International Journal of Population Data Science

Journal Website: www.ijpds.org

Cecilia: an R package to automate data cleaning of administrative datasets

Garcia, AF[1], Palfy, M[1], and Vasquez, SA[1]

[1]SA NT DataLink

Introduction

Data linkage has considerable potential to improve health and society. Linking vast and detailed information across multiple administrative population-scaled data sources enhances the quality of existing data, empowers population health research, and produces objective evidence to inform policy decisions. In this context, data cleaning is crucial to minimise linkage errors.

Objectives and Approach

As dealing with the heterogeneity of administrative datasets is an acknowledged time-consuming task, the objective is creating a public and open-source R package to automate and report steps of data cleaning in a reproducible fashion.

The package automatically assesses variables and reports relevant information and issues for linkage purposes, then cleans the dataset based on problems found, reassesses the variables and reports results again.

It has a default cleaning procedure based on years of accumulated linkage knowledge and an interactive exploratory session to check variables individually. The report also includes all settings from both default and interactive session.

Results

The package accurately detected, cleaned and reported potential linkage problems in variables such as names, addresses and dates in so far 15 actual populational datasets from multiple sources, with a diverse range of format, content, and inconsistencies. The entire process took minutes rather than hours.

The reports correctly gathered, organised and presented all relevant information for linkage, in all distinct sections of the hyperlinked document, such as those related to the dataset, individual variables or settings used for cleaning. The different types of information included text, figures, data dictionaries, and frequency tables of detected issues, such as non-alphanumeric characters, annotation terms or suffixes and prefixes.

The output datasets had all evaluated variables with cleaned data plus extra columns containing only issues themselves or problematic records.

Conclusion/Implications

The package accelerates the data cleaning of linkage variables, automating time-consuming steps, providing pertinent information for linkage as well as cleaned datasets. The complete process is time-efficient and reproducible.

As the output dataset contains variables with cleaned data and detected issues, it allows assessment of the level of cleaning performed.

http://dx.doi.org/10.23889/ijpds.v3i4.853