ISDS

Emerging Health Threats Journal 2011, **4**:s142. **doi:** 10.3134/ehtj.10.142

www.eht-journal.org

ABSTRACT

Classification of errors for quality assurance in the emerging infections program influenza hospitalizations surveillance system

A Pérez, T D'Mello, L Kamimoto, L Finelli, and MH Torres-Urquidy

Influenza Division, Centers for Disease Control and Prevention, Atlanta, GA, USA
E-mail: hvv9@cdc.gov

Objective

Introducing data quality checks can be used to generate feedback that remediates and/or reduces error generation at the source.[1] In this report, we introduce a classification of errors generated as part of the data collection process for the Emerging Infections Program (EIP)'s Influenza Hospitalization Surveillance Project at the Centers for Disease Control and Prevention (CDC). We also describe a set of mechanisms intended to minimize and correct these errors via feedback, with the collection sites.

Introduction

The CDC's Emerging Infections Program monitors and studies many infectious diseases, including influenza.[2] In 10 states in the US, information is collected for hospitalized patients with laboratory-confirmed influenza. Data are extracted manually by EIP personnel at each site, stripped of personal identifiers and sent to the CDC. The anonymized data are received and reviewed for consistency at the CDC before they are incorporated into further analyses. This includes identifying errors, which are used for classification.

Methods

We evaluated the most current dataset as of 24 August 2010, containing records for 6521 persons with influenza-associated hospitalizations from 1 September 2009 through 30 April 2010. We built fully automated software routines using SAS version 9.2 (SAS Institute Inc., Cary, NC, USA) to conduct quality assurance. For instance: when data about the patient age are not provided, our software identifies the missing information as an error. We generated our classification based on the characteristics of these errors using a data-driven approach (that is, clustering errors with similar properties). The classification was then discussed internally. Based on the common characteristics of the clusters we developed common definitions for each category in the classification. Finally, we measured the actual number of errors in the most current collected dataset and categorized the classification by type, with the most prominent ratio (predefined errors/errors in reports).

Table 1 Error classification

Name	*Predefined errors*	*Errors in reports*	*Ratio*
Data entry errors	43	761	18.12
Missing data	19	263	13.84
Integrity	18	129	7.17
Failure to meet case definition	4	28	7.00
Chronology	32	113	5.35

Results

The implementation of the error classification occurred during the preparation of the monthly report submitted to sites. The error classification was generated (Table 1). We found that 'Data Entry Errors' were the most prominent followed by 'Missing Data.' Other types of errors were identified as well.

Conclusions

Classification of errors allows for easier identification and prompt correction. In addition, it will allow us to improve subsequent versions of the software used to capture information and possibly minimize errors during capture.

Acknowledgements

This paper was presented as a poster at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA, on 1–2 December 2010.

References

1 Mullooly JP. The effects of data entry error: an analysis of partial verification. *Comput Biomed Res* 1990;**23:**259–67.
2 Dawood FS, Fiore A, Kamimoto L, Nowell M, Reingold A, Gershman K, *et al*. Influenza-associated pneumonia in children hospitalized with laboratory-confirmed influenza, 2003–2008. *Pediatr Infect Dis J* 2010;**29:**585–90.