COVID-19 in MS

Global Data Sharing Initiative

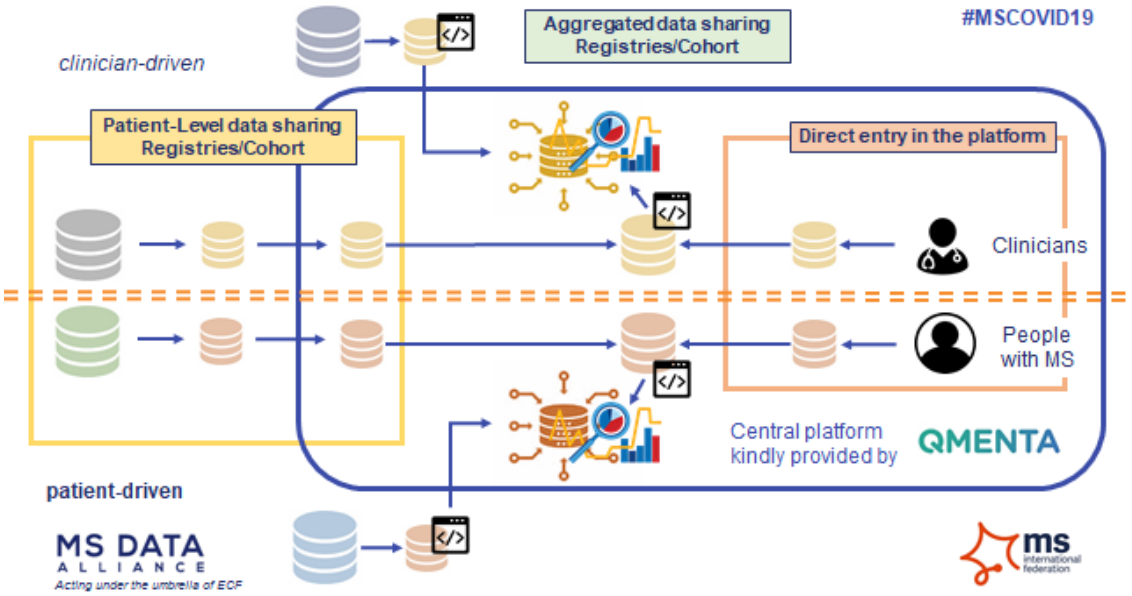
Analyses plan 2020

Important note: we have *fine-tuned* the script so we only query the data we really need for downstream analyses. Last year, we explored more options to define the final model. To reduce the privacy risk of the federated pipeline, we have eliminated many of these queries. This document only summarizes the scripts that will be leveraged into 2021.

Summary approach

Figure 1 summarizes the high-level overview of the approach. We recommend the implementation of a COVID-19 in MS core dataset in as many different data initiatives (registries/cohorts) as possible. The core variable set can be downloaded via the MSDA website (link). The following groups of variables were deemed important: COVID-19 infection, COVID-19 severity, COVID-19 treatment, demographic information, MS history and severity, information on DMT use and comorbidities, and selected lifestyle behaviours, particularly smoking.

Figure 1: high-level overview of the approach



Clinician-driven as well as patient-driven initiatives participate, but these data sources are currently analysed and interpreted separately because of 2 main reasons:

- To reduce the risk of duplicates (data entered by PwMS and data entered by clinicians on the same patients)
- First exploratory analyses have shown that patient-driven and clinician-driven initiatives are prone to different biases. Indeed, patient-driven initiatives mainly report milder COVID-19 bases and clinician-driven initiatives mainly report more sever COVID-19 cases in MS.

Three main “data-flows” are defined:

- “*Direct entry in the platform*”: We provide an option to directly enter data into our central platform (a distinction is made between a Clinician reported fast module for data entry by healthcare professionals and a Patient reported fast module for data entry by PwMS). The clinician-reported fast module might be of benefit to healthcare professionals who only have the resources to collect the core dataset.

- “*Patient-level data sharing via participating registries/cohorts*”: We invite all MS registries and cohorts to regularly share their COVID-19 core dataset (‘export’) into the central platform
- “*Aggregated data sharing via participating registries/cohorts*”: Some registries do not share patient-level data, but share results of specific scripts. We refer to these registries as “federated registries”. The script assumes the data is harmonized locally to the COVID-19 core dataset as described in the dictionary. After running the scripts locally, the counts are shared and combined with the counts of the data inside the central platform.

In 2020, we focused on exploratory and descriptive analysis because of following main reasons:

- The data counts were **too low** to allow trustworthy models that incorporate many confounding variables.

- A large proportion of the data was currently coming from the federated registries. Because we are currently working in the absence of a fully operational federated infrastructure, we are technically restricted to sharing “counts”. This means we are limited in the complexity of our federated analysis (e.g. properly running multivariate logistic regression models requires a more developed federated approach, which is currently not in place).

More detailed information on the stages of the pipeline

Data quality assessment and enhancement

To improve the quality of the data continuously over time, we have been working on setting up a data quality assessment and enhancement pipeline. This pipeline consists of two major parts:

- Unambiguously defining new variables that are used in downstream analysis (e.g. defining COVID-19 'suspected' and 'confirmed' cases, categorizing continuous variables to allow aggregation of the counts, …)
- Pre-defining PASS and FAIL criteria for variables (e.g. negative ages, unrealistically high numbers for height, …). Variables that "FAIL" are flagged and this information is to the registry custodians, allowing repair of failed variables in the next upload. Simultaneously, failed variables are cleaned and pre-processed so the records can still be incorporated into the downstream analysis. So in summary: Currently a 'fail' means the following action: "*set failed variable to missing, flag the variable, keep the entry (row)*"

IMPORTANT NOTE: we **always** assume the variables are properly mapped to the format as described in the dictionary. Automated format checks are incorporated when uploading patient-level data in the central platform. These automated format checks are currently not possible in our federated registries.
CALL-TO-ACTION to our **federated** registry custodians: please make sure the data is properly transformed before running the scripts.

Table A1 and A2 in the AppendixDocument summarizes the current assumptions and definitions that are incorporated in the scripts (same scripts are used in the central platform as sent to the federated registries). If you have any questions or comments, please do not hesitate to let us know (lotte.geys@uhasselt.be).

Feasibility study

Here, we are assessing the distribution, counts and completeness of the individual variables as well as the variations across the different contributing data sources. This will allow us to:

- learn more about the biases and inherent imperfection of the initiative. This makes it possible to interpret and contextualize the results
- assess the feasibility and power to address more fine-tune research questions.

We use an existing tool to profile the variables. This tool is a python package: pandas_profiling. Given a pandas dataframe, this tool can automatically generate a report giving basic statistics about the dataset. For example, it can give for each variable a quick summary, including the percentage of missing values, as well as basic descriptive statistics. It can also compute correlations and highlight pairs of variables having high correlation, for different correlation coefficients. This report can then be exported to a structured format, such as json or html.

Analysis - aggregation and visualisation

Scripts are called “query2_clinicians.py”” and “query2_patients.py” in the package “MSDA_Query2”

Request total counts (frequency procedure) on:
BY report_source
AND BY covid19_diagnosis
AND BY covid19_admission_hospital or covid19_icu_stay or covid19_ventilation or covid19_outcome_death or covid19_outcome_ventilation_or_ICU
AND BY dmt_type_overall
AND BY by age_in_cat
AND BY MS_type2
AND BY sex_binary
AND BY EDSS_in_cat2

We use this data to:

- build an interactive tool is created allowing user-friendly visualisation of these aggregated counts. This allows us to explore the risk factors visually.
- To run a (logistic) regression. The data source is added as a mixed effect.