ISDS 2012 Conference Abstracts

Data Quality: A Systematic Review of the Biosurveillance Literature

Tera Reynolds*[1], Ian Painter[2] and Laura Streichert[1]

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

Objective

To highlight how data quality has been discussed in the biosurveillance literature in order to identify current gaps in knowledge and areas for future research.

Introduction

Data quality monitoring is necessary for accurate disease surveillance. However it can be challenging, especially when "real-time" data are required. Data quality has been broadly defined as the degree to which data are suitable for use by data consumers [1]. When compromised at any point in a health information system, data of low quality can impair the detection of data anomalies, delay the response to emerging health threats [2], and result in inefficient use of staff and financial resources. While the impacts of poor data quality on biosurveillance are largely unknown, and vary depending on field and business processes, the information management literature includes estimates for increased costs amounting to 8-12% of organizational revenue and, in general, poorer decisions that take longer to make [3].

Methods

To fill an unmet need, a literature review was conducted using a structured matrix based on the following predetermined questions:

-How has data quality been defined and/or discussed?
-What measurements of data quality have been utilized?
-What methods for monitoring data quality have been utilized?
-What methods have been used to mitigate data quality issues?
-What steps have been taken to improve data quality?

The search included PubMed, ISDS and AMIA Conference Proceedings, and reference lists. PubMed was searched using the terms "data quality," "biosurveillance," "information visualization," "quality control," "health data," and "missing data." The titles and abstracts of all search results were assessed for relevance and relevant articles were reviewed using the structured matrix.

Results

The completeness of data capture is the most commonly measured dimension of data quality discussed in the literature (other variables include timeliness and accuracy). The methods for detecting data quality issues fall into two broad categories: (1) methods for regular monitoring to identify data quality issues and (2) methods that are utilized for *ad hoc* assessments of data quality. Methods for regular monitoring of data quality are more likely to be automated and focused on visualization, compared with the methods described as part of special evaluations or studies, which tend to include more manual validation.

Improving data quality involves the identification and correction of data errors that already exist in the system using either manual or automated data cleansing techniques [4]. Several methods of improving data quality were discussed in the public health surveillance literature, including development of an address verification algorithm that identifies an alternative, valid address [5], and manual correction of the contents of databases [6].

Communication with the data entry personnel or data providers, either on a regular basis (e.g., annual report) or when systematic data entry errors are identified, was mentioned in the literature as the most common step to prevent data quality issues.

Conclusions

In reviewing the biosurveillance literature in the context of the data quality field, the largest gap appears to be that the data quality methods discussed in literature are often *ad hoc* and not consistently implemented. Developing a data quality program to identify the causes of lower quality health data, address data quality problems, and prevent issues would allow public health departments to more efficiently and effectively conduct biosurveillance and to apply results to improving public health practice.

Keywords

Biosurveillance; Data quality; Literature review

Acknowledgments

We thank the ISDS Data Quality Workgroup for initiating this project, which was supported by CDC through contract with the Task Force for Global Health.

References

1. Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. JMIS. 1996:5–33.
2. Dixon BE, McGowan JJ, Grannis SJ. Electronic Laboratory Data Quality and the Value of a Health Information Exchange to Support Public Health Reporting Processes. Proc AMIA Symp. 2011;2011:322.
3. Redman TC. The impact of poor data quality on the typical enterprise. Commun ACM. 1998;41(2):79–82.
4. Maydanchik A. Data Quality Assessment. Technics Publications, LLC; 2007.
5. Zinszer K, Charland K, Jauvin C, et al. The influence of address errors on detecting outbreaks of campylobacteriosis. Emerg Health Threats J. 2011;4(s59):68–69.
6. Chen L, Dubrawski A, Waidyanatha N, Weerasinghe C. Automated detection of data entry errors in a real time surveillance system. Emerg Health Threats J. 2011;4(s69):9–10.

*Tera Reynolds
E-mail: treynolds@syndromic.org

ISDS Annual Conference Proceedings 2012. This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.