Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230288

OpenDataPsy: An Open-Data Repository with Standardized Storage and Description for Research in Psychiatry

Chloé SAINT-DIZIER[a,b], Antoine LAMER[a,b,c], Majda ZAANOUAR[b,c],
Alina AMARIEI[a] and Paul QUINDROIT[c,1]
[a] *Fédération régionale de recherche en psychiatrie et santé mentale (F2RSM Psy), Hauts-de-France, 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. Sharing health data could avoid duplication of effort in data collection, reduce unnecessary costs in future studies, and encourage collaboration and data flow within the scientific community. Several repositories from national institutions or research teams have making their datasets available. These data are mainly aggregated at spatial or temporal level, or dedicated to a specific field. The objective of this work is to propose a standardized storage and description of open datasets for research purposes. For this, we selected 8 publicly accessible datasets, covering the fields of demographics, employment, education and psychiatry. Then, we studied the format, nomenclature (i.e., files and variables names, modalities of recurrent qualitative variables) and descriptions of these datasets and we proposed on common and standardized format and description. We made available these datasets in an open gitlab repository. For each dataset, we proposed the raw data file in its original format, the cleaned data file in csv format, the variables description, the data management script and the descriptive statistics. Statistics are generated according to the type of variables previously documented. After one year of use, we will evaluate with the users if the standardization of the data sets is relevant and how they use the dataset in real life.

Keywords. Open data, psychiatry, data reuse, research

1. Introduction

Sharing health data could avoid duplication of effort in data collection, reduce unnecessary costs in future studies, and encourage collaboration and data flow within the scientific community. It would also improve the reproducibility and transparency of clinical research by allowing researchers to validate each other's results and reduce the impact of publication bias [1]. Many institutions are now making their datasets available. For example, in France, open platform contains French public data about demographics, COVID-19, elections, energy consumption, health and employment [2]. Researchers are increasingly encouraged to deposit their work on these platforms. These data were previously crossed for a study on the relationship between the immigrant rate and health

[1] Corresponding Author: Paul QUINDROIT, E-mail: paul.quindroit@univ-lille.fr.

status [3]. Institutions data are generally aggregated at a geographical level (e.g., a city or region) over a defined period (e.g., day, month, year). More specific repositories also propose individual data [4], from generic domain or specific data as hospital intensive care units [5–7]. The datasets are spread over different repositories, with heterogeneous formats, nomenclatures and descriptions.

The objective of this work is to propose a standardized storage and description of open datasets dedicated to psychiatry and mental health.

2. Methods

We have selected useful datasets for research in psychiatry available as open access on various national or international data sharing platform. They provide ecological variables which often supplement individual psychiatric variables (e.g., in a cohort) to gain information about their environment and adjust the statistical models. We studied the format, nomenclature (i.e., files and variables names, modalities of recurrent qualitative variables) and descriptions of these datasets.

We had to select a unique file format (i.e., type of file, column separator, decimal character, header, encoding), standardized files names, variables names and description. We also provided a standardized description of the variables composing the data file with a non ambiguous and meaningful variable name for printing and the type of variables (i.e., binary, qualitative, continuous quantitative, discrete quantitative). We did not take into account non relevant variables as unique identifier, qualitative variable with too much modalities and variables (e.g., zipcodes) corresponding to the statistical unit of the file. For each of the datasets, we kept the original dataset and we added the standardized file. We also provided a data management script when necessary. We developed a script to automatically generate descriptive statistics based on the cleaned and managed data file and the description of the variables. Statistics are generated according to the type of variables previously documented.

All the documents are available in a gitlab directory, with one folder per dataset and a README file to describe each folder [8].

3. Results

3.1. Source files

We selected 8 datasets, covering the fields of demographics, employment, education and psychiatry. We also provided a dataset giving a correspondence between two French zip codes nomenclatures. Raw files were available in heterogeneous formats (csv in French or International format, xls, xlsx), delimiting character in the case of flat files, decimal character ("," and ".") and encoding (LATIN-1, LATIN-3 and UTF8). Some datasets formats were often of poor quality, with empty lines at the top or bottom, specifying the source and the generation date. A file presenting social disadvantage indicator was generated based on several other variables: tax income, unemployment, workers, graduates.

3.2. Standardization

Each repository contains 5 files: the raw data file in its original format with a prefix for the type of file and a suffix for the years covered by the file, the cleaned data file in csv format, the variables description, the data management script and the descriptive statistics. The table 1 provides the standardized names and description of each file.

