Brainteaser

D9.6 Shared data package for the evaluation challenge and integration with EOSC (M42)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No GA101017598

Project Title	BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis
Grant Agreement No	GA101017598
Contract start date	01/01/2021
Contract duration	48 Months

Document ID	BRAINTEASER_D9.6 Shared data package for the evaluation challenge and integration with EOSC (M42)
Deliverable leader	UNIPD
Due date	30/06/2024
Deliverable date	25/09/2024
Dissemination level	PUBLIC

AUTHORS – CONTRIBUTORS

Name	Organization
Giovanni Birolo	University of Turin, Italy
Pietro Bosoni	University of Pavia, Italy
Guglielmo Faggioli	University of Padua, Italy
Helena Aidos	University of Lisbon, Portugal
Roberto Bergamaschi	University of Pavia, Italy
Paola Cavalla	“Città della Salute e della Scienza”, Italy
Adriano Chiò	University of Turin, Italy
Arianna Dagliati	University of Pavia, Italy
Mamede de Carvalho	University of Lisbon, Portugal
Giorgio Maria Di Nunzio	University of Padua, Italy
Piero Fariselli	University of Turin, Italy
Jose Manuel García Dominguez	Gregorio Marañon Hospital in Madrid, Spain
Marta Gromicho	University of Lisbon, Portugal
Alessandro Guazzo	University of Padua, Italy
Enrico Longato	University of Padua, Italy
Sara C. Madeira	University of Lisbon, Portugal
Umberto Manera	University of Turin, Italy
Stefano Marchesin	University of Padua, Italy
Laura Menotti	University of Padua, Italy
Gianmaria Silvello	University of Padua, Italy
Eleonora Tavazzi	IRCCS Foundation C. Mondino in Pavia, Italy
Erica Tavazzi	University of Padua, Italy
Isotta Trescato	University of Padua, Italy
Martina Vettoretti	University of Padua, Italy
Barbara Di Camillo	University of Padua, Italy
Nicola Ferro	University of Padua, Italy

PEER – REVIEWERS

Name	Organization
Borko Kostić	BELIT
Aleksandar Jovanović	BELIT

DOCUMENT HISTORY

Version	Date	Author/Organization	Modifications	Status
0.1	02/05/2024	UNIPD	Initial outline	Draft
0.2	01/06/2024	UNIPD	First draft	Draft
0.3	15/06/2024	UNIPD	Second draft	Draft
0.4	30/06/2024	UNIPD	Formatting	Draft
1.0	22/07/2024	UNIPD	Final draft for peer review	Draft
1.1	17/09/2024	UNIPD	Final draft	Final draft
2.0	25/09/2024	Maria F Cabrera / UPM	Final review and final version	Final

Disclaimer

This deliverable may be subject to final acceptance by the European Commission. The information and views set out in this document are those of the authors and do not necessarily reflect the official opinion of the European Commission. Neither the Commission nor any person acting on the Commission's behalf may hold responsible for the use which may be made of the information contained therein.

Copyright message

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Acronym	Meaning
ALS	Amyotrophic Lateral Sclerosis
ALSFRS	ALS Functional Rating Scale
ALSFRS-R	ALSFRS Revised
BTO	BrainTeaser Ontology
ESCO	European Skills, Competences, Qualifications and Occupations
FAIR	Findable, Accessible, Interoperable, Reusable
FVC	Forced Vital Capacity
MS	Multiple Sclerosis
NIV	Non Invasive Ventilation
PEG	Percutaneous Endoscopic Gastrostomy
RDF	Resource Description Framework

EXECUTIVE SUMMARY

The primary goal of this deliverable is to detail the steps we took to ingest, process, analyze, and prepare the training and testing datasets for the Intelligent Disease Progression Prediction (iDPP) lab at CLEF 2024.

The lab was organized into three tasks: two related to Amyotrophic Lateral Sclerosis (ALS) (Tasks 1 and 2), and one (Task 3) focused on evaluating the impact of environmental factors on the progression of Multiple Sclerosis (MS) and how to utilize environmental data for predictions. This year, unlike previous editions, the ALS data is prospective, collected specifically for the BRAINTEASER project, whereas previous years used retrospective data already available to clinical centers.

The first two datasets, used for Tasks 1 and 2, focus on ALS and include patients' static data as well as ALSFRS-R scores, both clinician-annotated and self-assessed, collected through the BRAINTEASER application. The third dataset, used for Task 3, builds on the MS dataset provided for iDPP@CLEF 2023, now augmented with environmental data.

We describe the data characteristics, the collection process, and some statistics of the final dataset. Data were ingested into the BRAINTEASER Ontology (BTO), so we also provide the SPARQL queries used to construct the MS datasets and extract environmental data from the BTO.

Finally, we outline the procedure followed to produce the datasets and divide them into training and test partitions for Tasks 1 and 2. Task 3 maintained the same partitions used in iDPP@CLEF 2023.

1 INTRODUCTION

ALS and MS are two severe and impactful diseases that cause progressive neurological impairment in those affected. The progression of both diseases is typically heterogeneous, leading to significant variability in treatment, outcomes, quality of life, and overall needs. This variability presents challenges for patients, clinicians, and caregivers alike. For instance, patients with ALS often require specific treatments such as NIV or PEG at certain stages, while those with MS may experience debilitating relapses that severely impact their quality of life.

Anticipating the needs of individuals with these diseases would be highly beneficial, yet the heterogeneity of the diseases makes it difficult to develop effective prognostic tools. This underscores the importance of creating automated tools to assist clinicians throughout all phases of disease progression and support personalized therapeutic decisions.

When developing new predictive approaches based on AI, researchers need a robust framework to design and evaluate various tasks, such as:

- Stratifying patients by phenotype throughout disease progression,
- Predicting disease progression in a probabilistic, time-dependent manner,
- Providing an explainable understanding of the mechanisms underlying MS and ALS.

It is crucial that these approaches are built on shared resources to ensure proper benchmarking, and comparable and reproducible experimentation. The iDPP Lab aims to provide an evaluation infrastructure for developing AI algorithms tailored to these needs. Unlike previous efforts, iDPP systematically addresses issues related to the application of AI in clinical practice for ALS and MS. Beyond defining risk scores based on the likelihood of short and long-term events, iDPP also offers clinicians structured and comprehensible data.

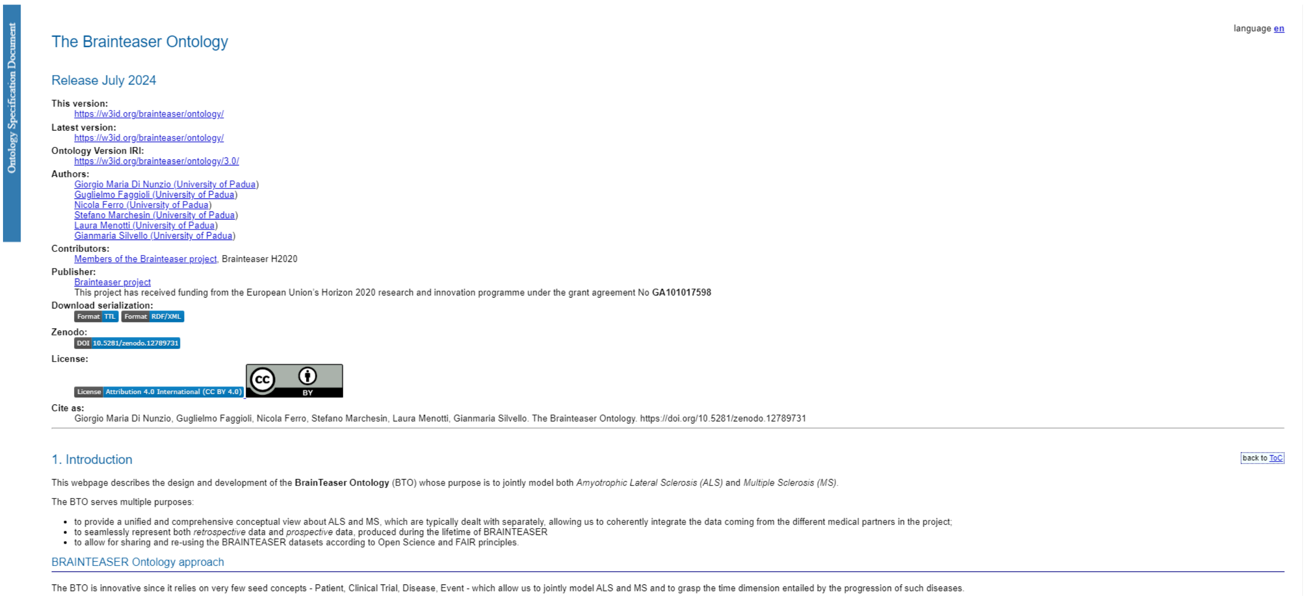
During this work, we describe the steps we took to ingest, process, analyze, and prepare the training and testing datasets for the Intelligent Disease Progression Prediction (iDPP) lab at CLEF 2024, as well as the data itself.

iDPP 2024 is the final iteration in a series of three evaluation challenges aimed at promoting reproducible and comparable AI-based approaches to predict the progression of ALS and MS. The first edition, iDPP 2022, focused on ALS, tasking participants with predicting the likelihood of patients requiring specific medical treatments based on their medical history. The second edition, iDPP 2023, expanded the dataset to include environmental data to assess its impact on patient needs and introduced a new task to predict the risk of MS worsening.

In this final release of the data, which further extends the 2023 dataset by incorporating environmental data for MS patients to evaluate the effects of pollution and external environments on disease progression. Additionally, two new tasks were introduced: predicting the progression of ALS using the ALSFRS-R scale based on clinical history and data from wearable devices and sensors.

2 SPARQL QUERIES

This section contains the queries used to construct the overall dataset that later is further processed to extract the datasets for each task.



Ontology Specification Document

language en

The Brainteaser Ontology

Release July 2024

This version:
https://w3id.org/brainteaser/ontology/

Latest version:
https://w3id.org/brainteaser/ontology/

Ontology Version IRI:
https://w3id.org/brainteaser/ontology/3.0/

Authors:
Giorgio Maria Di Nunzio (University of Padua)
Guglielmo Faggioli (University of Padua)
Nicola Ferro (University of Padua)
Stefano Marchesin (University of Padua)
Laura Menotti (University of Padua)
Gianmaria Silvello (University of Padua)

Contributors:
Members of the Brainteaser project, Brainteaser H2020

Publisher:
Brainteaser project
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No **GA101017598**

Download serialization:
Format TTL Format RDF/XML

Zenodo:
DOI 10.5281/zenodo.12789731

License:
License Attribution 4.0 International (CC BY 4.0)

Cite as:
Giorgio Maria Di Nunzio, Guglielmo Faggioli, Nicola Ferro, Stefano Marchesin, Laura Menotti, Gianmaria Silvello. The Brainteaser Ontology. https://doi.org/10.5281/zenodo.12789731

1. Introduction

back to ToC

This webpage describes the design and development of the **BrainTeaser Ontology** (BTO) whose purpose is to jointly model both *Amyotrophic Lateral Sclerosis (ALS)* and *Multiple Sclerosis (MS)*.

