Data Cleaning Report

The Effect of COVID-19 on Patients with Multiple Sclerosis Datathon September2023

Team Name: Analytics Ninjas Team Lead: Neenu Nair

**Team Members**: Archana Mukthavaram, Jothi Arumugam, Sreesudha Gadamsetty

# Introduction

The effect of COVID-19 on patients with **Multiple Sclerosis** (MS) has been a concern since the beginning of the COVID-19 pandemic. MS is an autoimmune disease that affects the central nervous system, and individuals with MS often face unique challenges when it comes to infections and immune-related conditions. This article explores the impact of COVID-19 on MS patients, touching upon disease severity, the influence of MS treatments (DMTs), and other comorbidities. It emphasizes the importance of informed decision-making and holistic care in managing these dual health challenges.

# Data Definition

Understanding the data is important before analyzing raw data to make conclusions.

As a part of better data definition, we segregated the entire COVID dataset into 4 groups. The 4 different sub-datasets are given below:

1. **Demographics**: Demographics-related details of patients.
2. **Treatment\_Analysis**: Data related to the Patient's treatment details
3. **Other\_Diseases:** Data related to other comorbidities of patients
4. **Covid\_Symptoms**: Data related to the COVID Symptoms of Patients

The table below shows the data definition of the Demographics group. Similarly, we defined all other groups. Please refer to the “Data Definition” sheet in this link [Analytics\_Ninjas\_COVID\_CleaningReport](https://docs.google.com/spreadsheets/d/1SqgMeyeMEf7vCSImNtowfmZU2kzYNYqJ80-Co7gFkBo/edit#gid=1386834576)for the entire data definition.

| **Column Name** | **Data Type** | **Nullable** | **Value Range** | **Data Definition** |
| --- | --- | --- | --- | --- |
| **Demographics** | | | | |
| secret\_name | Text | No | C/P\_(1-1500) | A unique identifier for the patient. "P\_" or "C\_" in the beginning indicates patient-reported and clinician-reported outcomes, respectively. |
| report\_source | Text | No | two unique values: "clinicians" for clinician-reported and "patients" for patient-reported. | Indicates the source from which the data is acquired. |
| age\_in\_cat | Whole Number | No | 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. | Indicates age in 4 different categories |
| bmi\_in\_cat2 | Text | Yes | “not\_overweight”: if BMI ≤ 30 kg/m2.  “overweight”: if BMI > 30 kg/m2. | This variable represents the body mass index (BMI) of the patient. |
| pregnancy | Text | Yes | yes, no | Pregnancy status of the patient. “Yes” if the person is pregnant and “No” if the person is not. |
| sex | Text | No | female, male | The biological sex of the patient. |
| current\_or\_former\_smoker | Text | Yes | yes, no | Indicates the smoking status of the patient. "Yes" means a patient is a smoker and/or has been a former smoker. "No" indicates otherwise. |
| has\_comorbidities | Text | No | yes, no | Indicates whether the person has any comorbidities. “Yes” indicates that there are comorbidities, and “No” indicates otherwise. |

# Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

## **3.1 Data Cleaning Analysis**

## Data Cleaning Tool Used **- Power Query Editor in PowerBI**

Dataset Used - GDSI\_OpenDataset\_Final

Size of original dataset -222KB (47 Columns,1141 rows)

Size of Cleaned dataset - 205KB (50 columns,1141 rows)

* Deleted 3 columns. Has\_comorbidities.1, has\_comorbidities.2(duplicate of has\_comorbidities) covid19\_confirmed\_case(Duplicate of column 'covid19\_diagnosis).
* Renamed 44 columns.
* Added 6 extra columns. W1\_RF, W2\_RF, W3\_RF, W4\_RF, COVID\_Score, COVID\_Status. The reason for adding the columns is that we want to show the COVID severity based on risk factors like age, BMI, COVID symptoms, and other diseases. We have given weightage to Risk factors as Low Risk, Moderate Risk, High Risk, and Critical risk. We have calculated the COVID\_score from the risk factors and given the COVID\_status as Mild, Moderate, Severe, and Critical.
* Total number of columns after data cleaning: 50
* Final Cleaned Dataset: Please refer to the “COVID\_Final\_Cleaned\_Dataset” sheet in this Link[Analytics\_Ninjas\_COVID\_CleaningReport](https://docs.google.com/spreadsheets/d/1SqgMeyeMEf7vCSImNtowfmZU2kzYNYqJ80-Co7gFkBo/edit#gid=958895000)

