ISDS 2018 Conference Abstracts

Beginner R methods for syndromic surveillance data validation

Elyse Kadokura*

WA State DOH, Shoreline, WA, USA

Objective

To share practical, user-friendly data validation methods in R that result in shorter validation time and simpler code.

Introduction

There are currently 123 healthcare facilities sending data to the Washington (WA) State syndromic surveillance program. Of these facilities, 30 are sending to the National Syndromic Surveillance Program's (NSSP) production environment. The remainder are undergoing validation or in queue for validation. Given the large number of WA healthcare facilities awaiting validation, staff within the state syndromic surveillance program developed methods in R to reduce the amount of time required to validate data from an individual facility.

Methods

The dplyr package and R Markdown file format were used to more rapidly conduct syndromic data validation. Dplyr, written by Hadley Wickham, was created for easy data manipulation.[1] The syntax of this package is user-friendly, providing a function for almost every common data manipulation task and utilizing the piping operator from the magrittr package. Data fields of interest for syndromic surveillance are classified as required (R), required but may be empty (RE), or optional (O). For R or RE data fields, dplyr can be used to check for patterns of missingness as well as verify that the correct value sets are being used for code fields. For character fields, dplyr can be used to pull samples of free-text, calculate word or character counts, or search for string patterns of interest.

R Markdown makes it easy for users to create reproducible reports in many different document types including HTML, PDF, and Word.[2] R Markdown files combine R code chunks and plain text to create easy-to-read, professional data validation reports that can be used internally or shared with data submitters for their review.

Results

The amount of time spent validating any single facility has decreased significantly. This has allowed the number of facilities undergoing data validation at one time to increase from 12 to 22. However, the length of time between beginning and completing data validation per facility has not decreased. While reporting data issues to facilities takes less time, the lag in the validation process still occurs while waiting for facilities to correct these issues at the feed origination.

Conclusions

In order to increase the number of healthcare facilities that are sending production quality data more quickly, more resources need to be directed at providing facilities with support on how to correct data issues rather than solely reporting the problems.

Keywords

validation; surveillance; syndromic

References

1. Anderson, S. dplyr and pipes: the basics [Internet]. 2014 [cited 2017 Oct 10]. Available from: http://seananderson.ca/2014/09/13/dplyr-intro.html
2. Broman, K. Knitr with R Markdown [Internet]. [cited 2017 Oct 10]. Available from: http://kbroman.org/knitr_knutshell/pages/Rmarkdown.html.

***Elyse Kadokura**
E-mail: elyse.kadokura@doh.wa.gov