Database Open Access

Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis

Hamza Khan , Lotte Geys , peer baneke , Giancarlo Comi , Liesbet Peeters

Published: Jan. 2, 2024. Version: 1.0.1

When using this resource, please cite: (show more options)
Khan, H., Geys, L., baneke, p., Comi, G., & Peeters, L. (2024). Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis (version 1.0.1). *PhysioNet*. RRID:SCR_007345. https://doi.org/10.13026/77ta-1866

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

Abstract

Multiple Sclerosis (MS) is an inflammatory autoimmune disease of the central nervous system, causing increased vulnerability to infections and disability among young adults. Ever since the coronavirus disease 2019 (COVID-19) outbreak, caused by severe acute respiratory syndrome coronavirus 2 infections, there have been concerns among people with MS (PwMS) about the potential interactions between various disease-modifying therapies and COVID-19. The COVID-19 in MS Global Data Sharing Initiative (GDSI) was initiated in 2020 to address these concerns. This paper focuses on the anonymisation and open-sourcing of a GDSI sub-dataset, comprising data entered by people with MS and clinicians using a fast data entry tool. The dataset includes demographics, comorbidities, hospital stay, and COVID-19 symptoms of PwMS. The dataset can be used to perform different statistical analyses to improve our understanding of COVID-19 in MS. Furthermore, this dataset can also be used within the context of educational activities to educate different stakeholders on the complex data science topics that were used within the GDSI.

Background

Multiple Sclerosis (MS) is a chronic neuroinflammatory autoimmune disease that affects the central nervous system. It results in varying degrees of functional loss due to demyelination and axonal damage [1]. People with MS are more prone to infections due to the combination of pathophysiology, treatment, and natural history of MS [2]. The COVID-19 and MS Global Data Sharing Initiative (GDSI) was established to investigate the effect of immunosuppressants or immune-modifying medications on COVID-19 or its outcomes in people with MS. The initiative aimed to scale up COVID-19 data collection efforts and provide the MS community with data-driven insights during the pandemic [3]. The GDSI identified variables related to COVID-19, severity, treatment, demographic information, MS history and severity, information on DMT use, comorbidities, and certain lifestyle behaviours, particularly smoking, as important to be included in the core dataset. The global MS community was engaged in sharing the documentation of the COVID-19 status of PwMS via a central platform provided by QMENTA [4].

Methods

This dataset was collected via a fast data entry tool that allowed clinicians, people with Multiple Sclerosis (PwMS), or their representatives to enter data directly into the central platform of the COVID-19 and MS Global Data Sharing Initiative. The tool provided a questionnaire based on predetermined variables and did not collect directly identifiable personal data to protect privacy. The tool has since been taken down as of February 3rd, 2022.

The dataset consists of a total of 1141 people with Multiple Sclerosis (PwMS), and to ensure the data's compliance with HIPAA guidelines, a thorough de-identification process was performed. After the data acquisition, a small cell risk assessment (SCRA) was carried out to categorize the variables into three categories: direct identifiers, sensitive variables, and indirect identifiers. Direct identifiers are unique variables that can identify a person without additional information, while sensitive variables are those that a respondent may not want to disclose. Indirect identifiers are non-unique variables that can re-identify an individual when combined with other indirect identifiers from another dataset.

Since no patient names were collected, the focus was on de-identifying dates and patient age. The dates in the 'stop_or_end_date_combined' column were shifted by a random number of days between -15 and 15 for each row to prevent re-identification through date information. Ages were categorized into four groups: 0 for ages between 0 and <18, 1 for ages between 18 and ≤50, 2 for ages between 51 and ≤70, and 3 for ages 71 or greater. By doing this, it was ensured that no specific ages above 90 were disclosed. After categorizing the variables and applying the necessary precautions, the dataset is de-identified and compliant with HIPAA guidelines while maintaining its utility for research purposes. Moreover, to ensure privacy preservation, techniques such as K-anonymity and diversity have been applied to the dataset.

The dataset includes a range of predetermined variables (n=47), such as sex, age category, MS type, EDSS score, smoking status, and BMI category. These variables provide insight into the patient's demographics, clinical characteristics, and COVID-19-related symptoms. A detailed description of the types of variables and their statistics has been provided in the Data Description section.

