Digital Professionalism in Health and Care: Developing the Workforce, Building the Future
P. Scott et al. (Eds.)

doi:10.3233/SHTI220912

Data Management for Health Data Reuse: Proposal of a Standard Workflow and a R Tutorial with Jupyter Notebook

Antoine LAMER[a,b,c1], Sanae AL MASSATI[b,c], Chloé SAINT-DIZIER[a,b], Emile FARES[a], Emmanuel CHAZARD[c] and Mathilde FRUCHART[c]
[a]*F2RSM Psy - Fédération régionale de recherche en psychiatrie et santé mentale Hauts-de-France, F-59350, Saint-André-Lez-Lille, France.*
[b]*Univ. Lille, Faculté Ingénierie et Management de la Santé, F-59000, Lille, France.*
[c]*Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France.*

Abstract. The data collected in the clinical registries or by data reuse require some modifications in order to suit the research needs. Several common operations are frequently applied to select relevant patients across the cohort, combine data from multiple sources, add new variables if needed and create unique tables depending on the research purpose. We carried out a qualitative survey by conducting semi-structured interviews with 7 experts in data reuse and proposed a standard workflow for health data management. We implemented a R tutorial based on a synthetic data set using Jupyter Notebook for a better understanding of the data management workflow.

Keywords. Education; Data Science; Data reuse; Data management; Programming

1. Introduction

In clinical research, data are collected using three approaches. In the classical approach, data are manually and prospectively collected with a Clinical Report Form (the CRF), according to the question addressed by the research protocol [1]. In the second approach, the clinical registry, data are collected prospectively as well, but without a predefined research question [2]. The last approach, defined as data reuse, has emerged with the increasing implementation of electronic health records (EHRs) and consists in reusing data automatically recorded through EHRs to address a research question defined after data recording [3].

The data collected using the CRF are presented in the optimal format for conducting statistical analysis, therefore they do not need any further computation. It is not the case with data collected with the other methods, where it is rare to use them directly without modifications to suit the research needs. Indeed, in these approaches, data are always multidimensional, heterogeneous and time-dependent. Several operations are generally realized to select the relevant patients across the cohort, combine information from

[1]Corresponding Author, Antoine Lamer, ULR 2694, 2 place de Verdun, F-59000, Lille, France; E-mail: antoine.lamer@univ-lille.fr.

multiple tables, add new variables if needed, and reconstruct a consolidated unique table to answer the research question [4]–[6]. These operations are difficult to perform manually, due to the large number of records and are implemented with a computer program, such as R, python or SQL. In particular, the R package *tidyverse* offers many functions for reformatting fields, crossing several tables, filtering, grouping and sorting records [7].

There are various resources to help to handle data management functions, such as packages documentation, tutorials and Massive Open Online Course (MOOC). However, beyond the implementation of a function in a standalone way to filter records for example, it is the whole chain, from the raw data to the final table, that must be mastered. The objective of this article is to highlight the main steps and operations required in the data management of heath data for secondary use. To do so, we interviewed experts in data reuse to define the main steps of data management. From that point on, we built a R tutorial to guide beginners and help them to discover the process and practice manipulating the main R functions.

2. Methods

We carried out a qualitative survey by conducting semi-structured interviews with 7 experts in data reuse who perform regularly data management, data cleaning and feature extraction. The objective of the interviews was to identify the main steps of the data management process and to propose a standard workflow. We asked experts to describe the usual operations they apply to transform raw data files into clean data sets before performing statistical analysis or data visualization. We have also asked them for the R packages and functions they use for these operations.

In a second step, we generated a virtual data set containing the data we usually find in a database: patients, hospital stays and drug administrations. This data set includes typical problems that are usually encountered when dealing with this type of projects, e.g., missing data, unformatted date field, duplicates, multiple tables. Figure 1 shows 5 records for each of these three tables, with the cardinality between tables.