The BTO serves multiple purposes:

- to provide a unified and comprehensive conceptual view about ALS and MS, which are typically dealt with separately, allowing us to coherently integrate the data coming from the different medical partners in the project;
- to seamlessly represent both *retrospective* data and *prospective* data, produced during the lifetime of BRAINTEASER
- to allow for sharing and re-using the BRAINTEASER datasets according to Open Science and FAIR principles.

BRAINTEASER Ontology approach

The BTO is innovative since it relies on very few seed concepts - Patient, Clinical Trial, Disease, Event - which allow us to jointly model ALS and MS and to grasp the time dimension entailed by the progression of such diseases.

Figure 1. A screenshot of the new version of the Brainteaser Ontology documentation available at https://brainteaser.dei.unipd.it/ontology/

2.1 Prefixes

```
PREFIX bto: <https://w3id.org/brainteaser/ontology/schema/>
PREFIX BTO_resource: <https://w3id.org/brainteaser/ontology/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssn: <http://purl.oclc.org/NET/ssnx/ssn#>
PREFIX NCIT: <http://purl.obolibrary.org/obo/NCIT_>
```

2.2 MS Dataset Queries

2.2.1 Static Variables

The subsequent query is used to extract static variables such as personal data, i.e. sex, ethnicity, residence, onset information, and diagnostic delay. For each patient, we return the following list of variables:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **sex:** The sex of the patient. Possible values: {female, male}.
- **residence_classification:** Residence classification. When not absent, possible values are: {Cities, Rural Area, Towns}.
- **ethnicity:** The ethnicity of a patient. When not absent, possible values are: {Black African, Caucasian, Hispanic}.

- **ms_in_pediatric_age:** Boolean variable that states whether the patient contracted MS in pediatric age (True) or not (False).
- **age_at_onset:** The patient's age at onset. Expressed in years.
- **age_at_baseline:** The patient's age at baseline. Expressed in years.
- **diagnostic_delay:** The Difference (in days) between the date of diagnosis and Time 0. When diagnosis date > Time 0, diagnostic_delay is set to NaN.
- **diagnosis_criteria_detail:** Criteria used to diagnose MS, e.g. MS according to McDonald 2001.
- **spinal_cord_symptom:** Boolean variable that states whether the patient had spinal cord symptoms (True) or not (False) at the onset.
- **brainstem_symptom:** Boolean variable that states whether the patient had brainstem symptoms (True) or not (False) at the onset.
- **eye_symptom:** Boolean variable that states whether the patient had eye symptoms (True) or not (False) at the onset.
- **supratentorial_symptom:** Boolean variable that states whether the patient had supratentorial symptoms (True) or not (False) at the onset.
- **other_symptoms:** Specifies what other symptoms (if any) the patient had at the onset. When not False, possible values are: {epilepsy, headache, psychic, RM+, sensory}.
- **centre:** This variable represents the medical centre from which the patient comes from.
- **diagnosis_criteria:** Criteria for the diagnosis. Possible values are: {Diagnosed MS, Neuromyelitis optica, NA}

```
SELECT ?patient_id ?sex ?residence_classification ?ethnicity
?ms_in_pediatric_age ?age_at_onset ?age_at_baseline ?diagnostic_delay
?diagnosis_criteria_detail ?spinal_cord_symptom ?brainstem_symptom
?eye_symptom ?supratentorial_symptom ?other_symptoms ?centre
?diagnosis_criteria
WHERE {

  ?p a bto:Patient;
     bto:undergo ?d;
     bto:undergo ?o;
     bto:ageBaseline ?age_at_baseline.

  BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

  # diagnosis information
  ?d a bto:Diagnosis.
  OPTIONAL{ ?d bto:diagnosticDelay ?diagnostic_delay.}
  OPTIONAL{ ?d bto:diagnosticCriteria ?diagnosis_criteria_detail.}
  OPTIONAL{ ?d bto:diagnosticCriteriaLabel ?diagnosis_criteria.}

  # onset information
  ?o a bto:Onset;
     bto:ageOnset ?age_at_onset.
  # brainstem_symptom (NCIT:C12441)
  BIND(IF(EXISTS{ ?o bto:anatomicalLocation
       NCIT:C12441.},"True"^^xsd:boolean,"False"^^xsd:boolean) AS
       ?brainstem_symptom)
  # eye_symptom (NCIT:C12401)
  BIND(IF(EXISTS{ ?o bto:anatomicalLocation
       NCIT:C12401.},"True"^^xsd:boolean,"False"^^xsd:boolean) AS
       ?eye_symptom)

  # spinal_cord_symptom (NCIT:C12464)
  BIND(IF(EXISTS{ ?o bto:anatomicalLocation
       NCIT:C12464.},"True"^^xsd:boolean,"False"^^xsd:boolean) AS
       ?spinal_cord_symptom)

  # supratentorial_symptom (NCIT:C12512)
  BIND(IF(EXISTS{ ?o bto:anatomicalLocation
       NCIT:C12512.},"True"^^xsd:boolean,"False"^^xsd:boolean) AS
       ?supratentorial_symptom)
  # other onset symptoms
  OPTIONAL {?o bto:otherOnset ?other_symptoms.}

  # residence classification
  OPTIONAL{
      ?p bto:residence ?residence.
      ?residence bto:placeLocation ?pl.
      ?pl bto:placeArea ?residence_classification.
  }

  # ethnicity
  OPTIONAL{
      ?p bto:hasEthnicity ?e.
      ?e rdfs:label ?ethnicity.
  }

  # MS in paediatric age
  OPTIONAL{ ?p bto:MSInPaediatricAge ?ms_in_pediatric_age. }

  # sex
  BIND(IF(
     EXISTS{ ?p bto:sex
     "Female"^^<http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral>.
   },"female","male") AS ?sex)

  # centre
  BIND(IF(
     EXISTS{
        ?p bto:enrolledIn ?ctp.
        ?ctp bto:inClinicalTrial ?ct.
        ?ct bto:conductedInClinic <https://www.mondino.it/>.
     },"pavia","turin") AS ?centre)
}
```

2.2.2 Expanded Disability Status Scale (EDSS)

This query returns information about the Expanded Disability Status Scale (EDSS) visits of each patient. The EDSS is a method of quantifying disability in multiple sclerosis and monitoring changes in the level of disability over time. It is widely used in clinical trials and in the assessment of people with MS. For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **pyramidal,cerebellar,brainstem,sensory,bowel_and_bladder,visual_function,cerebral_functions,ambulation:** value of each EDSS variable
- **edss_as_evaluated_by_clinician:** global EDSS assed by the clinician
- **week_from_baseline:** Time between the baseline and the data collection expressed in weeks

```
SELECT    ?patient_id    ?pyramidal    ?cerebellar    ?brainstem    ?sensory
?bowel_and_bladder   ?visual_function   ?cerebral_functions   ?ambulation
?edss_as_evaluated_by_clinician ?week_from_baseline
WHERE {

  ?p a bto:Patient;
     bto:undergo ?e.

  BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

  ?e bto:consists ?edss.

  ?edss a bto:EDSS;
        bto:privacyPreservingDate ?week_from_baseline.

  OPTIONAL { ?edss bto:pyramidalEDSS ?pyramidal.}
  OPTIONAL { ?edss bto:cerebellarEDSS ?cerebellar.}
  OPTIONAL { ?edss bto:brainstemEDSS ?brainstem.}
  OPTIONAL { ?edss bto:sensoryEDSS ?sensory.}
  OPTIONAL { ?edss bto:bowelAndBladderEDSS ?bowel_and_bladder.}
  OPTIONAL { ?edss bto:visualEDSS ?visual_function.}
  OPTIONAL { ?edss bto:cerebralEDSS ?cerebral_functions.}
  OPTIONAL { ?edss bto:ambulationEDSS ?ambulation.}
  OPTIONAL { ?edss bto:clinicallyEvaluatedEDSS
             ?edss_as_evaluated_by_clinician.}
}
```

2.2.3 Environmental Measurements

The environmental data are integrated with MS patients' positional information to investigate a possible relationship between the disease and pollutants. The subsequent query returns environmental information related to each patient. For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **week_from_baseline:** Time between the baseline and the data collection expressed in weeks.

- **<PM25, PM10, CO, NO2, O3, SO2, wind_speed, humidity, sealevel_pressure, global_radiation, precipitation, average_temperature, min_temperature, max_temperature>_num:** For each recorded variable, number of valid measurements in that week, ranging from 0 to 7.
- **<PM25, PM10, CO, NO2, O3, SO2, wind_speed, humidity, sealevel_pressure, global_radiation, precipitation, average_temperature, min_temperature, max_temperature>_mean:** average of valid measurements in that week; if there are no valid measurements in that week, it is denoted as NA.
- **<PM25, PM10, CO, NO2, O3, SO2>_who:** number of days spent over the respective World Health Organization (WHO) threshold in that week (computed as following); if there are no valid measurements in that week, it is denoted as NA.

2.2.4 Definition of rules based on WHO Guidelines (only for air pollutants).

The rules are established based on WHO's recommended air quality guideline levels for short-term (24 hours) exposure to various air pollutants. Each rule compares the concentration of a pollutant to its respective WHO guideline level. If the concentration exceeds the guideline level, the rule assigns a value of 1; otherwise, it assigns a value of 0. Here are the rules for different pollutants levels for short-term (24 hours) exposure to air pollutants (reference: https://www.who.int/publications/i/item/9789240034228)

- For Carbon Monoxide (CO): if CO ≤ 4 mg/m3 then the value is 0; otherwise, it is 1;
- For Nitrogen Dioxide (NO2): if NO 2 ≤ 25 µg/m3 then the value is 0; otherwise, it is 1;
- For Ozone (O3): if O 3 ≤ 100 µg/m3 then the value is 0; otherwise, it is 1;
- For Particulate Matter with a diameter of 2.5 micrometers or smaller (PM 2.5): if PM 2.5 ≤ 15 µg/m 3 then the value is 0; otherwise, it is 1
- For Particulate Matter with a diameter of 10 micrometers or smaller (PM 10): if PM 10 ≤ 45 µg/m3 , then the value is 0; otherwise, it is 1;
- For Sulfur Dioxide (SO2); if SO 2 ≤ 40 µg/m3 , then the value is 0; otherwise, it is 1.