Data Description

The file "GDSI_OpenDataset_Final.csv" consists of data collected from 1141 people with Multiple Sclerosis (PwMS) as part of the COVID-19 and MS Global Data Sharing Initiative. The dataset includes various categorical and numerical variables, such as sex, MS type, smoking status, age category, EDSS score, COVID-19 symptoms and BMI category. The data were collected to better understand the impact of COVID-19 on PwMS and support research and analysis related to the disease. The dataset has been further anonymized using K-anonymity and ℓ-diversity to protect the privacy of the patients and is available for use by the research community.

Following is the detailed description of the variables included in the dataset:

- **secret_name:** A unique identifier for the patient. "P_" or "C_" in the beginning indicates patient-reported and clinician-reported outcomes, respectively.
- **report_source:** Indicates the source from which the data is acquired. The variable has two unique values: "clinicians" for clinician-reported and "patients" for patient-reported. Possible values: “Patients” (92.63%), “Clinicians” (7.36%).
- **age_in_cat:** Indicates age in categories. Missing values for this field: 0.00%.
  - 0: if the age range is between 0 and <18.
  - 1: if the age range is between 18 and ≤50.
  - 2: if the age range is between 51 and ≤70.
  - 3: if the age range is 71 or greater.
- **bmi_in_cat2:** This variable represents the body mass index (BMI) of the patient. BMI is a statistical index that can estimate body fat in people of any age by dividing a person's weight in kilograms by the square of height in metres. The unique values in bmi_in_cat2 are "not_overweight" (75.02%) and "overweight" (2.97%). The possible missing values are 21.99%.
  - “not_overweight”: if BMI ≤ 30 kg/m^2.
  - “overweight”: if BMI > 30 kg/m^2.
- **covid19_admission_hospital:** Indicates the hospital admission status of the patient as a result of COVID-19. Has two unique values: “Yes” (1.31%) indicates admission in the hospital, and “No” (98.78%) indicates no admission. 0.00% missing values.
- **covid19_confirmed_case:** Confirmed COVID-19 diagnosis of the patient. Has two unique values: “Yes” (5.25%) if the diagnosis is positive and “No” (94.74%) if it's otherwise. 0.00% missing values.
- **covid19_diagnosis:** It shows the perceived COVID-19 diagnosis of the patient. It has three unique values: “not_suspected” (75.19%), “suspected” (19.54%) and “confirmed” (5.25%). 0.00% missing values.
- **covid19_has_symptoms:** Indicates the presence or absence of COVID-19 symptoms. Has two unique values: “Yes” (31.77%) if there are one or more symptoms associated with COVID-19, such as fever, dry cough, fatigue, pain, sore throat, shortness of breath, nasal congestion, loss of taste or smell and pneumonia. “No” (68.22%) if there are no symptoms. 0.71% possible missing values.
- **covid19_icu_stay:** Indicates whether the patient stayed in the intensive care unit (ICU) of the hospital as a result of COVID-19. “Yes” (0.35%) if the person stayed in the hospital intensive care unit (ICU) and “No” (99.38%) if the person did not stay in the ICU of the hospital. 0.26% possible missing values.
- **covid19_self_isolation:** Self-isolation status of the patient, advised either by the clinician or self-reported. “Yes” (39.87%) if the person self-isolated and “No” (58.8%) if the person did not self-isolate. Possible missing values are 1.31%.
- **covid19_sympt_chills:** Presence of chills as a COVID-19 symptom. “Yes” (8.85%) indicates the presence of chills as a symptom, and “No” (12.88%) indicates otherwise. 78.26% possible missing values.
- **covid19_sympt_dry_cough:** Presence of dry cough as a COVID-19 symptom. “Yes” (17.17%) if there are symptoms of dry cough and “No” (7.44%) if there are no symptoms of it. 75.37% possible missing values.
- **covid19_sympt_fatigue:** Presence of fatigue as a COVID-19 symptom. “Yes” (20.94%) if there are symptoms of fatigue and “No” (4.55%) if there are no symptoms of it. 74.49% possible missing values.
- **covid19_sympt_fever:** Presence of fever as a COVID-19 symptom. “Yes” (12.35%) if there are symptoms of fever and “No” (12.62%) if there are no symptoms of it. 75.02% possible missing values.

