An R package for reproducibility research in data extraction

Yinxiaohe Sun, Mark K Salloway, Reuben Chong Wee Ng, Yilin Ning, Ying Chen, Stefano Calza, Chuen Seng Tan

***Abstract*—Reusing data to address multiple research questions is becoming a common process with increasing usage of electronic health records (EHRs) for public health research. Facilitating access to data stored as tables in a database can be a cumbersome process when it requires translation of researchers' data requirements specified on a paper into an electronic extraction. We developed an automatic data extraction workflow implemented as an R package, called RDataXMan, which aims to reduce translation and coding errors.**

I. Introduction

Currently, a manual data extraction process for reusing the EHR database [1] for research purposes is performed by our Data Science and Analytics (DaSA) team with expertise in Oracle and R. This extraction process creates a barrier between users and the database, especially among users with limited database and coding experience.

To ensure efficiency and reproducibility [2-3] of the data extraction step, we developed an automatic data extraction workflow as an R package (see Fig. 1), called RDataXMan (i.e., **R**'s **Data** e**X**traction **Man**agement). The package is written in R, a popular language among data scientists and analyst, because it can easily be extended to other database system and perform data analysis. The following principles were used when developing the R package:

- Reduce translation of users' requirement from the paper application form into code (e.g., generate electronic request forms).
- Utilize applications familiar to users to indicate data requirements (e.g., request form is an excel file with filtering and sorting options).

*Research supported by the Centre for Health Services and Policy Research SBRO10/NS01G from the National University Health Systems Pte Ltd.

Y.Sun and M,K,Salloway are with the Centre for Health Services and Policy Research, National University of Singapore (NUS)/ National University Health System (NUHS) (email: ephsy@nus.edu.sg/ephsmk@nus.edu.sg).

R.C.W.Ng is with the Lee Kuan Yew School of Public Policy, NUS (email:spprng@nus.edu.sg)

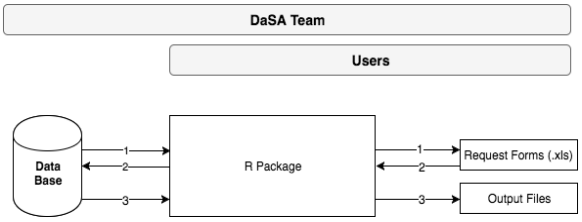
Y. Ning is with the Department of Surgery, Yong Loo Lin School of Medicine, NUS/NUHS (e-mail: e0013210@u.nus.edu).

Y.Chen is with the Saw Swee Hock School of Public Health, NUS/NUHS (e-mail: yingc@u.nus.edu).

S.N. Calza is with is with Department of Molecular and Translational Medicine, University of Brescia, and Department of Medical Epidemiology and Biostatistics, Karolinska Institutet (e-mail: stefano.calza@unibs.it).

C.S.Tan is with the Saw Swee Hock School of Public Health, NUS/NUHS (phone: +65-6601-3206, e-mail: ephtcs@nus.edu.sg).

Fig.1 Proposed new data extraction process



- Minimize input of information by users at each step (e.g., include a sheet in the excel file containing the specifications of arguments at the first step).
- Facilitate reproducibility research by utilizing a simple data extraction management workflow and file organization.

II. Process

We briefly describe the steps in our automated data extraction R package:

Step 1: DaSA team generates the electronic request forms from the package, which allows users to specify their inclusion criterion and variable lists requirements.
Step 2: Users indicate the items they want with an "x" in the selection column of the request forms, and submit the request forms to the database via the R package
Step 3: The package extracts and returns the data with its summary statistics as output files to users.

III. Conclusion

The R package developed translates the requirements of users into immediately executable data extraction steps. This empowers users with limited database experience to utilize complex EHRs datasets. Future work will include plans to make the R package accessibility to a range of databases besides Oracle, and allow multiple inclusion criteria and tables for variable for selection.

References

[1] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.," *AMIA Summits Transl. Sci. Proc.*, vol. 2010, pp. 1–5, 2010.

[2] "Journals unite for reproducibility," *Nature*, vol. 515, no. 7525, pp. 7–7, Nov. 2014.

[3] M. BAKER, "Is there a reproducibility crisis?," *Nature*, vol. 533, pp. 452–454, 2016.