2.2.5 Unit of measure of variables.

- PM25, PM10, CO, NO2, O3, SO2: micrograms per cubic meter (µg/m3)
- Wind speed: m/s
- Relative humidity: %
- Sea level pressure: hPa
- Global radiation: W/m²
- Precipitation sum: mm
- Average, Minimum, Maximum temperature: °C

```
SELECT ?patient_id ?week_from_baseline ?PM25_num ?PM25_mean ?PM25_who
?PM10_num ?PM10_mean ?PM10_who ?CO_num ?CO_mean ?CO_who ?NO2_num ?NO2_mean
?NO2_who ?O3_num ?O3_mean ?O3_who ?SO2_num ?SO2_mean ?SO2_who
?wind_speed_num ?wind_speed_mean ?humidity_num ?humidity_mean
?sealevel_pressure_num ?sealevel_pressure_mean ?global_radiation_num
?global_radiation_mean ?precipitation_num ?precipitation_mean
?average_temperature_num ?average_temperature_mean ?min_temperature_num
?min_temperature_mean ?max_temperature_num ?max_temperature_mean
```

```
WHERE {

   ?p a bto:Patient;
      bto:residence ?residence.

   BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

   ?residence bto:placeLocation ?pl.
   ?pl a bto:Place.
   ?sd a ssn:SensingDevice;
       bto:coveredPlace ?pl.

   # PM2.5
   OPTIONAL{
       ?PM25 a pollution:PM2.5_concentration;
             ssn:isProducedBy ?sd;
             bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{?PM25 bto:validMeasurements ?PM25_num.}
       OPTIONAL{ ?PM25 bto:averageMeasurement ?PM25_mean.}
       OPTIONAL{ ?PM25 bto:whoThresholdExceeded ?PM25_who.}
   }
   # PM10
   OPTIONAL{
       ?PM10 a pollution:PM10_concentration;
             ssn:isProducedBy ?sd;
             bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{ ?PM10 bto:validMeasurements ?PM10_num.}
       OPTIONAL{ ?PM10 bto:averageMeasurement ?PM10_mean.}
       OPTIONAL{ ?PM10 bto:whoThresholdExceeded ?PM10_who.}
   }
   # Carbon Monoxide (CO)
   OPTIONAL{
       ?CO a bto:CO_concentration;
           ssn:isProducedBy ?sd;
           bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{ ?CO bto:validMeasurements ?CO_num.}
       OPTIONAL{ ?CO bto:averageMeasurement ?CO_mean.}
       OPTIONAL{ ?CO bto:whoThresholdExceeded ?CO_who.}
   }
   # Nitrogen Dioxide (NO2)
   OPTIONAL{
       ?NO2 a pollution:NO2_concentration;
            ssn:isProducedBy ?sd;
            bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{ ?NO2 bto:validMeasurements ?NO2_num.}
       OPTIONAL{ ?NO2 bto:averageMeasurement ?NO2_mean.}
       OPTIONAL{ ?NO2 bto:whoThresholdExceeded ?NO2_who.}
   }
   # Ozone (O3)
   OPTIONAL{
       ?O3 a pollution:O3_concentration;
           ssn:isProducedBy ?sd;
           bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{ ?O3 bto:validMeasurements ?O3_num.}
```

```
        OPTIONAL{ ?O3 bto:averageMeasurement ?O3_mean.}
        OPTIONAL{ ?O3 bto:whoThresholdExceeded ?O3_who.}
    }
    # Sulphur Dioxide (SO2)
    OPTIONAL{
        ?SO2 a pollution:SO2_concentration;
             ssn:isProducedBy ?sd;
             bto:privacyPreservingDate ?week_from_baseline.
        OPTIONAL{ ?SO2 bto:validMeasurements ?SO2_num.}
        OPTIONAL{ ?SO2 bto:averageMeasurement ?SO2_mean.}
        OPTIONAL{ ?SO2 bto:whoThresholdExceeded ?SO2_who.}
    }
    # Wind Speed
    OPTIONAL{
        ?wind_speed a bto:WindSpeed;
                    ssn:isProducedBy ?sd;
                    bto:privacyPreservingDate ?week_from_baseline.
        OPTIONAL{ ?wind_speed bto:validMeasurements ?wind_speed_num.}
        OPTIONAL{ ?wind_speed bto:averageMeasurement ?wind_speed_mean.}
    }
    # Humidity
    OPTIONAL{
        ?humidity a bto:Humidity;
                  ssn:isProducedBy ?sd;
                  bto:privacyPreservingDate ?week_from_baseline.
        OPTIONAL{ ?humidity bto:validMeasurements ?humidity_num.}
        OPTIONAL{ ?humidity bto:averageMeasurement ?humidity_mean.}
    }
    # Sea Level Pressure
    OPTIONAL{
        ?sealevel_pressure a bto:AtmosphericPressure;
                           ssn:isProducedBy ?sd;
                           bto:privacyPreservingDate ?week_from_baseline.
        OPTIONAL{          ?sealevel_pressure          bto:validMeasurements
?sealevel_pressure_num.}
        OPTIONAL{         ?sealevel_pressure         bto:averageMeasurement
?sealevel_pressure_mean.}
    }
    OPTIONAL{
        ?global_radiation a bto:GlobalRadiation;
                          ssn:isProducedBy ?sd;
                          bto:privacyPreservingDate ?week_from_baseline.
        OPTIONAL{          ?global_radiation          bto:validMeasurements
?global_radiation_num.}
        OPTIONAL{          ?global_radiation         bto:averageMeasurement
?global_radiation_mean.}
    }
    OPTIONAL{
        ?precipitation a bto:Precipitation;
                       ssn:isProducedBy ?sd;
                       bto:privacyPreservingDate ?week_from_baseline.
        OPTIONAL{ ?precipitation bto:validMeasurements ?precipitation_num.}
        OPTIONAL{            ?precipitation            bto:averageMeasurement
?precipitation_mean.}
```

```
   }
   OPTIONAL{
       ?average_temperature a bto:MeanTemp;
                            ssn:isProducedBy ?sd;
                            bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{        ?average_temperature        bto:validMeasurements
?average_temperature_num.}
       OPTIONAL{        ?average_temperature        bto:averageMeasurement
?average_temperature_mean.}
   }
   OPTIONAL{
       ?min_temperature a bto:MinTemp;
                        ssn:isProducedBy ?sd;
                        bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{          ?min_temperature          bto:validMeasurements
?min_temperature_num.}
       OPTIONAL{          ?min_temperature          bto:averageMeasurement
?min_temperature_mean.}
   }
   OPTIONAL{
       ?max_temperature a bto:MaxTemp;
                        ssn:isProducedBy ?sd;
                        bto:privacyPreservingDate ?week_from_baseline.
       OPTIONAL{          ?max_temperature          bto:validMeasurements
?max_temperature_num.}
       OPTIONAL{          ?max_temperature          bto:averageMeasurement
?max_temperature_mean.}
   }
}
```

2.2.6 Outcomes

This query returns information concerning relapses, e.g., the patient who had a relapse and the date in which such an event occurred with respect to the baseline. For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **week_from_baseline:** Time between the baseline and the relapse expressed in weeks.

```
SELECT ?patient_id ?week_from_baseline
WHERE {

  ?p a bto:Patient;
     bto:undergo ?r.

  BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

  ?r bto:relapse ?relapse.

  ?relapse a bto:Relapse;
           bto:privacyPreservingDate ?week_from_baseline.
}
```

2.3 ALS Dataset Queries

2.3.1 Static Variables

The subsequent query is used to extract static variables such as personal data, e.g., sex and diagnostic delay. For each patient, we return the following list of variables:

- patient_id: Hashed code that uniquely identifies a patient.
- sex: The sex of the patient. Possible values: {female, male}.
- diagnostic_delay: Delay between onset and diagnosis. Expressed in years.
- age_at_diagnosis: The patient's age at diagnosis. Expressed in years.
- FVC: FVC measured at diagnosis.
- weight: weight measured at diagnosis.
- bmi: BMI measured ad diagnosis.

```
SELECT ?patient_id ?sex ?diagnostic_delay ?age_at_diagnosis ?FVC ?weight
?bmi
WHERE {

  ?p a bto:Patient;
     bto:undergo ?d.

  BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

  # Diagnosis Information
  ?d a bto:Diagnosis;
     bto:diagnosticDelay ?diagnostic_delay;
     bto:ageDiagnosis ?age_at_diagnosis.

  OPTIONAL{
     ?d bto:consists ?pft.
     ?pft a bto:PulmonaryFunctionTest;
          bto:pulmonaryFVCRel ?FVC.
  }
  OPTIONAL{
     ?d bto:consists ?ce.
     ?ce a bto:ClinicalEvaluation.
     OPTIONAL{?ce bto:weight ?weight.}
     OPTIONAL{?ce bto:bodyMassIndex ?bmi.}
  }

  BIND(IF(
  EXISTS{
  ?p bto:sex
     "Female"^^<http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral>.
  },"F","M") AS ?sex)
}
```

2.3.2 ALSFRS-R

The subsequent query is used to extract information about ALSFRS-R questionnaires. We recall that tasks 1 and 2 of iDPP@CLEF 2024 considers both ALSFRS-R questionnaires compiled by clinicians or by patients using the BRAINTEASER app. In both cases, we return:

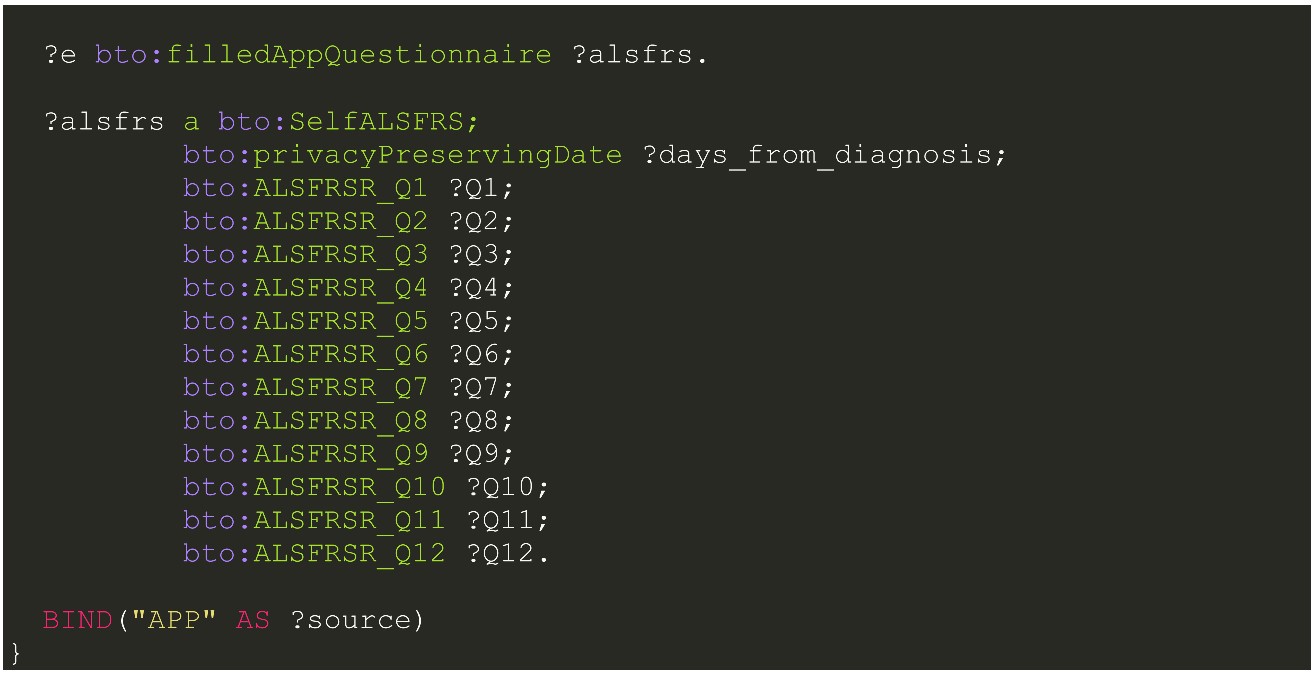
- **patient_id:** Hashed code that uniquely identifies a patient.
- **days_from_diagnosis:** Time since diagnosis expressed in days.
- **source:** Who provided the annotation. Possible values {CT, APP}. CT indicates a value assessed by a clinician, APP indicates a self-assessed value.
- **Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12:** value for the ALSFRS-R questionnaire item.