- **covid19_sympt_loss_smell_taste:** Presence of loss of smell and taste as a COVID-19 symptom. “Yes” (7.17%) if there are symptoms of loss of smell and taste and “No” (13.75%) if there are no symptoms of loss of smell and taste. 75.37% possible missing values.
- **covid19_sympt_nasal_congestion:** Presence of nasal congestion as a COVID-19 symptom. “Yes” (13.67%) if there are symptoms of nasal congestion and “No” (9.46%) if there are no symptoms of it. 76.86% possible missing values.
- **covid19_sympt_pain:** Presence of pain as a COVID-19 symptom. “Yes” (16.3%) if there are symptoms of pain and “No” (7.88%) if there are no symptoms of it. 75.81% possible missing values.
- **covid19_sympt_pneumonia:** Presence of pneumonia as a COVID-19 symptom. “Yes” (1.22%) if there are symptoms of pneumonia and “No” (18.75%) if there are no symptoms of it. 80.01% possible missing values.
- **covid19_sympt_shortness_breath:** Presence of shortness of breath as a COVID-19 symptom. “Yes” (9.02%) if there are symptoms of shortness of breath and “No” (13.40%) if there are no symptoms of it. 77.56% possible missing values.
- **covid19_sympt_sore_throat:** Presence of sore throat as a COVID-19 symptom. “Yes” (14.89%) if there are symptoms of sore throat and “No” (9.02%) if there are no symptoms of it. 76.07% possible missing values.
- **covid19_ventilation:** Indicates whether the patient used a ventilator unit for ventilation during their hospital stay. “Yes” (0.35%) if ventilation was used, and “No” (99.21%) indicates otherwise. 0.43% possible missing values.
- **current_dmt:** Indicates the status of the disease-modifying therapy (DMT) at the time of data entry of the patient. There are three unique values in the variable: “yes” (77.12%), “no” (8.85%) and “never_treated” (14.02%). 0.00% possible missing values. The values are measured as follows:
  - yes: if the person is currently on a DMT
  - no: if the person is not currently on a DMT
  - never_treated: if the person has never been on a DMT
- **dmt_glucocorticoid:** Describes the status of intake of glucocorticoid. “Yes” (4.2%) if the person is taking glucocorticoid, and “no” (89.65%) states otherwise. 6.13% possible missing values.
- **edss_in_cat2:** Indicates the category in which the Expanded Disability Status Scale (EDSS) lies. The EDSS is one of the most commonly used disability assessment tools [23]. The EDSS is an ordinal scale from 0 to 10 that runs in increments of 0.5, with 0 indicating normal neurological status and 10 indicating death due to MS. Even though it has a few shortcomings, such as low reliability and sensitivity to change, it is still one of the preferred scales [23,24]. The variables had two unique values: “1” (0.00%) and “0” (45.8%). 54.16% possible missing values. The values were calculated as follows:
  - 0: if the EDSS is between 0 and < = 6.
  - 1: if the EDSS is > 6.

