Plan Overview

A Data Management Plan created using DMPTool

Title: Metabolomic Evaluation of SARS-COV-2 Infections

Creator: Danilo Bernardineli

Affiliation: Non Partner Institution

Principal Investigator: Danilo Bernardineli

Data Manager: Danilo Bernardineli

Funder: Digital Curation Centre (dcc.ac.uk)

Funding opportunity number: 50621

Template: Digital Curation Centre

Last modified: 04-09-2020

Copyright information:

Metabolomic Evaluation of SARS-COV-2 Infections

Tabular metabolomics dataset for 450 patients

* 2 samples per patient

    * The first one is on the onset of internment

    * The second one is on discharge or aggravation of the condition

* Each sample contains 99 primary metabolites

* Format: CSV

* Expected minimum volume: 900 rows / 100 columns

Tabular patient metadata with categorizations

* Complication degree

    * No complications

    * Complications with no intensive support

    * Complications with intensive support

* Age range

    * Between 10 and 39 years old

    * Between 40 and 59 years old

    * Between 60 and 79 years old

    * Older than 80 years

* Format: CSV

* Expected minimum volume: 450 rows / 3 columns

The metabolomics dataset will be acquired through sputum collection and the standard protocol around it.

The patient metadata will be collected through administrative questionnaires by the associated hospitals, and transformed into an tabular CSV file.

Once retrieved, the files are going to be upload to an public Git repository at https://gitlab.com/ion_ia/covid19_metabolomics, which has version control by default. Any changes on the files, or a posterior QA processes, will be documented on the commit log and the CHANGELOG.md at the root directory.

The data will accompany an README.md containing the context and directions for it. Also, the public repository

containing the data will have all pertinent documentation, like the analysis code, the research project, and the published content around it.

Also, regarding metadata, we'll adopt an multiple header standard, were the first lines of the CSV files indicates variable name, group and subgroup.

The ethics around the retrieval of data will be regulated through an ethical agreement between the researchers and the health providers.

As for data sensitivity, it will be handled through tokenization and anonimization of all possible patient identifiers, like bar-codes and internal ids. The dictionary for mapping those tokens will be stored on an private repository.

The non-sensitive data will be donated to the Brazilian public health provider (SUS - Sistema Único de Saúde), and all the generated data and analysis code will be published in an to-be defined public commons license, like MIT License.

The data will be stored and backed up in an Gitlab repository, which contemplates and takes care of all storage, security and backup issues.

The repository with the metabolomics and metadata will be public, however the token dictionary for identifying the patients will be secured on an private repository on which only selected IONMedicine colaborators will have access.

All data on the public repository (metabolomics and patient metdata), is to be retained and shared on long-term. We vouch for storing it for at least 5 years, and the data is going to be aggregated into large-scale metabolomics datasets, like MetabolomeXchange, allowing for contribution on large-scale aggregated studies.

As for the data on the private repository (patient identity dictionary), they are going to be retained for 1 year, with an deletion scheduled for afterwards.

The data will be stored on Gitlab, which is an consolidated provider of code and data repositories. No efforts, costs or time required.

Through the https://gitlab.com/ion_ia/covid19_metabolomics git repository, and through pushing it on data finder mechanisms, like DataCite, Kaggle, and others. We intent to get an persistent identifier through DataCite.

The data will be available as soon we have it. If needed, modifications to the data will be handled through the Git versioning.

No

The responsible for the data management will be the Artificial Intelligence division of IONMedicine. Tasks will be divided accordingly.

The creation and maintenance of the data repositories, as well as the publishing of it on several data providers like Kaggle, will be attribution of Danilo L. Bernardineli

No additional resources needed