CPRDTools: a CPRD data wrangling toolkit

Abstract

The analysis of large scale data, and moreover the analysis of large scale electronic health records data is become more commonplace. The ease and ability of modern data generation and capture means there is the potential across almost all industries to capture more data now, than ever before and that equates to large data resources, with CPRD no exception. These large data resources, as attractive and appealing as they may be to researchers, pose a significant problem - how to manage all that data? CPRDTools is a collection of wrapper R functions intended to simplify the loading, extraction and management of Clinical Practice Research Datalink (CPRD) GOLD specific and associated electgronic health records data. Allowing for the loading of CPRD and non-CPRD data into a SQLite based database, providing an efficient, secure and updatable repository for the data, keeing the origional source files intact. Through data queries, user-defined data are drawn allowing for subsetting, joining and filtering in a signle step, creating alaysis ready data.

Introduction

The CPRD is one of the largest longitudinal medical records databases in the world, supported by the National Institute of Health (NIHR) and the Medicines and Healthcare products Regulatory Agency (MHRA). It was first established in 1987 as the Value Added Medical Products (VAMP) dataset, this grew into the General Practice Research Database (GPRD) in 1993, before its final transition into CPRD in 2012 (Herrett et al. 2015).

PRD GOLD data are comprised of ten separate datasets: patient, practice, staff, consultation, clinical, additional clinical details, referral, immunisation, test and therapy (Padmanabhan 2017). These datasets contain their specific data and are linkable through a unique linkage key field, where key does not imply importance but a unique variable contained in two datasets allowing them to be joined, such as the CPRD-assigned and anonymised unique patient identifier `patid`.

Due to the size of CPRD, data extracted for research are spread over multiple text (`.txt`) files within each dataset to enable file transfer. This means that for CPRD clinical data, for instance, a researcher may receive their requested clinical data broken up over 25 individual text files. This is done to aid with file completeness and reduce turn-around times if errors are found. If an error occurred during the transfer of data between the data owner and the researcher

or in the extraction of the requested data by the data owner, the error can potentially be limited to only select files, requiring only their replacement with the corrected/error-free versions.

Often these text files are additionally zipped, or compressed, requiring that these files first be uncompressed or unzipped. These multiple files from each dataset (clinical, referral etc.) then need to be grouped together and amalgamated into a single `table`. Tables are a collection of data of the same shape, from the same dataset. In CPRD, each separate dataset (patient, practice, clinical etc.) forms a table. These tables are then stored in the SQLite database.

SQLite is an opensource, SQL based database engine (SQLite 2017). The use of a SQLite database provides an efficient storage solution, allowing for the loading, updating and maintenance of the database, all while retaining the original *raw* data files unaltered. An SQlite database permits for rapid data extraction through the use of data queries, drawing the required data from the database, allowing filtering, sub-setting, limiting and the joining of data in a single execution step.

This document aims to provide a simple and introductory overview of `CPRDTools` and its application to arbitrary (fictional) data. This data though are provided in the manner in which many CPRD data extract are received, where data are spread over multiple files and often located in sub-folders.

`CPRDTools` are loaded using:

```
library(devtools)
install_github("JamesCFSchmidt/CPRDTools")
library(CPRDtools)
```

CPRDTools overview

The functions within 'CPRDTools' broadly fall into three groups, categorised by their general application area: (1) - loading, (2) - maintenance and other tasks and (3) - extraction. **Loading** encompasses all functions used in the reading, converting and writing of data into the database including a function used to list all available files and all available CPRD files in a specified location, functions used in database maintenance, query speed improvements and date conversion functions fall under **maintenance and other tasks**, and finally, functions used to, and in the process of, drawing and retrieving data from the database fall within **extraction**.

Loading

References

Herrett, Emily, Arlene M. Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd van Staa, and Liam Smeeth. 2015. "Data Resource Profile: Clinical Practice Research Datalink (CPRD)." *International Journal of Epidemiology* 44 (3): 827–36. https://doi.org/10.1093/ije/dyv098.

Padmanabhan, Shivani. 2017. "CPRD Gold Data Specification." https://cprdcw.cprd.com/_docs/CPRD_GOLD_Full_Data_Specification_v2.0.pdf.

SQLite. 2017. *SQLite*. Charlotte, North Caolina: Hipp, Wyrick and Company, Inc. https://www.sqlite.org/index.html.