Using machine learning to model the prognosis of multiple sclerosis patients

A Data Management Plan created using DMPonline.be

Creators: Robbe D'hondt, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1S38023N

ID: 196786

Start date: 01-11-2022

End date: 31-10-2026

Project abstract:

Multiple sclerosis is a complex disease with a highly heterogeneous disease course. In this project, we will develop machine learning methods to study this disease course using a combination of demographical, clinical, genomic, and radiomic data. This study is done on the patient level (individualized prognosis models), on the group level (patient stratification models), and on the population level (identification of prognostic biomarkers). With this study, we address pressing needs from both the clinicians, who want to start the right treatment for the right patient as soon as possible, as well as from the pharmacological industry, which wants to develop new treatments as cost-efficiently as possible. In the process, we will contribute to the machine learning literature on genetic data analysis, multi-target learning (i.e. combining outcomes of different data types), recurrent event survival analysis, and clustering (i.e. dynamic over time). Specifically, the models developed in this thesis will need to deal with sporadic and irregularly sampled time series, missing values (even in the outcome space), and confounding factors. To maximize the probability of integration in clinical practice, we additionally ensure that the models are explainable, reliable, and trustworthy.

Last modified: 26-04-2023

Using machine learning to model the prognosis of multiple sclerosis patients
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
UZ Leuven	Clinical, radiomic, and genomic data for ~700 patients	Reuse existing data	Digital	Observational	.csv	<100GB	
MSBase	Clinical data for ~50,000 patients	Reuse existing data	Digital	Observational	.csv	<1GB	
MultipleMS	Genetic data for ~20,000 patients	Reuse existing data	Digital	Observational	.csv	<100GB	
Processed data	Preprocessed data, ready for machine learning modeling	Generate new data	Digital	Compiled/aggregated data	.csv	<1GB	
Analysed data	Machine learning models, graphs, tables, text	Generate new data	Digital	Compiled/aggregated data	.pkl, .pdf, .txt	<1GB	
Code	Computer programming code (data analysis, machine learning algorithms)	Generate new data	Digital	Software	.txt	<100MB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- UZ Leuven: The laboratory for neuro-immunology in UZ Leuven, led by the PI's professor An Goris (my co-promotor) and professor Bénédicte Dubois (neurologist who collected the data).
- MSBase: An international online registry combining data of multiple sclerosis patients from hundreds of centers around the world. Access is granted through a research proposal with as PI's professor Bart Van Wijmeersch and professor Liesbet Peeters.
- MultipleMS: A horizon 2020 project led by Karolinska Institute in Sweden.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

Multiple sclerosis patient data.
Ethical approval references:

- UZ Leuven: S59940 and S60222 (Ethics Committee Research UZ/KU Leuven).
- MSBase: CME2020/069 (Committee for Medical Ethics UHasselt).
- MultipleMS: To be obtained later in the project (when specific agreements on the collaboration have been made).

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We will process pseudonymized retrospective data from multiple sclerosis patients, held by hospitals. This involves demographical

characteristics, measurements taken at hospital visits, genetic data (10^7 SNPs), radiomic features (information from MRI scans).
Privacy compliance references:

- UZ Leuven: G-2020-2436.
- MSBase: G-2020-2741.
- MultipleMS: To be obtained later in the project (when specific agreements on the collaboration have been made).

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

The developed software may result in valorisation potential, e.g. for companies developing multiple sclerosis treatments. In that case, the exact software license that will be used should be discussed with LRD and will be updated later.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Exploitation restrictions depend on the dataset:

- UZ Leuven: No restrictions.
- MSBase: We can only work on the work packages and tasks as described in the data use agreement (project 2020-015). This agreement further stipulates the authorship rules for publications resulting from working on this data.
- MultipleMS: To be updated later.

Dissemination of the patient data that is being reused in this work is not possible, not for any of the three datasets. The patient data remains the property of the medical centres in each case.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Documentation will include information on how to pre-process the hospital data (inclusion/exclusion criteria for patients, which variables to select, how to deal with missing data, etc.), as well as the exact validation procedure (train/test split, cross validation, evaluation measures) and parameter optimization procedure. Although a detailed methodology section will be required in any publication about the project, we will have a more elaborate description of the exact protocol in an extra document, this may be a readme.txt file or a lab notebook. For the generated computer programs, detailed source code documentation and a manual will be added.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

The metadata of the generated software includes programming language, author, version, date of creation,... Such metadata is automatically created by a git repository like KU Leuven GitLab.