Patient

patient_id	birth_date	sex	location_id
1	19900102	M	2
2	19870204	F	1
3	19601020	F	2
4	19671205	F	3
5	19561128	M	1

1 n

Hospital stay

hospital_stay_id	patient_id	hospital	admission_date	discharge_date
1	1	Hospital A	20200103	20200103
2	1	Hospital C	20210224	20210224
3	2	Hospital A	20210201	20210201
4	3	Hospital A	20210810	20210810
5	3	Hospital A	20210812	20210812

1 n

Treatment

treatment_id	hospital_stay_id	treatment	posology	unit	treatment_date
1	1	T1	2	mg	20200103
2	1	T2	1	mg	20200103
3	2	T1	2	mg	20200224
4	3	T2	2	mg	20210201
5	3	T1	1	mg	20210201

Figure 1. Experimental data set.

Based on the steps and the operations identified in the interviews, we have developed a R tutorial. Its main objective is to manage the three data tables in order to answer the

following research goals: in the context of a hospitalization, we aim to compare the duration of hospitalizations, the number of drug administrations and the drugs administered with the sex of the patient, and different age groups (i.e. <18, 18-34, 35-54, 55-74, >=75 years old). The tutorial includes the following elements: context and research question, exploration of the original data sets, definition of the data management plan and implementation of the data management. Each step of the implementation is mentioned with an explanation of the R function, an illustration, a R block code and the display of the result.

This tutorial was implemented using Jupyter Notebook, an interactive computational laboratory notebook, which can work with codes in many different programming languages such as Python, Java, R, or Julia [8]. Jupyter Notebook allows the smooth integration of code and narrative text (in Markdown syntax) into a single document that can be executed and edited immediately. Jupyter Notebook is accessed through a web browser which makes it practical to use, both locally and with remote access on a server. In the case of remote access, no installation is required on the user's computer except a web browser, additional packages could be installed beforehand by an expert user on the server. These points encourage the use of notebooks by beginners, especially in courses [9].

3. Results

After summarizing the interviews, we identified the following main steps in the data management process:

- Exploration of source data: for each source table, we identify the statistical unit and the type of each variable and we evaluate the data quality (e.g., missing data, duplicates, abnormal values).
- Definition of the optimal data table required to carry out the analysis: statistical unit (e.g., one row per patient, one row per hospital stay), relevant variables and inclusion criteria.
- Data management workflow to transform the raw data to the optimal data table in the form of a flat file.
- Implementation of the workflow.

Experts suggested 16 typical operations they used when implementing the data management plan. These operations are usually performed with functions from the following R packages: *base*, *utils*, *dplyr*, *lubridate*, *stringr* and *tidyr*. These operations can be functionally grouped into the following categories : exploring raw data, formatting, adapting the data structure, merging of tables, and customizing the final table. The table 1 displays these operations, as well as the related R functions and packages.

Table 1. Typical data management operations and appropriate R functions

Steps	Data management operations	R functions
Exploring raw data	Identify the statistical unit of each table and the fields which characterize it.	utils::str utils::head
Formatting	Transform strings into numbers or dates.	lubridate::ymd
	Remove duplicate records	base::unique
	Rename columns homogeneously	dplyr::rename
Adapting the data structure	Group rows by a grouping column and summarize values of other columns to obtain one value for each modality of the grouping column.	dplyr::group_by dplyr::summarize
	Pivot row-oriented records into column-oriented records.	tidyr::pivot_wider
Merging of tables	Combine several tables to obtain a single table with information available.	base::merge
Customizing the final table	Filter records according to inclusion criteria or quality indicators.	dplyr::filter base::is.na
	Add new variables based on existing ones. E.g. cut a quantitative variable, compute a delay between two dates, extract a subvalue from a string.	dplyr::mutate base::cut lubridate:: stringr::str_replace
	Select relevant variables and delete the unnecessary ones.	dplyr::select

Based on the interviews, we propose a tutorial that covers the data management functions in the following order (Figure 2). First, source tables are reformatting, when needed, records are deduplicated and variables are renamed homogeneously. Then, the tables structures are adapted to fit the statistical unit of the study (here, the hospital stay). After that, tables are merged to obtain a single table. Finally, new variables are calculated in the final tables, records are filtered to apply the inclusion criteria of the study and useless variables are dropped to keep only the relevant ones.