2.3.2.1 *Clinical ALSFRS-R Query.*

```
SELECT ?patient_id ?days_from_diagnosis ?source ?Q1 ?Q2 ?Q3 ?Q4 ?Q5 ?Q6 ?Q7
?Q8 ?Q9 ?Q10 ?Q11 ?Q12
WHERE {

  ?p a bto:Patient;
     bto:undergo ?e.

  BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

  ?e bto:consists ?alsfrs.

  ?alsfrs a bto:ALSFRS;
          bto:privacyPreservingDate ?days_from_diagnosis;
          bto:alsfrs1 ?Q1;
          bto:alsfrs2 ?Q2;
          bto:alsfrs3 ?Q3;
          bto:alsfrs4 ?Q4;
          bto:alsfrs5 ?Q5;
          bto:alsfrs6 ?Q6;
          bto:alsfrs7 ?Q7;
          bto:alsfrs8 ?Q8;
          bto:alsfrs9 ?Q9;
          bto:alsfrs10 ?Q10;
          bto:alsfrs11 ?Q11;
          bto:alsfrs12 ?Q12.

  BIND("CT" AS ?source)
}
```

2.3.2.2 Patient ALSFRS-R Query.

```
SELECT ?patient_id ?days_from_diagnosis ?source ?Q1 ?Q2 ?Q3 ?Q4 ?Q5 ?Q6 ?Q7
?Q8 ?Q9 ?Q10 ?Q11 ?Q12
WHERE {

  ?p a bto:Patient;
     bto:undergo ?e.

  BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)
```



```
  ?e bto:filledAppQuestionnaire ?alsfrs.

  ?alsfrs a bto:SelfALSFRS;
          bto:privacyPreservingDate ?days_from_diagnosis;
          bto:ALSFRSR_Q1 ?Q1;
          bto:ALSFRSR_Q2 ?Q2;
          bto:ALSFRSR_Q3 ?Q3;
          bto:ALSFRSR_Q4 ?Q4;
          bto:ALSFRSR_Q5 ?Q5;
          bto:ALSFRSR_Q6 ?Q6;
          bto:ALSFRSR_Q7 ?Q7;
          bto:ALSFRSR_Q8 ?Q8;
          bto:ALSFRSR_Q9 ?Q9;
          bto:ALSFRSR_Q10 ?Q10;
          bto:ALSFRSR_Q11 ?Q11;
          bto:ALSFRSR_Q12 ?Q12.

  BIND("APP" AS ?source)
}
```

2.3.3 Sensors Data

The subsequent query is used to extract sensor data measurements. Due to the high number of variables, we split the queries based on related variables.

2.3.3.1 Beat-To-Beat Measurements.

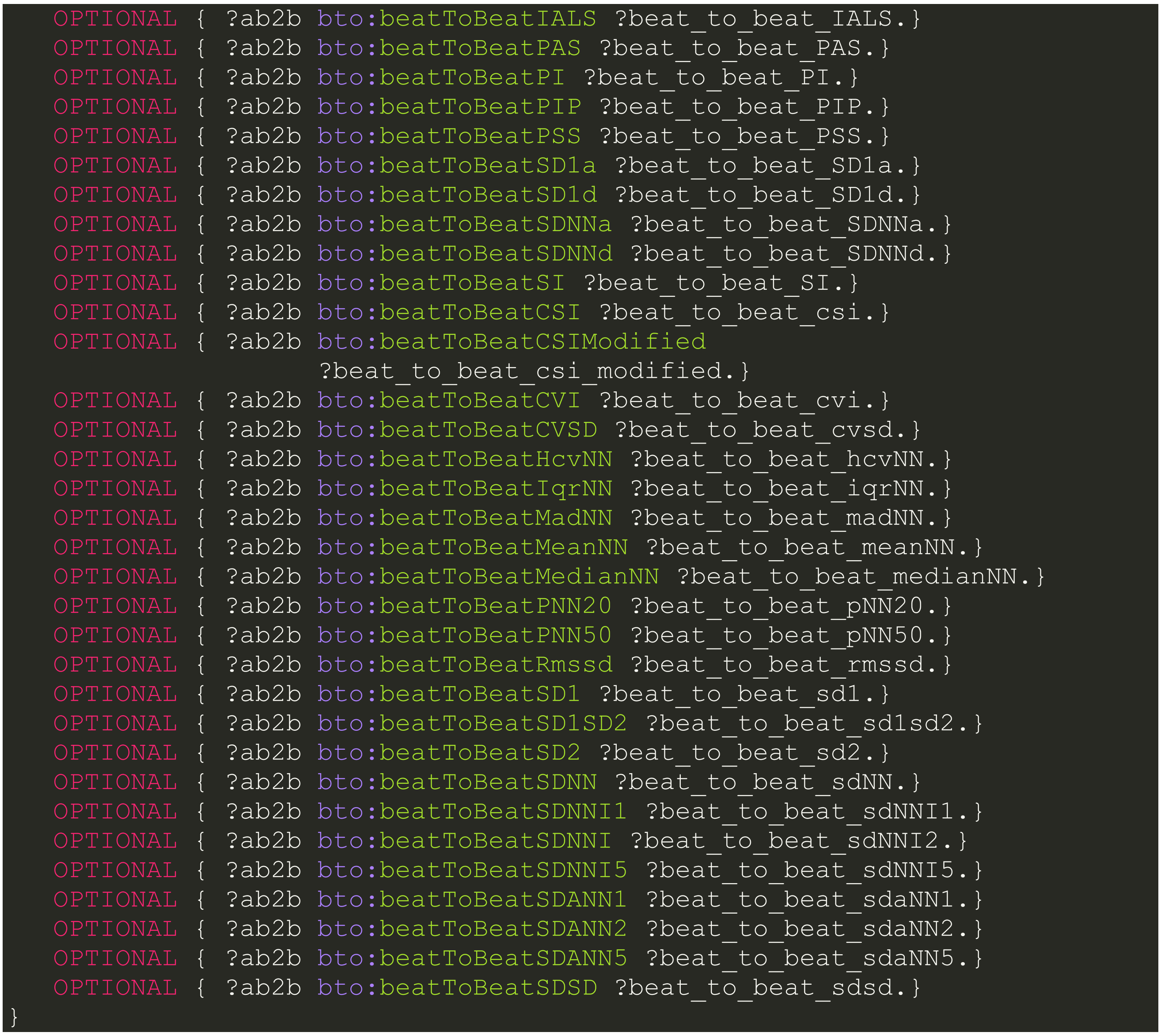
For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **days_from_diagnosis:** Time since diagnosis expressed in days.
- **beat_to_beat_AI**: Area Index, defined as the cumulative area of the sectors corresponding to the points that are located above Line of Identity (LI) divided by the cumulative area of sectors corresponding to all points in the Poincaré plot except those that are located on LI.
- **beat_to_beat_C2a**: The contributions of heart rate acceleration in long-term Heart Rate Variability (HRV).
- **beat_to_beat_C2d**: The contributions of heart rate deceleration in long-term HRV.
- **beat_to_beat_Ca**: The total contributions of heart rate accelerations to HRV.
- **beat_to_beat_Cd**: The total contributions of heart rate accelerations to HRV.
- **beat_to_beat_GI**: Guzik's Index, defined as the distance of points above line of identity (LI) to LI divided by the distance of all points in Poincaré plot to LI except those that are located on LI.
- **beat_to_beat_HTI**: The HRV triangular index, measuring the total number of Inter-beat (RR) intervals divided by the height of the RR intervals histogram.
- **beat_to_beat_IALS**: Inverse of the average length of the acceleration/deceleration segments.
- **beat_to_beat_PAS**: Percentage of NN intervals in alternation segments

- beat_to_beat_PI: Porta's Index, defined as the number of points below LI divided by the total number of points in Poincaré plot except those that are located on LI.
- beat_to_beat_PIP: Percentage of inflection points of the RR intervals series.
- beat_to_beat_PSS: Percentage of short segments
- beat_to_beat_SD1a: Short-term variance of contributions of accelerations (shortenings of RR intervals).
- beat_to_beat_SD1d: Short-term variance of contributions of decelerations (prolongations of RR intervals).
- beat_to_beat_SDNNa: Total variance of contributions of accelerations (shortenings of RR intervals).
- beat_to_beat_SDNNd: Total variance of contributions of decelerations (prolongations of RR intervals).
- beat_to_beat_SI: Slope Index, defined as the phase angle of points above LI divided by the phase angle of all points in Poincaré plot except those that are located on LI.
- beat_to_beat_csi: The Cardiac Sympathetic Index, calculated by dividing the longitudinal variability of the Poincaré plot by its transverse variability.
- beat_to_beat_csi_modified: The modified CSI obtained by dividing the square of the longitudinal variability by its transverse variability.
- beat_to_beat_cvi: The Cardiac Sympathetic Index, equal to the logarithm of the product of longitudinal variability of the Poincaré plot by its transverse variability.
- beat_to_beat_cvsd: The root mean square of the sum of successive differences (RMSSD) divided by the mean of the RR intervals (MeanNN).
- beat_to_beat_hcvNN: The median absolute deviation of the RR intervals (MadNN) divided by the median of the absolute differences of their successive differences (MedianNN).
- beat_to_beat_iqrNN: The interquartile range (IQR) of the RR intervals.
- beat_to_beat_madNN: The median absolute deviation of the RR intervals.
- beat_to_beat_meanNN: The mean of the RR intervals.
- beat_to_beat_medianNN: The median of the absolute values of the successive differences between RR intervals.
- beat_to_beat_pNN20: The proportion of RR intervals greater than 20ms, out of the total number of RR intervals.
- beat_to_beat_pNN50: The proportion of RR intervals greater than 50ms, out of the total number of RR intervals.
- beat_to_beat_rmssd: The square root of the mean of the sum of successive differences between adjacent RR intervals.
- beat_to_beat_sd1: SD1 is a measure of the spread of RR intervals on the Poincaré plot perpendicular to the line of identity. It is an index of short-term RR interval fluctuations, i.e., beat-to-beat variability. It is equivalent (although on another scale) to RMSSD, and therefore it is redundant to report correlations with both.

