) or by including meta-data on the sources for derived data elements and tracing back to meta-data stored about the completeness of the data sources (a 'pull' method). Both approaches offer advantages linked to the sparseness of the data, the frequency of updates of previously missing data sources and the length of the baseline data used in analytics. In particular, high frequency of updates and sparse data favors the pull method, while longer baselines favor the push method

We have also designed several methods for visualizing the completeness level of data elements in counts tables and time series graphs.

References

1. Mandl KD, et. al. Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience. JAMIA 2004; 11(2):141-150.