## **3.2 Data Modelling**

## Data modeling is the process of creating a simplified diagram of data elements it contains, to represent the data and how it flows. Using table referencing we created 4 tables named Demographics, Treatment\_Analysis, Other\_Diseases, and Covid\_Symptoms. Used “PatientID” as the Primary key for all 4 tables with a one-to-one relationship from demographics to all other Tables.

## 

**3.3 Data Cleaning Approaches And Challenges**

**Data Cleaning Approaches**:

* Removed duplicates (3 columns)
* Renamed columns
  + - Secret\_name to ‘PatientID’
    - Sex to ‘Gender’
    - covid19\_sympt\_xxx to Xxx
    - Covid19\_xxx to COVID-19\_Xxx
    - Com\_xxx to Xxx
* Format / Change data type
  + - Capitalized the first letter of each word with ‘\_’ as the delimiter between the words
    - Changed the data type of the columns with only ‘yes’ and ‘no’ to “Whole Number” by replacing the values to ‘1’ and ‘0’
    - Changed the data type of ‘year\_onset’ to Date->Year
* Missing data
  + - Left blank values unmodified as updating these values from relevant fields will affect the outcome of the analysis.

Please refer to the “Data Cleaning” sheet in this link [Analytics\_Ninjas\_COVID\_CleaningReport](https://docs.google.com/spreadsheets/d/1SqgMeyeMEf7vCSImNtowfmZU2kzYNYqJ80-Co7gFkBo/edit#gid=1386834576)for the detailed data cleaning.

**Challenges:**

* The ‘DMT\_End\_Date’ column has date values and is in the format of “DD/MM/YYYY”. So if we try to change the data type to ‘Date’, some values are throwing errors, and fixing them will lead to loss of data.
* After creating the Data Model, we were not able to access the fields of one table from another. Later resolved it by referring to them using the DAX function ‘RELATED’.

# 4. Conclusion

The raw COVID dataset has been examined in a proper way with detailed data definition. The data cleaning and Data Modelling process has been carried out successfully within a specified time frame with the help of Power BI- Power query editor. The entire cleaning procedure has been documented and the final cleaned dataset has been extracted from Power BI and stored in an Excel file for evaluation.

# 5. References

1. [Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis](https://physionet.org/content/patient-level-data-covid-ms/1.0.0/)
2. [Types of Multiple Sclerosis](https://www.nationalmssociety.org/What-is-MS/Types-of-MS#:~:text=Four%20disease%20courses%20have%20been,secondary%20progressive%20MS%20(SPMS))
3. [Expanded Disability Status Scale (EDSS) | MS Trust](https://mstrust.org.uk/a-z/expanded-disability-status-scale-edss)
4. [Disease Modifying Therapies For MS - DMTs](https://www.mssociety.org.uk/about-ms/treatments-and-therapies/disease-modifying-therapies#:~:text=Disease%20Modifying%20Therapies%20(DMTs)%20are,treatment%20that%20uses%20stem%20cells)
5. [Underlying Medical Conditions Associated with Higher Risk for Severe COVID-19: Information for Healthcare Professionals | CDC](https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html)
6. [A Novel Scoring System for Prediction of Disease Severity in COVID-19](https://www.frontiersin.org/articles/10.3389/fcimb.2020.00318/full)
7. [Clinical Spectrum of SARS-CoV-2 Infection](https://www.covid19treatmentguidelines.nih.gov/overview/clinical-spectrum/)