OJPHI

ISDS 2015 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

An R Script for Assessment of Data Quality in the BioSense Locker Database

Serena Rezny* and Stacey Hoferka

Illinois Department of Public Health, Chicago, IL, USA

Objective

To describe an R script developed to assess and produce reports on data quality in the BioSense locker database.

Introduction

Syndromic surveillance requires reliable, accurate, and complete healthcare encounter data to assess patterns of illness and respond to public health events. Illinois implemented syndromic surveillance statewide in response to Meaningful Use reporting objectives. To address the need for continuous, automated assessment following initial on-boarding of facility Emergency Department data, we developed an R script to assess the quality of data in the private BioSense locker database.

This script builds upon and adapts from scripts previously developed for syndromic surveillance[1, 2] and data quality assessment[3].

Methods

This R script examines identifying variables in the HL7 messages from the locker, aggregates messages into ED visits based on these identifiers, processes the aggregated data to calculate metadata for each visit, and computes various data quality metrics. Results are displayed in the console and written to an HTML file.

Given a user-specified time period and list of contributing facilities, this R script assembles a MySQL query and executes it to retrieve messages from the database meeting the specified criteria.

This script first examines the identifiers Unique_Visiting_ID, Patient_Visit_ID, Unique_Patient_ID, and FacilityID_UUID. It calculates the number of distinct values of UVID, PVID, and (FID, PVID) which appear, then calculates the number of UVID corresponding to each (FID, PVID) pair and vice versa. Finally, it finds instances of more than one PVID matching a single (FID, PID) pair on a single date and tallies the number of such occurrences at each facility.

This script aggregates messages into patient visits based on the ordered pair (FacilityID_UUID, Patient_Visit_ID); this pair of HL7 data elements as defined in the PHIN messaging guide uniquely identifies patient visits using a facility-generated identifier. Any distinct values present in a single field in different messages from the same visit are concatenated to ensure capture of new or revised information submitted throughout the patient visit. In addition, several metadata variables are calculated: the first and last values of key date fields are determined, and the longest and latest entries in reason-for-visit data elements (Chief_Complaint, Triage_Notes, diagnosis variables) and the date of the first non-null diagnosis and disposition are identified. Indicator variables reflecting multiple values (within a single visit) for age, PID, MRN, UVID, and admit date are also created.

A new data table of these aggregated visits and visit-level metadata variables is saved as an R object and exported as a pipe-delimited text file.

Finally, the R script produces an HTML file which displays message-level and visit-level data quality metrics. At the message level, results on the identifier variables are provided. Several tables displaying visit-level data quality results by facility are provided, including a table of percent completeness of several key demographic and clinical variables, a table displaying percent of visits having multiple values of demographic and clinical variables that should be single-valued, and frequency tables of visits by age group and patient class. The script can also be easily extended to calculate and display other results or additional data quality metrics that may be developed in the future.

This script was developed with input from the BioSense Data Quality workgroup and was demonstrated on conference calls that led to the sharing among jurisdictions of results regarding identifiers.

Results

Since January 2013, Illinois has performed on-boarding and validation to incorporate 156 hospitals into the BioSense Platform. This data quality assessment script, scheduled to run quarterly, produces a facility-level report that is shared with hospital and vendor contacts to support re-engagement and improvements following initial onboarding validation.

Conclusions

Moving forward, IDPH will utilize this script in its local HL7 integration project to improve data quality and facility monitoring on a continuous basis.

Keywords

R programming; data quality; syndromic surveillance

Acknowledgments

We would like to thank Rosa Ergas, Harold Gil, Farah Naz, and Hailin Yu.

References

1. CDC NSSP, shared_biosense_functions.r [Computer software].
2. CDC NSSP, ILI.r [Computer software].
3. Gil H, Rennick M, Wiedeman C. UCEP-Biosense_DQ-5_12_14.r [Computer software].

***Serena Rezny**
E-mail: 'serena.rezny@illinois.gov

ISDS Annual Conference Proceedings 2015. 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.