- **beat_to_beat_sd1sd2**: The ratio between short- and long-term fluctuations of the RR intervals (SD1 divided by SD2).
- **beat_to_beat_sd2**: Measure of the spread of RR intervals on the Poincaré plot along the line of identity. It is an index of long-term RR interval fluctuations.
- **beat_to_beat_sdNN**: The standard deviation of the RR intervals.
- **beat_to_beat_sdNNI1**: The mean of the standard deviations of RR intervals extracted from 1-minute segments of time series data.
- **beat_to_beat_sdNNI2**: The mean of the standard deviations of RR intervals extracted from 2-minutes segments of time series data.
- **beat_to_beat_sdNNI5**: The mean of the standard deviations of RR intervals extracted from 5-minutes segments of time series data.
- **beat_to_beat_sdaNN1**: The standard deviation of average RR intervals extracted from 1-minute segments of time series data.
- **beat_to_beat_sdaNN2**: The standard deviation of average RR intervals extracted from 2-minutes segments of time series data.
- **beat_to_beat_sdaNN5**: The standard deviation of average RR intervals extracted from 5-minutes segments of time series data.
- **beat_to_beat_sdsd**: The standard deviation of the successive differences between RR intervals.

```
SELECT    ?patient_id    ?measure_days_from_diagnosis    ?beat_to_beat_AI
?beat_to_beat_C2a   ?beat_to_beat_C2d   ?beat_to_beat_Ca   ?beat_to_beat_Cd
?beat_to_beat_GI  ?beat_to_beat_HTI  ?beat_to_beat_IALS  ?beat_to_beat_PAS
?beat_to_beat_PI  ?beat_to_beat_PIP  ?beat_to_beat_PSS  ?beat_to_beat_SD1a
?beat_to_beat_SD1d        ?beat_to_beat_SDNNa         ?beat_to_beat_SDNNd
?beat_to_beat_SI       ?beat_to_beat_csi       ?beat_to_beat_csi_modified
?beat_to_beat_cvi         ?beat_to_beat_cvsd          ?beat_to_beat_hcvNN
?beat_to_beat_iqrNN       ?beat_to_beat_madNN        ?beat_to_beat_meanNN
?beat_to_beat_medianNN      ?beat_to_beat_pNN20        ?beat_to_beat_pNN50
?beat_to_beat_rmssd        ?beat_to_beat_sd1         ?beat_to_beat_sd1sd2
?beat_to_beat_sd2         ?beat_to_beat_sdNN         ?beat_to_beat_sdNNI1
?beat_to_beat_sdNNI2      ?beat_to_beat_sdNNI5       ?beat_to_beat_sdaNN1
?beat_to_beat_sdaNN2 ?beat_to_beat_sdaNN5 ?beat_to_beat_sdsd
WHERE {
   ?p a bto:Patient;
      bto:undergo ?e.
   BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

   ?e bto:hasRegisteredActivity ?ab2b.
   ?ab2b a bto:BeatToBeatData;
         bto:privacyPreservingDate ?measure_days_from_diagnosis.

   OPTIONAL { ?ab2b bto:beatToBeatAI ?beat_to_beat_AI.}
   OPTIONAL { ?ab2b bto:beatToBeatC2a ?beat_to_beat_C2a.}
   OPTIONAL { ?ab2b bto:beatToBeatC2d ?beat_to_beat_C2d.}
   OPTIONAL { ?ab2b bto:beatToBeatCa ?beat_to_beat_Ca.}
   OPTIONAL { ?ab2b bto:beatToBeatCd ?beat_to_beat_Cd.}
   OPTIONAL { ?ab2b bto:beatToBeatGI ?beat_to_beat_GI.}
   OPTIONAL { ?ab2b bto:beatToBeatHTI ?beat_to_beat_HTI.}
```



```
    OPTIONAL { ?ab2b bto:beatToBeatIALS ?beat_to_beat_IALS.}
    OPTIONAL { ?ab2b bto:beatToBeatPAS ?beat_to_beat_PAS.}
    OPTIONAL { ?ab2b bto:beatToBeatPI ?beat_to_beat_PI.}
    OPTIONAL { ?ab2b bto:beatToBeatPIP ?beat_to_beat_PIP.}
    OPTIONAL { ?ab2b bto:beatToBeatPSS ?beat_to_beat_PSS.}
    OPTIONAL { ?ab2b bto:beatToBeatSD1a ?beat_to_beat_SD1a.}
    OPTIONAL { ?ab2b bto:beatToBeatSD1d ?beat_to_beat_SD1d.}
    OPTIONAL { ?ab2b bto:beatToBeatSDNNa ?beat_to_beat_SDNNa.}
    OPTIONAL { ?ab2b bto:beatToBeatSDNNd ?beat_to_beat_SDNNd.}
    OPTIONAL { ?ab2b bto:beatToBeatSI ?beat_to_beat_SI.}
    OPTIONAL { ?ab2b bto:beatToBeatCSI ?beat_to_beat_csi.}
    OPTIONAL { ?ab2b bto:beatToBeatCSIModified
                     ?beat_to_beat_csi_modified.}
    OPTIONAL { ?ab2b bto:beatToBeatCVI ?beat_to_beat_cvi.}
    OPTIONAL { ?ab2b bto:beatToBeatCVSD ?beat_to_beat_cvsd.}
    OPTIONAL { ?ab2b bto:beatToBeatHcvNN ?beat_to_beat_hcvNN.}
    OPTIONAL { ?ab2b bto:beatToBeatIqrNN ?beat_to_beat_iqrNN.}
    OPTIONAL { ?ab2b bto:beatToBeatMadNN ?beat_to_beat_madNN.}
    OPTIONAL { ?ab2b bto:beatToBeatMeanNN ?beat_to_beat_meanNN.}
    OPTIONAL { ?ab2b bto:beatToBeatMedianNN ?beat_to_beat_medianNN.}
    OPTIONAL { ?ab2b bto:beatToBeatPNN20 ?beat_to_beat_pNN20.}
    OPTIONAL { ?ab2b bto:beatToBeatPNN50 ?beat_to_beat_pNN50.}
    OPTIONAL { ?ab2b bto:beatToBeatRmssd ?beat_to_beat_rmssd.}
    OPTIONAL { ?ab2b bto:beatToBeatSD1 ?beat_to_beat_sd1.}
    OPTIONAL { ?ab2b bto:beatToBeatSD1SD2 ?beat_to_beat_sd1sd2.}
    OPTIONAL { ?ab2b bto:beatToBeatSD2 ?beat_to_beat_sd2.}
    OPTIONAL { ?ab2b bto:beatToBeatSDNN ?beat_to_beat_sdNN.}
    OPTIONAL { ?ab2b bto:beatToBeatSDNNI1 ?beat_to_beat_sdNNI1.}
    OPTIONAL { ?ab2b bto:beatToBeatSDNNI ?beat_to_beat_sdNNI2.}
    OPTIONAL { ?ab2b bto:beatToBeatSDNNI5 ?beat_to_beat_sdNNI5.}
    OPTIONAL { ?ab2b bto:beatToBeatSDANN1 ?beat_to_beat_sdaNN1.}
    OPTIONAL { ?ab2b bto:beatToBeatSDANN2 ?beat_to_beat_sdaNN2.}
    OPTIONAL { ?ab2b bto:beatToBeatSDANN5 ?beat_to_beat_sdaNN5.}
    OPTIONAL { ?ab2b bto:beatToBeatSDSD ?beat_to_beat_sdsd.}
}
```

2.3.3.2 Heart Rate Measurements.

For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **days_from_diagnosis:** Time since diagnosis expressed in days.
- **heart_rate_baseline**: The baseline heart rate (at stimulus onset).
- **heart_rate_linear_trend**: The parameter corresponding to the linear trend of heart rate.
- **heart_rate_max_time**: The time at which maximum heart rate occurs.
- **heart_rate_maximum**: The maximum heart rate after stimulus onset.
- **heart_rate_mean**: The mean heart rate after stimulus onset.
- **heart_rate_min_time**: The time at which minimum heart rate occurs.
- **heart_rate_minimum**: The minimum heart rate after stimulus onset.