- **pregnancy:** Pregnancy status of the patient. “Yes” (0.43%) if the person is pregnant and “No” (74.75%) if the person is not. Possible missing values are 24.80%.
- **sex:** The biological sex of the patient. “Male” (20.77%) for males and “Female” (79.22%) for females. Possible missing values are 0.00%.
- **ms_type2:** The type of Multiple Sclerosis phenotype. This variable has three unique values: “relapsing_remitting” (79.4%), “progressive_MS” (9.11%) and “other” (11.48%). Possible missing values are 0.00%. The values were calculated as follows:
  - relapsing_remitting: if the type of MS is relapsing-remitting MS (RRMS)
  - progressive_MS: if the type of MS is secondary progressive MS (SPMS) or primary progressive MS (PPMS)
  - other: if the type of MS is a clinically isolated syndrome (CIS) or empty or “not_sure” in case the patient or clinician was not sure.
- **current_or_former_smoker:** Indicates the smoking status of the patient. "Yes" (43.38%) indicates whether a patient is a smoker and/or has been a former smoker. "No" (0.00%) indicates otherwise. Possible missing values are 0.00%.
- **dmt_type_overall:** Indicates the specific type of DMT the person was on during data entry. The unique values in the variable are: “Currently on another drug not listed” (16.21%), “Currently on dimethyl fumarate” (12.7%), “Currently on fingolimod” (10.16%), “Currently not using any DMT” (8.85%), “Currently on interferon” (8.23%), “Currently on ocrelizumab” (7.36%), “Currently on natalizumab” (5.69%), “Currently on glatiramer” (5.6%), “Currently on teriflunomide” (5.52%), “Currently on cladribine” (3.06%), “Currently on rituximab” (1.31%), “Currently on alemtuzumab” (1.13%). Possible missing values are 14.11%. The values were calculated as follows:
  - “No information on DMT use”: if there is no information in the present or past history of DMT use by the person.
  - “currently not using any DMT”: if the person is currently not using any DMT but has used DMT in the past or has not used DMT at all.
  - “currently on interferon”: if the current DMT of the person is on interferon.
  - “currently on glatiramer”: if the current DMT of the person is on glatiramer.
  - “currently on natalizumab”: if the current DMT of the person is on natalizumab.
  - “currently on fingolimod”: if the current DMT of the person is on fingolimod.
  - “currently on dimethyl fumarate”: if the current DMT of a person is on fumarate.
  - “currently on teriflunomide”: if the current DMT of the person is on teriflunomide.
  - “currently on alemtuzumab”: if the current DMT of the person is on alemtuzumab.
  - “currently on cladribine”: if the current DMT of the person is on cladribine.
  - “currently on siponimod”: if the current DMT of a person is on siponimod.
  - “currently on rituximab”: if the current DMT of the person is on rituximab.
  - “currently on ocrelizumab”: if the current DMT of the person is on ocrelizumab.

    - “currently on another drug not listed”: if the person is on DMT other than the above-listed ones.
- **duration_treatment_cat:** The duration of treatment of MS. The variable has two unique values: “0” (3.59%) and “1” (3.59%). Possible missing values are 80.54%. The unique values were calculated as follows:
    - 0: if the duration of treatment is less than 11 years.
    - 1: if the duration of treatment is 11 years or more.
- **stop_or_end_date_combined:** Date in dd/mm/yyyy format indicating the stopping of DMT. Possible missing values are 28.13%.
- **covid19_outcome_levels_2:** The outcome of COVID-19. The variable has three unique values: “0” (98.68%), “1” (0.87%), and “2” (0.43%). Possible missing values are 0.00%. The unique values were calculated as follows:
    - 0: If the person has COVID-19 but has not been hospitalised.
    - 1: The person has COVID-19 and has been hospitalised.
    - 2: The person has COVID-19, has been hospitalised, has been in the intensive care unit and/or was in a ventilation facility.
    - 3: The person died due to COVID-19 (not present in this dataset).
- **has_comorbidities:** Indicates whether the person has any comorbidities. “Yes” (28.74%) indicates that there are comorbidities, and “No” (71.25%) indicates otherwise. Possible missing values are 0.00%.
- **com_cardiovascular_disease:** Indicates the presence of cardiovascular comorbidities. “Yes” (1.13%) indicates that there are comorbidities, and “No” (22.08%) indicates otherwise. Possible missing values are 76.77%.
- **com_chronic_kidney_disease:** Indicates whether the person has any chronic kidney disease. “Yes” (0.35%) indicates that there are chronic kidney comorbidities, and “No” (22.61%) indicates otherwise. Possible missing values are 77.03%.
- **com_chronic_liver_disease:** Indicates whether the person has any chronic liver disease. “Yes” (0.87%) indicates that there are chronic liver comorbidities, and “No” (22.43%) indicates otherwise. Possible missing values are 76.68%.
- **com_diabetes:** Indicates whether the person has any diabetes. “Yes” (1.48%) indicates that there is diabetes, and “No” (21.99%) indicates otherwise. Possible missing values are 76.51%.
- **com_hypertension:** Indicates whether the person has hypertension. “Yes” (4.64%) indicates that there is hypertension, and “No” (19.63%) indicates otherwise. Possible missing values are 75.72%.
- **com_immunodeficiency:** Indicates whether the person has an immunodeficiency. “Yes” (2.54%) indicates that there is immunodeficiency, and “No” (20.42%) indicates otherwise. Possible missing values are 77.03%.
- **com_lung_disease:** Indicates whether the person has any lung disease. “Yes” (2.80%) indicates that there is a lung disease, and “No” (20.50%) indicates otherwise. Possible missing values are 76.68%.
- **com_malignancy:** Indicates whether the person has any malignancy. “Yes” (1.05%) indicates that there is malignancy, and “No” (21.91%) indicates otherwise. Possible missing values are 77.03%.
- **com_neurological_neuromuscular:** Indicates whether the person has any neurological and/or neuromuscular comorbidity. “Yes” (2.19%) indicates that there is neurological and/or neuromuscular comorbidity, and “No” (21.12%) indicates otherwise. Possible missing values are 76.68%.
- **comorbidities_other:** Indicates names of other comorbidities that the patient might have that are not mentioned in the column names. Possible missing values are 79.66%

