COACTION PUBLISHING

An information visualization approach to improving data quality

Atar Baer*

Public Health—Seattle & King County, Seattle, WA, USA; Department of Epidemiology, University of Washington, Seattle, WA, USA

Objective

We sought to develop a method for visualizing data quality over time.

Introduction

The Public Health—Seattle & King County (PHSKC) syndromic surveillance system has been collecting emergency department (ED) data since 1999. These data include hospital name, age, sex, zip code, chief complaint, diagnoses (when available), disposition and a patient and visit key. Data are collected for 19 of 20 King County EDs, for visits that occurred the previous day. Over time, various problems with data quality have been encountered, including data drop-offs, missing data elements, incorrect values of fields, duplication of data, data delays and unexpected changes in files received from hospitals. In spite of close monitoring of the data as part of our routine syndromic surveillance activities, there have occasionally been delays in identifying these problems. Since the validity of syndromic surveillance is dependent on data quality, we sought to develop a visualization to help monitor data quality over time, in order to improve the timeliness of addressing data quality problems.

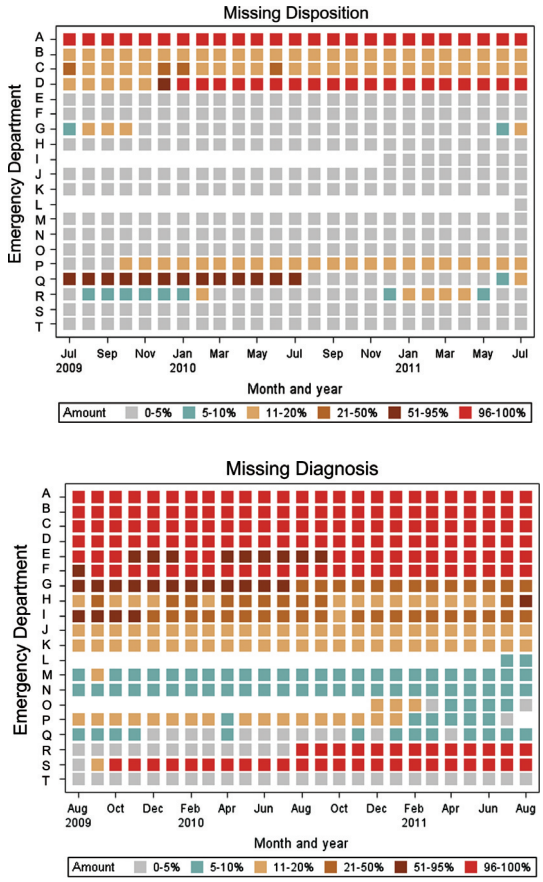

Fig. 1. Heatmaps showing percent of records at each ED missing patient disposition (top) and diagnosis (bottom) data over time. The heatmaps provide a visual cue for when data quality has changed (color shifts), where data are routinely unavailable (stable color over time) or missing (white space). The heatmaps are generated monthly for each of 13 data quality measures.

Methods

SAS version 9.2 (Carey, NC) was used to create two groups of visualizations: (1) a separate heatmap for each hospital, showing how each individual ED performs on each of 13 data quality measures and (2) a separate heatmap for each data quality measure, showing how data quality varies by ED. The heatmaps summarize data by month and year, though other visualizations (e.g., daily or weekly) are also possible. For each row on the heatmap, a color change indicates that data quality has shifted over time. Blocks with stable color over time suggests that there has not been a change in data quality. White space on the heatmap highlights periods of time where data were not recorded by the system and can provide a visual cue for newly added EDs, hospital closures or data drop-offs. The heatmaps are generated monthly for each of 13 data quality measures. SAS code for generating the heatmaps will be provided at the session.

Results

Two heatmaps are provided as examples of our visualization approach (see Fig. 1). Since applying this visualization to our syndromic data, PHSKC has identified several data quality errors that are likely to have gone undetected or been slow to detect otherwise, including out of range ages and sudden data drop-offs. Consequently, we have adopted the methodology to other nonsyndromic data sources, including notifiable condition reporting to the health department.

Conclusions

Syndromic surveillance systems commonly encounter problems with data quality. These problems can result in imprecise counts and can adversely affect detection of trends, outbreaks and situational awareness. The heatmap visualizations have been a useful tool for PHSKC to identify problems with data quality in a timely manner. The code can be easily adapted to display other data quality measures, stratifications and data sources beyond the ED setting.

Keywords

Data quality; public health practice; visualization

***Atar Baer**
E-mail: atar.baer@kingcounty.gov

 Citation: Emerging Health Threats Journal 2011, **4**: 11014 - DOI: 10.3402/ehtj.v4i0.11014