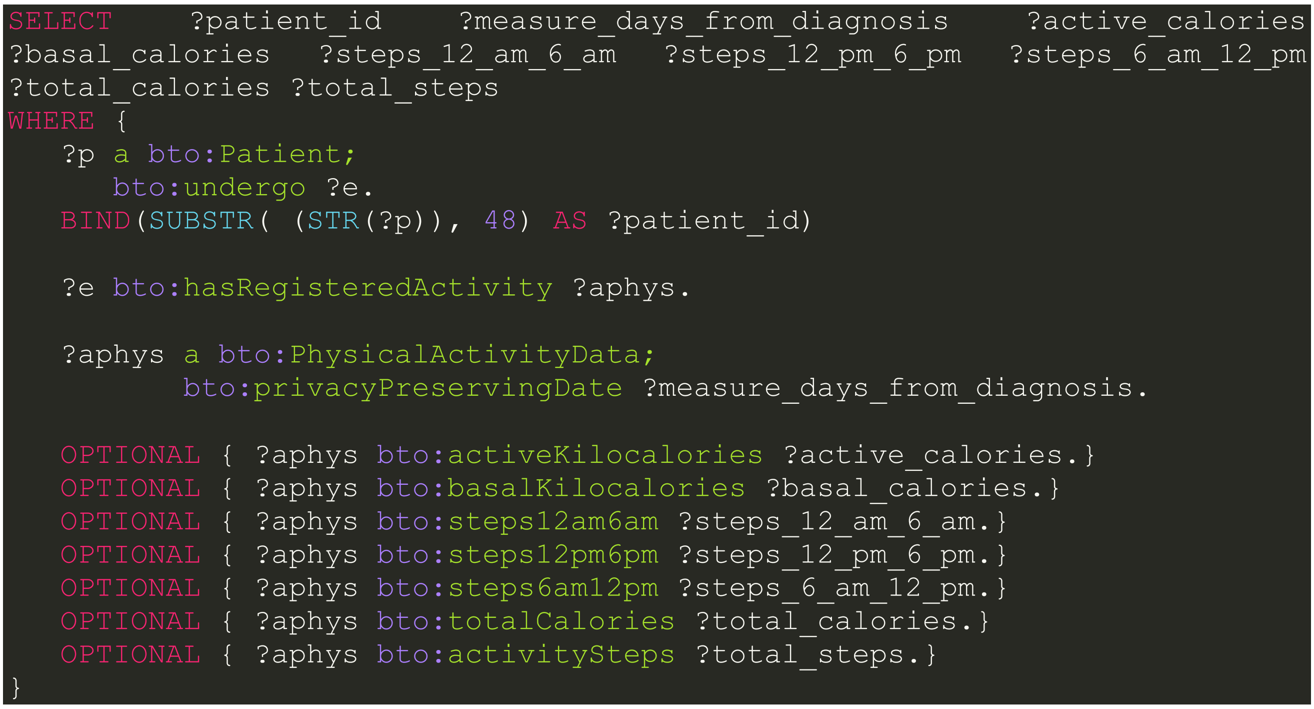
- **heart_rate_quadratic_trend**: The parameter corresponding to the curvature of heart rate.
- **heart_rate_r2**: The quality of the quadratic model.
- **heart_rate_std**: The standard deviation of the heart rate after stimulus onset.

```
SELECT   ?patient_id   ?measure_days_from_diagnosis   ?heart_rate_baseline
?heart_rate_linear_trend      ?heart_rate_max_time      ?heart_rate_maximum
?heart_rate_mean          ?heart_rate_min_time          ?heart_rate_minimum
?heart_rate_quadratic_trend ?heart_rate_r2 ?heart_rate_std
WHERE {
   ?p a bto:Patient;
      bto:undergo ?e.
   BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

   ?e bto:hasRegisteredActivity ?ah.

   ?ah a bto:HeartRateMeasurement;
       bto:privacyPreservingDate ?measure_days_from_diagnosis.

   OPTIONAL { ?ah bto:baselineHeartRate ?heart_rate_baseline.}
   OPTIONAL { ?ah bto:linearTrendHeartRate ?heart_rate_linear_trend.}
   OPTIONAL { ?ah bto:maxTimeHeartRate ?heart_rate_max_time.}
   OPTIONAL { ?ah bto:maxHeartRate ?heart_rate_maximum.}
   OPTIONAL { ?ah bto:averageHeartRate ?heart_rate_mean.}
   OPTIONAL { ?ah bto:minTimeHeartRate ?heart_rate_min_time.}
   OPTIONAL { ?ah bto:minHeartRate ?heart_rate_minimum.}
   OPTIONAL { ?ah bto:quadraticTrendHeartRate
                  ?heart_rate_quadratic_trend.}
   OPTIONAL { ?ah bto:r2HeartRate ?heart_rate_r2.}
   OPTIONAL { ?ah bto:stdHeartRate ?heart_rate_std.}
}
```

2.3.3.3 Physical Activity Measurements.

For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **days_from_diagnosis:** Time since diagnosis expressed in days.
- **active_calories**: Total active calories.
- **basal_calories**: Total basal calories.
- **steps_12_am_6_am**: Total steps within the considered segment of data and between 12 AM and 6 AM daily.
- **steps_12_pm_6_pm**: Total steps within the considered segment of data and between 12 PM and 6 PM.
- **steps_6_am_12_pm**: Total steps within the considered segment of data and between 6 AM and 12 PM.
- **total_calories**: Sum of active and basal calories.
- **total_steps**: Total steps within the considered segment of data.



```
SELECT    ?patient_id    ?measure_days_from_diagnosis    ?active_calories
?basal_calories  ?steps_12_am_6_am  ?steps_12_pm_6_pm  ?steps_6_am_12_pm
?total_calories ?total_steps
WHERE {
   ?p a bto:Patient;
      bto:undergo ?e.
   BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

   ?e bto:hasRegisteredActivity ?aphys.

   ?aphys a bto:PhysicalActivityData;
          bto:privacyPreservingDate ?measure_days_from_diagnosis.

   OPTIONAL { ?aphys bto:activeKilocalories ?active_calories.}
   OPTIONAL { ?aphys bto:basalKilocalories ?basal_calories.}
   OPTIONAL { ?aphys bto:steps12am6am ?steps_12_am_6_am.}
   OPTIONAL { ?aphys bto:steps12pm6pm ?steps_12_pm_6_pm.}
   OPTIONAL { ?aphys bto:steps6am12pm ?steps_6_am_12_pm.}
   OPTIONAL { ?aphys bto:totalCalories ?total_calories.}
   OPTIONAL { ?aphys bto:activitySteps ?total_steps.}
}
```

2.3.3.4 Respiratory Measurements.

For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **days_from_diagnosis:** Time since diagnosis expressed in days.
- **respiration_ApEn**: The approximate entropy.
- **respiration_DFA_alpha1**: The "short-term" fluctuation value generated from Detrended Fluctuation Analysis i.e. the root mean square deviation from the fitted trend of the breath-to-breath intervals. Will only be computed if there are more than 160 breath cycles in the respiratory rate.
- **respiration_DFA_alpha2**: The long-term fluctuation value. Will only be computed if there are more than 640 breath cycles in the respiratory rate.
- **respiration_RMSSD**: The root mean square of successive differences of the breath-to-breath intervals.
- **respiration_SD1**: A measure of the spread of breath-to-breath intervals on the Poincaré plot perpendicular to the line of identity. It is an index of short-term variability.
- **respiration_SD2**: SD2 is a measure of the spread of breath-to-breath intervals on the Poincaré plot along the line of identity. It is an index of long-term variability.
- **respiration_SD2SD1**: The ratio between short- and long-term fluctuations of the breath-to-breath intervals (SD2 divided by SD1).
- **respiration_SDBB**: The standard deviation of the breath-to-breath intervals.
- **respiration_SDSD**: The standard deviation of the successive differences between adjacent
- **respiration_SampEn**: The sample entropy.

- **respiration_alpha1_DimMean**: Multifractal DFA. Dimmean is the mean of singularity dimensions.
- **respiration_alpha1_DimRange**: Multifractal DFA. DimRange is the range of singularity dimensions, corresponding to the height of the singularity spectrum.
- **respiration_alpha1_ExpMean**: Multifractal DFA of short-term fluctuations. ExpMean is the mean of singularity exponents.
- **respiration_alpha1_ExpRange**: Multifractal DFA of short-term fluctuations. ExpRange is the range of singularity exponents, corresponding to the width of the singularity spectrum.
- **respiration_alpha2_DimMean**: Multifractal DFA. Dimmean is the mean of singularity dimensions.
- **respiration_alpha2_DimRange**: Multifractal DFA. DimRange is the range of singularity dimensions, corresponding to the height of the singularity spectrum.
- **respiration_alpha2_ExpMean**: Multifractal DFA of long-term fluctuations. ExpMean is the mean of singularity exponents.
- **respiration_alpha2_ExpRange**: Multifractal DFA of long-term fluctuations. ExpRange is the range of singularity exponents, corresponding to the width of the singularity spectrum.

```
SELECT    ?patient_id    ?measure_days_from_diagnosis    ?respiration_ApEn
?respiration_DFA_alpha1     ?respiration_DFA_alpha2     ?respiration_RMSSD
?respiration_SD1  ?respiration_SD2  ?respiration_SD2SD1  ?respiration_SDBB
?respiration_SDSD      ?respiration_SampEn      ?respiration_alpha1_DimMean
?respiration_alpha1_DimRange                    ?respiration_alpha1_ExpMean
?respiration_alpha1_ExpRange                    ?respiration_alpha2_DimMean
?respiration_alpha2_DimRange                    ?respiration_alpha2_ExpMean
?respiration_alpha2_ExpRange
WHERE {
   ?p a bto:Patient;
      bto:undergo ?e.
   BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

   ?e bto:hasRegisteredActivity ?aresp.

   ?aresp a bto:RespiratoryData;
         bto:privacyPreservingDate ?measure_days_from_diagnosis.
   OPTIONAL { ?aresp bto:respirationApEn ?respiration_ApEn.}
   OPTIONAL { ?aresp bto:respirationDFAalpha1 ?respiration_DFA_alpha1.}
   OPTIONAL { ?aresp bto:respirationDFAalpha2 ?respiration_DFA_alpha2.}
   OPTIONAL { ?aresp bto:respirationRMSSD ?respiration_RMSSD.}
   OPTIONAL { ?aresp bto:respirationSD1 ?respiration_SD1.}
   OPTIONAL { ?aresp bto:respirationSD2 ?respiration_SD2.}
   OPTIONAL { ?aresp bto:respirationSD2SD1 ?respiration_SD2SD1.}
   OPTIONAL { ?aresp bto:respirationSDBB ?respiration_SDBB.}
   OPTIONAL { ?aresp bto:respirationSDSD ?respiration_SDSD.}
   OPTIONAL { ?aresp bto:repirationSampEn ?respiration_SampEn.}
   OPTIONAL { ?aresp bto:respirationAlpha1DimMean
                     ?respiration_alpha1_DimMean.}
   OPTIONAL { ?aresp bto:respirationAlpha1DimRange
                     ?respiration_alpha1_DimRange.}
   OPTIONAL { ?aresp bto:respirationAlpha1ExpMean
                     ?respiration_alpha1_ExpMean.}
   OPTIONAL { ?aresp bto:respirationAlpha1ExpRange
                     ?respiration_alpha1_ExpRange.}
   OPTIONAL { ?aresp bto:respirationAlpha2DimMean
                     ?respiration_alpha2_DimMean.}
   OPTIONAL { ?aresp bto:respirationAlpha2DimRange
                     ?respiration_alpha2_DimRange.}
   OPTIONAL { ?aresp bto:respirationAlpha2ExpMean
                     ?respiration_alpha2_ExpMean.}
   OPTIONAL { ?aresp bto:respirationAlpha2ExpRange
                     ?respiration_alpha2_ExpRange.}
}
```

2.3.3.5 SPO2 Measurements.

For each patient, we return:

