OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Streamlining Foodborne Disease Surveillance with Open-Source Data Management Software

Michael Judd* and Karen Wong

Centers for Disease Control and Prevention, Atlanta, GA, USA

Objective

The "ledsmanageR", a data management platform built in R, aims to improve the timeliness and accuracy of national foodborne surveillance data submitted to the Laboratory-based Enteric Disease Surveillance (LEDS) system by automating the data processing, validating, and reporting workflow.

Introduction

The National Surveillance Team in the Enteric Diseases Epidemiology Branch of the Centers for Disease Control and Prevention (CDC) collects electronic data from all state and regional public health laboratories on human infections caused by *Campylobacter, Salmonella*, Shiga toxin-producing *E. coli*, and *Shigella* in LEDS. These data inform annual estimates of the burden of illness, assessments of patterns in bacterial subtypes, and can be used to describe trends in incidence. Robust digital infrastructure is required to process, validate, and summarize data on approximately 60,000 infections annually while optimizing use of financial and personnel resources.

Methods

We leveraged the robust and extensible programming facilities of the R programming language and the active community of R users to develop a data integration, processing, and reporting pipeline for LEDS via an internal software package we named "ledsmanageR". We designed all data retrieval, cleaning, and provisioning algorithms using tools from RStudio software packages[1–3] and tracked changes to source code and data using CDC's internal Gitlab server. We automated data validation requests to reporting partners by generating customizable emails directly from the R console[4]. We streamlined the data reconciliation process using OpenRefine[5], a point-and-click tool for cleaning big data. We automated generation of annual reports, a process that was previously manual, using parameterized RMarkdown documents. Staff epidemiologists performed design and implementation internally, requiring no external consulting.

Results

Developing our free and open-source software platform for national foodborne surveillance data management has saved the Enteric Diseases Epidemiology Branch thousands of dollars because we no longer depend on proprietary software requiring annual licensing fees. This transition occurred without any disruption in surveillance operations. Partial automation of email-based data validation and annual report generation processes reduced employee time requirement from one full-time position to one part-time position. The modular nature of ledsmanageR permitted LEDS to collect an expanded set of data elements with no changes to the core data processing and reporting workflow.

Conclusions

We developed and implemented a flexible tool that helps maintain the integrity of surveillance data and reduces the need for manual data cleaning, which can be laborious and error-prone. The user-friendly design features of ledsmanageR demonstrate that data management can be optimized using programming skills that are increasingly common among epidemiologists. Our work on improving the accuracy and efficiency of enteric disease surveillance has served as a proof of concept for plans to streamline data processing for other surveillance systems.

Keywords

R; Open Source; Surveillance

References

1. Wickham, H. (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse
2. Wickham, H; Hester, J; Francois R. (2016). readr: Read Tabular Data. R package version 1.1.1. https://CRAN.R-project.org/package=readr
3. Wickham, H and Miller, E. (2017). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 1.1.0. https://CRAN.R-project.org/package=haven
4. Premraj, R. (2016). mailR: A Utility to Send Emails from R. R package version 0.6. https://github.com/rpremraj/mailR
5. Ham, K. (2013). OpenRefine (version 2.5). Free, open-source tool for cleaning and transforming data. Journal of the Medical Library Association: JMLA, 101(3), 233. http://openrefine.org.

***Michael Judd**
E-mail: vuj4@cdc.gov