Singapore Healthcare Management 2019

Reduce, Reuse, Recycle: Using R to create reproducible codes

Hanis Binte Abdul Kadir, Chia Sing Yi
Health Services Research Unit, Singapore General Hospital

Singapore General Hospital
SingHealth

INTRODUCTION

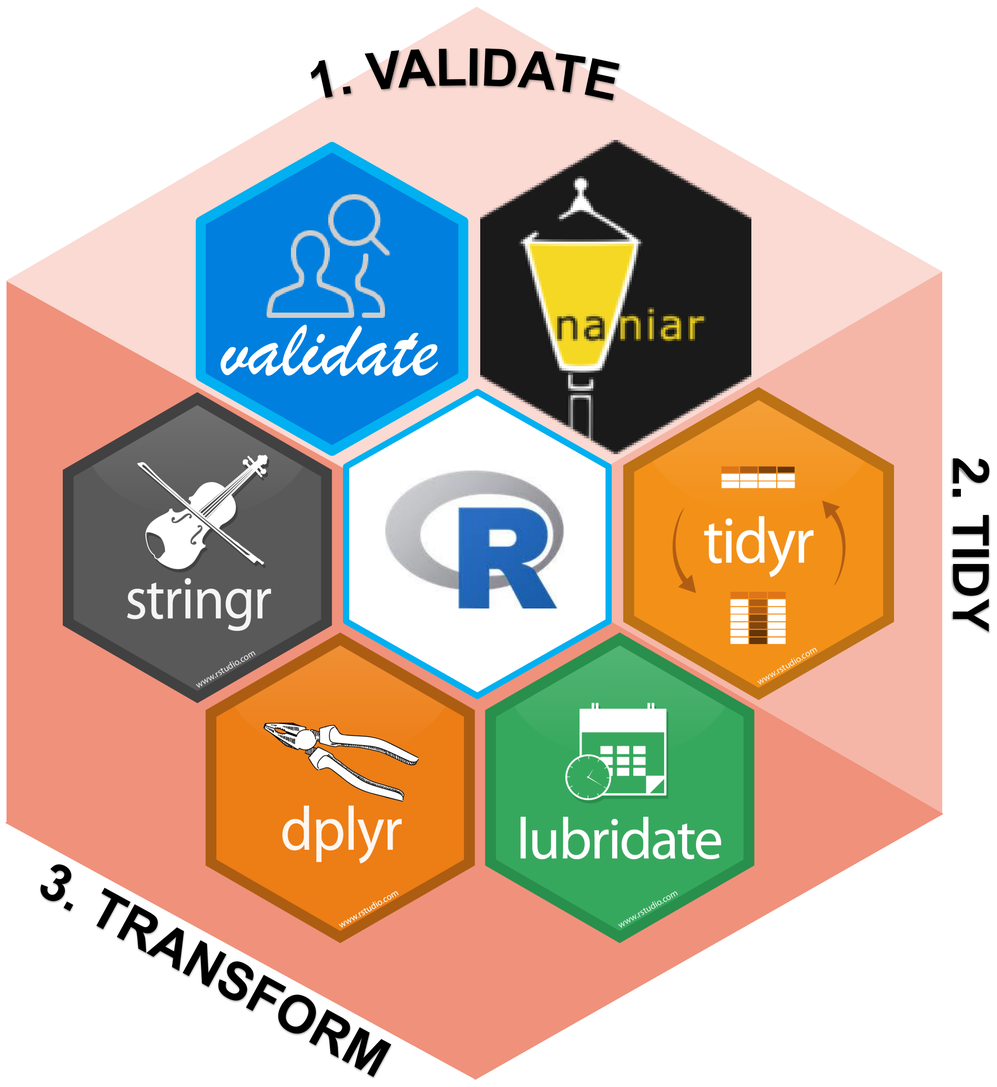
Electronic medical record is a major source of data for research, quality improvement and clinical audit. However, raw data extracted from the databases are often incomprehensible and further data processing is required to transform the data into usable form in order to derive insights.

AIM As data processing has become an integral part of these studies, our objective was to create reproducible codes to automate this process.

METHODOLOGY

The codes were written in R programming language. Data processing includes validating, tidying and transforming of data. 6 key packages were used in performing these steps.

	Package	Function	Example
1. VALIDATE	{naniar}	to assess if missingness in data is as expected	NT-proBNP is not a commonly ordered lab test, hence a high proportion of missing results is expected. On the other hand, height and weight are usually measured for every clinic visit. If there is a high proportion of missing values, the study team should inform the data extraction team to investigate.
	{validate}	to check data against domain knowledge and extraction logic	
2. TIDY	{tidyr}	to reshape data and present it in a tidy form	Diagnosis codes are presented in long format and require reshaping of data in order to derive the individual diagnoses for each record.
3. TRANSFORM	{dplyr}	to efficiently perform advanced data manipulation operations	Medication dosage instructions are captured as free text (e.g. Take 2 tablets 3 times a day). To calculate dosage per day, the quantity and frequency have to be extracted and transformed into numerical variables.
	{stringr}	to manipulate text data and obtain pieces of information from them	
	{lubridate}	to handle dates and times easily	



In order to acquire reproducibility, parameters were sensibly placed in the functions. This allows flexibility to accommodate different projects' requirements.

RESULTS

REDUCE TIME SPENT

- Creating reproducible codes reduces the amount of time spent in performing common tasks
- Reproducible codes can be used as many times, without having to re-write them or use copy-and-paste
- The use of parameters in these codes allows for very minor modifications when applying across different projects with similar requirements, or across different time points of the same project
- For example, we could specify the year as an input parameter, and change the year value accordingly to generate yearly reports

REDUCE HUMAN ERROR

- As the same algorithm is applied, the chance of making incidental mistakes when using copy-and-paste is eliminated
- The codes need to be updated at only one place, instead of many
- For example, if we had used copy-and-paste, and updated the variable name, we may have changed it at some places, but miss a few others
- This also ensures standardisation across different members in the team

CONCLUSION

There are many advantages to adopting reproducible R codes. We can validate, tidy, and transform data easily and efficiently since only minimal changes are required to process from one data to another. This does not only reduce duplication of effort and the amount of time spent, but also offers consistency and minimises human errors.

Acknowledgement: https://CRAN.R-project.org/