- **patient_id:** Hashed code that uniquely identifies a patient.
- **days_from_diagnosis:** Time since diagnosis expressed in days.
- **spo2_AOD100**: Cumulative area of desaturations under the 100% SpO2 level as baseline and normalized by the total recording time
- **spo2_AODmax**: The area under the oxygen desaturation event curve, using the maximum SpO2 value as baseline and normalized by the total recording time
- **spo2_AV**: Average of the pulse oximetry.
- **spo2_CA**: Integral SpO2 below the 90% SpO2 level normalized by the total recording time
- **spo2_CT**: Percentage of the time spent below the 90% oxygen saturation level.
- **spo2_DI**: Delta Index.
- **spo2_M**: Percentage of the pulse oximetry 90% below median oxygen saturation.
- **spo2_MED**: Median of the pulse oximetry.
- **spo2_Min**: Minimum value of the pulse oximetry.
- **spo2_ODI**: The average number of desaturation events per hour (int).
- **spo2_P**: Percentile (90th).
- **spo2_POD**: Percentage of oxygen desaturation events
- **spo2_RG**: SpO2 range (difference between the max and min value).
- **spo2_SD**: Standard deviation of the pulse oximetry.
- **spo2_ZC**: Number of zero-crossing points.

```
SELECT ?patient_id ?measure_days_from_diagnosis ?spo2_AOD100 ?spo2_AODmax
?spo2_AV ?spo2_CA ?spo2_CT ?spo2_DI ?spo2_M ?spo2_MED ?spo2_Min ?spo2_ODI
?spo2_P ?spo2_POD ?spo2_RG ?spo2_SD ?spo2_ZC
WHERE {
   ?p a bto:Patient;
      bto:undergo ?e.
   BIND(SUBSTR( (STR(?p)), 48) AS ?patient_id)

   ?e bto:hasRegisteredActivity ?aspo2.
   ?aspo2 a bto:SPO2Data;
          bto:privacyPreservingDate ?measure_days_from_diagnosis.

   OPTIONAL { ?aspo2 bto:spo2AOD100 ?spo2_AOD100.}
   OPTIONAL { ?aspo2 bto:spo2AODMax ?spo2_AODmax.}
   OPTIONAL { ?aspo2 bto:spo2AV ?spo2_AV.}
   OPTIONAL { ?aspo2 bto:spo2CA ?spo2_CA.}
   OPTIONAL { ?aspo2 bto:spo2CT ?spo2_CT.}
   OPTIONAL { ?aspo2 bto:spo2DI ?spo2_DI.}
   OPTIONAL { ?aspo2 bto:spo2M ?spo2_M.}
   OPTIONAL { ?aspo2 bto:spo2MED ?spo2_MED.}
   OPTIONAL { ?aspo2 bto:spo2Min ?spo2_Min.}
   OPTIONAL { ?aspo2 bto:spo2ODI ?spo2_ODI.}
   OPTIONAL { ?aspo2 bto:spo2P ?spo2_P.}
   OPTIONAL { ?aspo2 bto:spo2POD ?spo2_POD.}
   OPTIONAL { ?aspo2 bto:spo2RG ?spo2_RG.}
   OPTIONAL { ?aspo2 bto:spo2SD ?spo2_SD.}
   OPTIONAL { ?aspo2 bto:spo2ZC ?spo2_ZC.}
}
```

3 RELEASED DATASETS

Compared to the previous deliverable (D9.5), we release three datasets: two new datasets for ALS and an extension of the iDPP 2023 dataset concerning MS. More in detail, the two new ALS datasets comprise a common training part with 52 training patients, whose ALSFRS-R scores were both annotated by the clinicians and self-assessed. Concerning the test sets, 21 and 11 patients were included in them for Task 1 and Task 2, respectively.

Regarding MS, the part of the dataset concerning static variables and MS-related information is the same as the one used for iDPP 2023. The major improvement regards environmental data that have been added to the dataset.

3.1 Tasks 1 and 2: ASL Dataset with Clinical or self-assessed ALSFRS-R

The datasets for Task 1 and Task 2 were collected from ALS-diagnosed patients recruited during the BRAINTEASER project from three centers in Lisbon, Madrid, and Turin. At recruitment, patients were given a commercial fitness tracker (the Garmin VivoActive 4 smartwatch), and data from its sensors was collected during a follow-up period with a median duration of 270 days. Patients were encouraged to wear the watch as much as they were comfortable with, ideally all the time, both while awake and sleeping. Each day of data for each patient was summarized into a vector of 90 statistics related to heart rate and beat-to-beat interval, respiration rate, and nocturnal pulse oximetry. Sensor data was not available every day for each patient.

During the same period, disease progression was assessed by their clinician using the ALSFRS-R questionnaire (roughly every three months, following standard clinical practice). Patients also used the same questionnaire to self-assess their progression through a smartphone app developed specifically by the BRAINTEASER project. They were prompted for the assessment once per month, though the actual frequency varied and depended on patient compliance.

3.1.1 Creation of the datasets

Patients with insufficient data were excluded from the challenge dataset. Specifically, this included those with less than three months of follow-up data, those with more than 50% of sensor data missing, and those without at least two clinical or self-assessed ALSFRS-R evaluations. After applying these criteria, a dataset of 83 patients was obtained, with a median of 254 days of sensor data per patient. These patients and their data were then divided into a training group (common to both Tasks 1 and 2) and two task-specific testing groups.

3.1.2 Split into training and test

The patients were split into three groups:

- **training** patients with at least two clinical and two self-assessed ALSFRS-R evaluations;
- **test-ct** patients with at least two clinical but without two self-assessed ALSFRS-R evaluations;

- **test-app** patients with at least two self-assessed but without two clinical ALSFRS-R evaluations.

The training set thus included 52 patients with a median of 3.5 clinical and 5 self-assessed ALSFRS-R evaluations (189 and 301 in total, respectively). The test-ct set (the test set for Task 1) included 21 patients, whose first clinical ALSFRS-R evaluations were included as features and the second evaluations were the prediction target. The test-app set (the test set for Task 2) included 11 patients and was built in the same way using the self-assessed ALSFRS-R evaluations. The full available sensor data for all patients was included in both the training and test datasets, while only the clinical (resp. self-assessed) ALSFRS-R evaluations were included for Task 1 (resp. Task 2). A comparative description of the datasets is shown in Table 1.

Table 1. Comparison between training and test populations for Task 1 and 2. Continuous variables are presented as median (interquartile range); categorical variables as count (percentage on available data), for each level. "Sensor adherence" is the ratio of days with available sensor data during the whole sensor follow-up

Variable	Level	Task 1/2 Train	Task 1 test	Task 2 test
Sex	Female	11 (21.15%)	9 (42.86%)	4 (36.36%)
Sex	Male	41 (78.85%)	12 (57.14%)	7 (63.64%)
Diagnostic delay (months)	median (IQR)	0.8 (0.4-1.3)	0.9 (0.4-1.8)	1.0 (0.4-1.6)
Age at diagnosis	median (IQR)	56 (49-64)	62 (57-66)	60 (52-66)
FVC	median (IQR)	85 (79-95)	84 (79-98)	92 (79-113)
Weight	median (IQR)	75 (64-81)	67 (60-71)	65 (60-70)
BMI	median (IQR)	25 (23-27)	24 (22-26)	22 (21-25)
ALSFR-R CT (count)	median (IQR)	3.5 (2.0-5.0)	-	-
ALSFR-R APP (count)	median (IQR)	5.0 (3.0-8.0)	-	-
Sensor follow-up (months)	median (IQR)	9.8 (5.2-13.6)	8.9 (5.3-14.2)	5.9 (5.5-8.3)
Sensor adherence	median (IQR)	98% (89%-100%)	98% (85%-100%)	100% (99%-100%)

3.2 Task 3: MS Dataset

The dataset used for Task 3 in iDPP@CLEF 2024 is structured similarly to those from iDPP@CLEF 2023, though some features (e.g., evoked potentials, MRIs) were not included, and certain records have been filtered based on the purpose of the task.

3.2.1 Updates over IDPP@CLEF 2023

In the 2024 dataset, EDSS data before January 1, 2013 (aligned with the start of environmental data collection) were filtered, and patients without EDSS follow-ups were removed. Additionally, patients who did not experience a relapse after their first non-filtered EDSS follow-up (i.e., the baseline for each patient) were excluded.

The dataset has been expanded to incorporate environmental data, which includes information on patients' exposure to various air pollutants identified as significant public health risks in the latest World Health Organization (WHO) global air quality guidelines, such as particulate matter (PM) - encompassing both $PM_{2.5}$ (particles with an aerodynamic diameter of 2.5 micrometers or less) and PM_{10} (particles with an aerodynamic diameter of 10 micrometers or less) - as well as ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and several weather factors (including wind speed, relative humidity, sea level pressure, global radiation, precipitation, and average, minimum, and maximum temperatures).

Air pollutant data from public monitoring stations were collected daily from the European Air Quality Portal using the DiscoMap tool[1]. The geographical coordinates (longitude and latitude) of each monitoring station were matched to specific postcodes, identifying the nearest station to each patient's residence postcode. Instead, weather data were gathered daily from the European Climate Assessment and Dataset station network, which provides access to the E-OBS dataset, a daily gridded land-only observational dataset over Europe[2]. Each grid was matched with the nearest monitoring station using Euclidean distance based on geographical coordinates. This approach ensured that air pollution and weather data were aligned with the same spatial and temporal granularity. Daily environmental measurements were aggregated into weekly averages from each patient's baseline. As additional features, the number of days per week spent over the respective WHO recommended air quality guideline levels for short-term (24 hours) exposure was computed for each air pollutant.

Finally, a subset of 380 MS patients from the Turin and Pavia research centers was selected for Task 3 in iDPP@CLEF 2024, compared to 550 patients for Task 1 and 638 for Task 2 in iDPP@CLEF 2023. The resulting MS dataset[3] includes static variables with demographic and clinical information, EDSS scores with corresponding Functional System (FS) sub-scores, environmental measurements, and the outcome time, representing the week of the first relapse occurrence after the baseline for each patient. EDSS follow-ups are reported between the baseline and the outcome time, while environmental measurements span from January 1, 2013, to December 30, 2023. It is important to note that environmental data may have gaps due to availability. When considering only environmental data preceding the outcome time, the median number of weeks available for each patient is 59, with an interquartile range of 103.25 weeks. The distributions of air pollutant concentrations (measured in micrograms per cubic meter), averaged across patients over these weeks, are depicted in the boxplots of Figure 2.