Usage Notes

The dataset acquisition pipeline was developed using Python programming language, and the data is provided in a CSV format, which makes it compatible with a variety of data analysis tools and software packages. The pipeline utilized popular Python libraries, including matplotlib 3.6.0, pandas 1.5.3, NumPy 1.24, and SciPy 1.0 for data aggregation, statistical analysis (e.g., bias checks using chi-squared test), and visualization. Jupyter Notebook 5.0 served as the interface for the pipeline.

To assist the user community in both the collection and analysis of the data, the code and tools developed for this dataset are available through GitHub [5]. Users can access the GitHub repository, which can help them to reproduce the analyses, adapt the code for their specific needs, and collaborate with other researchers.

Ethics

This study followed the ethical guidelines and received approval from the ethics committee of Hasselt University (The Ethical committee UHasselt, CME2020/025 AMD3).

Acknowledgements

We would like to thank all people with MS and clinicians for their time invested in providing the data within the context of the GDSI. We are grateful for the continued support of the sponsors of the Multiple Sclerosis Data Alliance and Multiple Sclerosis International Federation. Finally, we thank QMENTA for the use of the central data platform.

Conflicts of Interest

MSIF receives income from a range of corporate sponsors, recently including Biogen, BristolMyersSquibb, Coloplast, Janssen, Merck, Mylan, Novartis, Roche and Sanofi – all of which is publicly disclosed. In addition, the authors involved in this research have no conflicts of interest to report that are relevant to this study.

References

1. Calabresi PA. Diagnosis and management of multiple sclerosis. Am Fam Physician. 2004 Nov 15;70(10):1935–44.
2. Montgomery S, Hillert J, Bahmanyar S. Hospital admission due to infections in multiple sclerosis patients. Eur J Neurol. 2013 Aug;20(8):1153–60.
3. Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: A global data sharing initiative. Mult Scler Houndmills Basingstoke Engl. 2020 Sep;26(10):1157–62.
4. Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis. Neurology. 2021 Nov 9;97(19):e1870–85.
5. https://github.com/hky365/Global-Data-Sharing-Initiative- [Accessed May 10 2023]

Contents

Share

Access

Access Policy:
Anyone can access the files, as long as they conform to the terms of the specified license.

License (for files):

Discovery

DOI (version 1.0.1):
https://doi.org/10.13026/77ta-1866

DOI (latest version):
https://doi.org/10.13026/3cmh-xa68

Corresponding Author

Liesbet Peeters
Biomedical Research Institute, Data Science Institute, University MS Center, UHasselt.
liesbet.peeters@uhasselt.be

Versions

1.0.0 - May 30, 2023

1.0.1 - Jan. 2, 2024

Files

Total uncompressed size: 237.4 KB.

Access the files

- Download the ZIP file (23.4 KB)
- Download the files using your terminal:
  `wget -r -N -c -np https://physionet.org/files/patient-level-data-covid-ms/1.0.1/`
- Download the files using AWS command line tools:
  `aws s3 sync --no-sign-request s3://physionet-open/patient-level-data-covid-ms/1.0.1/ DESTINATION`

Folder Navigation: <base>

Name	Size	Modified
GDSI_OpenDataset_Final.csv	221.2 KB	2023-05-09
LICENSE.txt	14.5 KB	2023-12-28
README.txt	1.5 KB	2023-12-28
SHA256SUMS.txt	245 B	2024-01-02

Maintained by the MIT Laboratory for Computational Physiology

Supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Heart Lung and Blood Institute (NHLBI), and NIH Office of the Director under NIH grant numbers U24EB037545 and R01EB030362