Standardization of the cleaned data file complies with the following characteristics:

- File format: csv, column separator ",", decimal character "."
- Variable names: lowercase, no special characters except the underscore
- Standard modalities for sex (M/F)
- Date in YYYY-MM-DD (year, month, day)
- Frequent Standardization of variable names for the most frequent ones (e.g., insee_code, postal_code).

Each dataset was documented with a README file at the root of the folder. The elements of this file are described in table 2, The figure 1 represents the folder dedicated to the finess *(FIchier National des Etablissements Sanitaires et Sociaux, French designation for the national register of health and social establishments)* dataset, a list of all French health facilities. The figure 2 is the subdirectory containing the raw finess dataset and the R data management file.

Table 1. Standardized name and description of files available for each dataset

Files	Description
raw_XXX_year.xslx	Source file without modification
dm_XXX_year.R	Data management script for cleaning and standardizing the raw data file
cleaned XXX_year.csv	Cleaned and standardized file
variables_XXX_year.csv	Description of the variables
desc_XXX_year.html	Descriptive statistics

Table 2. Documentation items contained README file

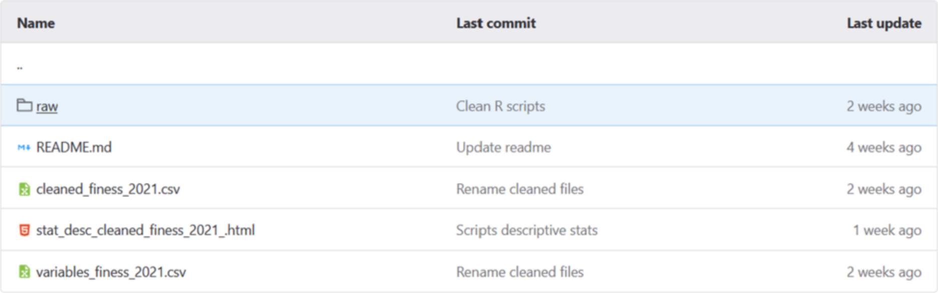
Section	Description
Description	Context (e.g., country and date of generation), productor (e.g., institution, company), statistical unit, type of data
Overview	Display of the five first rows
License and reuse conditions	Name of the license and/or link with the source page with use conditions
Raw data source	Link to the source page where the data were downloaded
References	Link to data productor website or bibliographic references

Table 3. Statistical indicators and graphics appropriate for each type of variable

Type of variable	Statistics	Graphics
Binary variable	Count, percentage, percentage of missing data	Doughnut
Qualitative variable	Count, percentage, percentage of missing data	Barplot
Continuous quantitative	Median IDR, min, max, percentage of missing data	Histogram and density
Discrete quantitative	Median IDR, min, max, percentage of missing data	Barplot
Date	Min, max, number of events per date with quartile	-

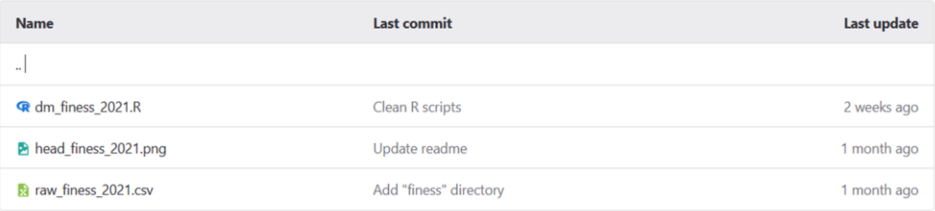
3.3. Automated descriptive statistics

For each type of variable, we propose appropriate statistical indicators and graphics, described in table 3. The figure 3 represents descriptive statistics for the variable “Number of bed-days for full time hospitalization”, available in the dataset dedicated to annual hospital activity in psychiatry.