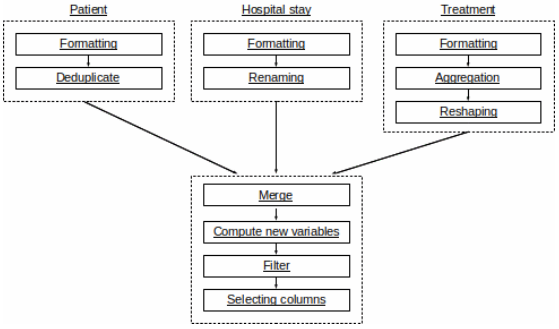

Figure 2. Standard workflow of data management operations.

The tutorial, the data and associated illustrations are available online in our gitlab repository [10]. The figure 3 display a section of the tutorial.

4. Discussion and conclusion

Based on a qualitative survey involving experts in data reuse, we synthesized and illustrated the main operations usually performed when data managing with a R. For each of these operations, we proposed and illustrated a R function. The tutorial is available

online and open to improvements. The next step will be to deploy and test the tutorial with students.

Compared to Rmarkdown, which also integrates code and narrative text, Jupyter Notebook prevents the need to work locally and to have to install the necessary libraries himself. We did not focus on all functionalities available for each function, and we preferred to give an extensive overview of the process. Obviously, there are other ways to do the transformations we have proposed in the tutorial. The content of this tutorial could be adapted for other learning materials, such as ou datacamp, MOOC. This tutorial could also be adapted in SQL and Python, two programming languages widely spread in data science.

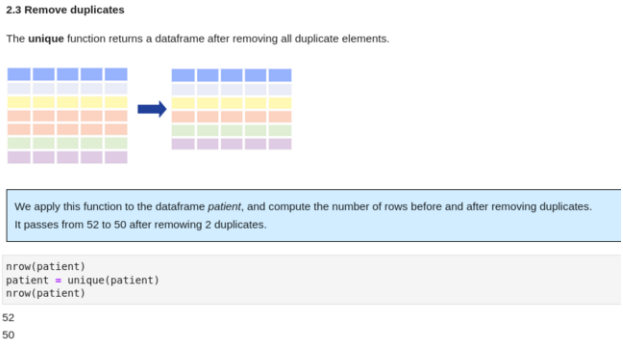

Figure 3. A preview of the tutorial with the explanation of the unique function, the appropriate illustration and the execution of a R chunk.

References

[1] B. Krishnankutty, S. Bellary, N. B. R. Kumar, and L. S. Moodahadu, "Data management in clinical research: An overview," *Indian J. Pharmacol.*, vol. 44, no. 2, pp. 168–172, 2012, doi: 10.4103/0253-7613.93842.

[2] H. Mauch, J. Kaur, C. Irwin, and J. Wyss, "Design, implementation, and management of an international medical device registry," *Trials*, vol. 22, p. 845, Nov. 2021, doi: 10.1186/s13063-021-05821-5.

[3] C. Safran, "Reuse of Clinical Data," *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 52–54, Aug. 2014, doi: 10.15265/IY-2014-0013.

[4] A. Lamer, M. Jeanne, G. Ficheur, and R. Marcilly, "Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse," *Stud. Health Technol. Inform.*, vol. 221, pp. 102–106, 2016.

[5] E. Chazard *et al.*, "Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features," *Stud. Health Technol. Inform.*, vol. 255, pp. 15–19, 2018.

[6] E. Chazard *et al.*, "'Book Music' Representation for Temporal Data, as a Part of the Feature Extraction Process: A Novel Approach to Improve the Handling of Time-Dependent Data in Secondary Use of Healthcare Structured Data," *Stud. Health Technol. Inform.*, vol. 290, pp. 567–571, Jun. 2022, doi: 10.3233/SHTI220141.

[7] "Tidyverse." https://www.tidyverse.org/ (Accessed Jun. 25, 2022).

[8] "Project Jupyter." https://jupyter.org (Accessed Jan. 16, 2022).

[9] K. M. Mendez, L. Pritchard, S. N. Reinke, and D. I. Broadhurst, "Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing," *Metabolomics*, vol. 15, no. 10, p. 125, 2019, doi: 10.1007/s11306-019-1588-0.

[10] "Resources · main · health_data_science / health_data_science_tutorials · GitLab," *GitLab*. https://gitlab.com/d8096/health_data_science_tutorials/-/tree/main/ressources (Accessed Jun. 28, 2022).