3. Data storage & back-up during the research project

Where will the data be stored?

- UZLeuven: on the shared Windows network drive in Leuven.
- MSBase: on my hard drive.
- MultipleMS: to be discussed later.
- Processed data: In the same locations as the raw data, see 3 bullet points above.
- Analysed data: on my hard drive.
- Code: on my hard drive and KU Leuven GitLab. The code may be made available upon publication of an article describing the developed method.

How will the data be backed up?

- UZLeuven: no backup. This is under the responsibility of the team in Leuven. I am legally not allowed to make copies of the data.
- MSBase: no backup. In case of corruption/loss, a new copy can be requested. This is the safest and most convenient option.
- MultipleMS: to be discussed later.
- Processed data: same procedures as for the raw data, see 3 bullet points above.
- Analysed data: continuously backed up on OneDrive for Business (as this data is no longer as sensitive as the patient data described above).
- Code: continuously backed up on OneDrive for Business. KU Leuven GitLab naturally acts as a double backup. This triple versioning is appropriate because the code is the most important data in this project, as all processed / analysed data can easily be reconstructed using the code and the raw data.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Storage: my hard drive has a capacity of 500 GB, which is more than sufficient to store all the data.

Backup: our OneDrive for Business license offers 2TB of storage, which is even greater than my hard drive capacity. Furthermore, KU Leuven GitLab offers sufficient storage for the double backup of the <100 MB of code generated in this project.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

- Servers in Leuven (for the UZLeuven data): responsibility of the team in Leuven.
- Hard drive of my laptop: has a bitlocker, laptop itself is password-protected.
- OneDrive for Businesss: secured by two-factor authentication.
- KU Leuven GitLab: secured by two-factor authentication.

The latter 2 storage media are supported by the KU Leuven infrastructure, and are ensured to stay within a datacenter in Europe.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

No substantial costs for storage are expected.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All generated data will be retained for minimum 10 years (KU Leuven RDM policy).

The reused patient data can legally not be preserved by us after the research project. Responsibles for the preservation of this data are as follows:

- My co-promotor professor An Goris for the UZ Leuven data.

- Professor Bart Van Wijmeersch for the MSBase data.
- Karolinska Institute for the MultipleMS data.

Where will these data be archived (stored and curated for the long-term)?

The generated software data will be archived on KU Leuven's K drive, which is a drive specifically for archiving. It may also be made available, e.g. on a GitHub repository.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The expected costs for preservation are negligible and will be covered by the budget of the PI.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

- Raw data reused in this project and processed data will not be made available for legal and ethical reasons. The processed data can easily be recovered from the raw data using the software code generated during this project.
- Analysed data will be included in a publication, possibly as supplemental materials.
- Code may be made available on a project-per-project basis, e.g. in a hosted Git repository. Access to these repositories will not be restricted, as they do not consist of sensitive personal data (although care must be taken to curate the analysed data, e.g. graphs and tables, to fullfill this condition).

If access is restricted, please specify who will be able to access the data and under what conditions.

Not applicable.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects
- Yes, Ethical aspects

We are not allowed to (re)share the existing data that is being reused here and the processed data derived from it.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

In an Open Access repository.

When will the data be made available?

Upon publication of the research results.

Which data usage licenses are you going to provide? If none, please explain why.

The exact license of the software has to be decided, and will be discussed with LRD. For internal use in the research group, the software will be stored on the network drives.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

A DOI can be generated for any GitHub repository.

What are the expected costs for data sharing? How will these costs be covered?

No costs for data sharing are expected, creation of a public GitHub repository is free.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Myself (as the PhD student).

Who will manage data storage and backup during the research project?

Myself (as the PhD student).

Who will manage data preservation and sharing?

PI.

Who will update and implement this DMP?

PI and myself (as the PhD student).