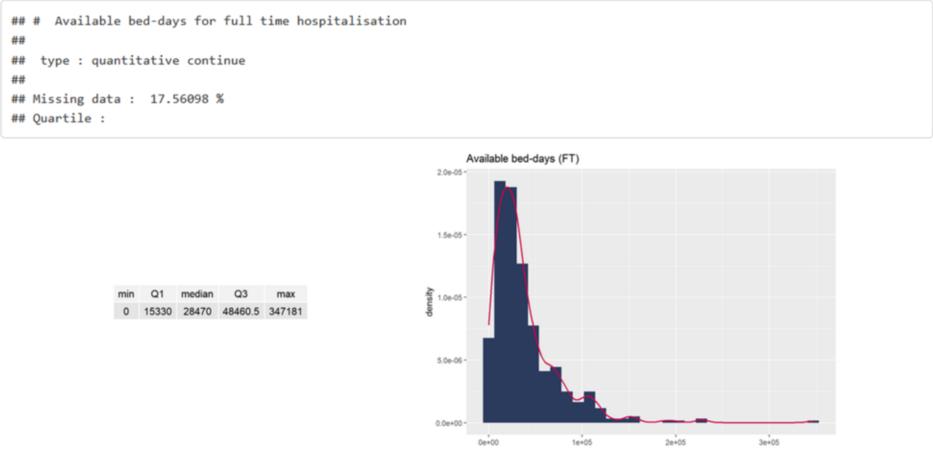
Name	Last commit	Last update
..		
raw	Clean R scripts	2 weeks ago
README.md	Update readme	4 weeks ago
cleaned_finess_2021.csv	Rename cleaned files	2 weeks ago
stat_desc_cleaned_finess_2021_.html	Scripts descriptive stats	1 week ago
variables_finess_2021.csv	Rename cleaned files	2 weeks ago

Figure 1. Repository for finess dataset, a list of all French health facilities.



Name	Last commit	Last update
..		
dm_finess_2021.R	Clean R scripts	2 weeks ago
head_finess_2021.png	Update readme	1 month ago
raw_finess_2021.csv	Add "finess" directory	1 month ago

Figure 2. Subdirectory containing the raw finess dataset and the R data management file.



```
## #  Available bed-days for full time hospitalisation
##
##  type : quantitative continue
##
## Missing data :  17.56098 %
## Quartile :
```

min	Q1	median	Q3	max
0	15330	28470	48460.5	347181

Figure 3. Automated descriptive statistics for the dataset dedicated to annual hospital activity.

4. Discussion and Conclusions

In this study, we implemented OpenDataPsy, a standardized storage and description of open datasets dedicated to psychiatry and mental health. These datasets are directly ready to be used, after being transformed into a common standard format. The description of variables allowed to automatically generate descriptive statistics of each dataset, in common format.

It remains to manage the recovery of the data sets when they are updated on the initial source site. We also need to spread the repository inside the psychiatric community. After several months of use, we will evaluate with the users if the standardization of the data sets is relevant and how they use the dataset in real life. In the future, we will feed the directory with new datasets. In particular, we plan to integrate free text data coming from forum and social media. It will be necessary to anonymise this data before sharing.

References

[1] Shahin MH, Bhattacharya S, Silva D, Kim S, Burton J, Podichetty J, Romero K, Conrado DJ. Open data revolution in clinical research: opportunities and challenges. Clinical and Translational Science. 2020 Jul;13(4):665-74. doi:10.1111/cts.12756.

[2] Inventaires thématiques de données - data.gouv.fr, (n.d.). https://www.data.gouv.fr/fr/pages/thematiques-a-la-une/ (accessed November 17, 2022).

[3] Perrot J, Hamel JF, Lamer A, Levaillant M. The Relationship between the Immigrant Rate and Health Status in the General Population in France. Journal of Personalized Medicine. 2021 Jun 30;11(7):627. doi:10.3390/jpm11070627.

[4] Banzi R, Canham S, Kuchinke W, Krleza-Jeric K, Demotes-Mainard J, Ohmann C. Evaluation of repositories for sharing individual-participant data from clinical studies. Trials. 2019 Dec;20:1-0. doi:10.1186/s13063-019-3253-3.

[5] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation. 2000 Jun 13;101(23):e215-20.. doi:10.1161/01.cir.101.23.e215.

[6] Johnson A, Bulgarelli L, Pollard T, Celi LA, Mark R, Horng IV S. MIMIC-IV-ED. PhysioNet. 2021. doi:10.13026/S6N6-XD98.

[7] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJ, Kompanje EJ, Kaplan L, Bailey H, Kesecioglu J, Cecconi M, Churpek M, Clermont G. Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force: Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med. 2021 Jun 1;49(6):e563-77. doi:10.1097/CCM.0000000000004916.

[8] health_data_science / open_data_psy · GitLab, GitLab. (n.d.). https://gitlab.com/d8096/open_data_psy (accessed January 4, 2023).