[1] https://discomap.eea.europa.eu/Index

[2] https://www.ecad.eu/download/ensembles/download.php

[3] https://brainteaser.dei.unipd.it/challenges/idpp2024/assets/other/ms/ms-variables-description.txt

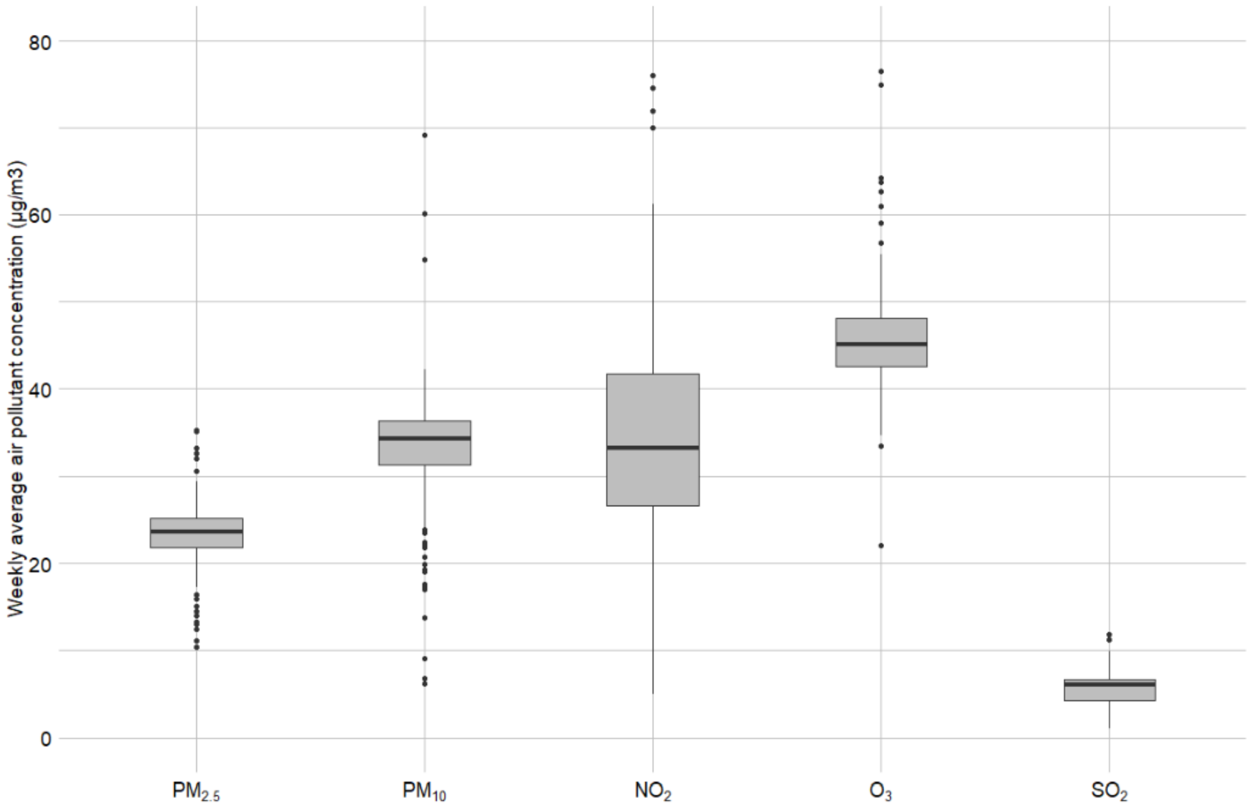


Figure 2. Boxplots of weekly air pollutant concentrations averaged across patients.

3.2.2 Split into training and test

The dataset was split into a training set (70\%) and a test set (30\%), with subjects stratified by outcome time to ensure an even distribution across both sets. The distribution of static data, including demographic and clinical information, and EDSS were verified to be similar in both training and test sets. Additionally, since environmental exposure is considered, the distribution of patients from the two clinical centres and their residence classification (Cities, Rural Areas, and Towns) was checked to be balanced.

Statistical tests, including the Kruskal-Wallis test for continuous variables and the Chi-squared test for categorical and ordinal variables, were performed to assess the appropriateness of the stratification. Special attention was given to sparsely observed levels in categorical variables to ensure rare levels appeared only in the training set if at all. Table 2 provides a comparison of variable distributions between the training and test sets, confirming that the split meets the best-practice quality standards.

Table 2. Comparison between training and test populations for MS task. Continuous variables are presented as median (interquartile range); categorical variables as count (percentage on available data), for each level.

		Train	Test
Sex	Female	148 (74.37%)	54 (66.67%)

		Train	Test
	Male	51 (25.63%)	27 (33.33%)
Ethnicity	Caucasian	181 (90.96%)	77 (95.06%)
	Hispanic	2 (1.00%)	-
	Black African	2 (1.00%)	-
	NA	14 (7.04%)	4 (4.94%)
Residence classification	Cities	53 (26.63%)	20 (24.69%)
	Rural Area	52 (26.13%)	22 (27.16%)
	Towns	94 (47.24%)	39 (48.15%)
Centre	Pavia	129 (64.82%)	58 (71.61%)
	Turin	70 (35.18%)	23 (28.39%)
Occurrence of MS in pediatric age	FALSE	176 (88.44%)	77 (95.06%)
	TRUE	23 (11.56%)	4 (4.94%)
Age at onset	median (IQR)	28 (22-36)	30 (24-34)
Age at baseline	median (IQR)	38 (31-47)	38 (33-47)
Diagnostic delay	median (IQR)	12 (4-47)	12 (3-28)
Spinal cord symptom	FALSE	143 (71.86%)	54 (66.67%)
	TRUE	56 (28.14%)	27 (33.33%)
Brainstem symptom	FALSE	146 (73.37%)	57 (70.37%)
	TRUE	53 (26.63%)	24 (29.63%)
Eye symptom	FALSE	148 (74.37%)	59 (72.84%)
	TRUE	51 (25.63%)	22 (27.16%)
Supratentorial symptom	FALSE	140 (70.35%)	50 (61.73%)

		Train	Test
	TRUE	59 (29.65%)	31 (38.27%)
Other symptoms	FALSE	197 (99.00%)	80 (98.77%)
	Sensory	1 (0.50%)	1 (1.23%)
	Epilepsy	1 (0.50%)	-
EDSS	median (IQR)	2.0 (1.5-3.0)	2.0 (1.5-3.5)
	NA	3 (0.36%)	0 (0.00%)
Outcome time	median (IQR)	59 (24-122)	53 (25-130)

4 ZENODO REPOSITORY

The data described in this deliverable are publicly available according to the FAIR principles on Zenodo, at the link: https://zenodo.org/records/12789962.

The datasets are shared in two formats:

- RDF (serialized in Turtle[4]) modelled according to the BRAINTEASER Ontology (BTO);
- CSV, as shared during the iDPP@CLEF 2022 and 2023 challenges, split into training and test.

Each format corresponds to a specific folder in the datasets, where a dedicated README file provides further details on the datasets.

4.1 ALS Data

In the remainder, we report the guidelines on how to use the data available on the Zenodo repository.

4.1.1 RDF Data

We share a privacy-preserving snapshot of the BRAINTEASER Knowledge Base (KB) which contains the information used to generate the datasets associated with tasks 1 and 2 of the iDPP@CLEF 2024 (http://brainteaser.dei.unipd.it/challenges/idpp2024/) challenge, focused on predicting the progression of Amyotrophic Lateral Sclerosis (ALS).

The BRAINTEASER KB has been produced following the BRAINTEASER Ontology (BTO) (https://w3id.org/brainteaser/ontology), developed by the BRAINTEASER project, which ensures the consistency of the represented data. Moreover, several checks have been

[4] https://www.w3.org/TR/turtle/

performed to ensure that all the instances are clean, contain proper values in the expected ranges, and do not have contradictions.

4.1.1.1 Data Overview

Starting from this RDF dataset, it is possible to generate the same CSV files associated with tasks 1 and 2 of the iDPP@CLEF 2024 (http://brainteaser.dei.unipd.it/challenges/idpp2024/) challenge. The only exception is the file **test-target.csv**, which can only be partially generated due to the structure of the ontology. We provide some SPARQL queries that will produce the desired CSV file.

In **./data/**, a set of RDF files can be found. In particular, **brtALS.ttl** contains all patients information.

brtALS.ttl file includes 2,201,367 statements and comprises 86 patients, 247 clinical ALSFRS-R, 323 self assessed ALSFRS-R and 21,456 rows of sensors data.

4.1.1.2 Details on the queries

In **./queries/**, a set of documents describing each SPARQL query that generates the desired CSV file can be found. In particular:

ALS_query_alsfrs_APP.txt generates the files `alsfrs.csv`, containing all the events about self assessed ALSFRS-R.

ALS_query_alsfrs_CT.txt generates the files `alsfrs.csv`, containing all the events about clinical ALSFRS-R.

query_ALS_static.txt generates the file `static.csv`, containing all the patient data.

4.1.1.3 Sensors Data

Due to the high number of variables in the mentioned CSV file, we provide a set of queries which will generate a set of CSV files that can be merged together to compose **sensor.csv**. In particular:

ALS_query_sensors_b2b.txt extracts all sensors data about beat-to-beat measurements.

ALS_query_sensors_heartrate.txt extracts all sensors data about heart rate measurements.

ALS_query_sensors_physical.txt extracts all sensors data about physical activity measurements.

ALS_query_sensors_respiratory.txt extracts all sensors data about respiratory measurements.

ALS_query_sensors_spo2.txt extracts all sensors data about SPO2 measurements.

4.1.2 ALS Tasks

We will evaluate proposals of different approaches to predict the ALSFRS-R scores for ALS patients.

Participants are asked to predict the ALSFRS-R scores based on the patient data collected over an average of nine months via a dedicated app developed by the BRAINTEASER project and sensor data collected from the sensors of a fitness smartwatch in the context of clinical trials in Turin, Pavia, Lisbon, and Madrid, fully anonymized.

- Task1: Predicting ALSFRS-R Score from Sensor Data (ALS): It focuses on predicting the twelve scores of the ALSFRS-R (ALS Functional Rating Scale - Revised), assigned by medical doctors roughly every three months, from the sensor data collected via the app. The ALSFRS-R is a somehow "subjective" evaluation usually performed by a medical doctor and this task will help in answering a currently open question in the research community, i.e. whether it could be derived from objective factors.
- Task2: Predicting Patient Self-assessment Score from Sensor Data: It focuses on predicting the self-assessment score assigned by patients from the sensor data collected via the app. Self-assessment scores correspond to each of the ALSFRS-R scores but, while the latter ones are assigned by medical doctors during visits, the former ones are assigned via auto-evaluation by patients themselves using the provided app. If the self-assessment performed by patients, more frequently than the assessment performed by medical doctors every three months or so, can be reliably predicted by sensor and app data, we can imagine a proactive application which, monitoring the sensor data, alerts the patient if an assessment is needed.

4.1.3 ALS Tabular Data

We share the datasets associated with task 1 and 2 of the iDPP@CLEF 2024 (http://brainteaser.dei.unipd.it/challenges/idpp2024/) challenge. Tasks 1 and 2 share the same training dataset, comprising the following data:

- Static patient data, including demographic and clinical information;
- All the ALSFRS-R evaluations (twelve scores for each) from the medical doctor assessment or self-assessed through the app, each with the time of collection;
- All available sensor data, summarized by day, each with the time of collection.

All times are expressed in days between the diagnosis and the collection. Note that sensor data is not available for every day and may contain gaps due to lack of adherence from the patient and/or technical problems in the data collection. However, finding ways to work around these limitations is part of the challenge.

The test data will contain the same static and sensor data for other patients not in the training set. Only the first ALSFRS-R evaluation will be available, together with the time of the target ALSFRS-R evaluation whose scores are to be predicted.

For more information about tasks, please see the iDPP@CLEF 2024 Participation Guidelines (http://brainteaser.dei.unipd.it/challenges/idpp2024/).

4.1.3.1 Training Dataset

Tasks 1 and 2 share the same training dataset, available in **./data/Task1/train/** and **./data/Task2/train/**. It comprises 52 patients, 189 clinical ALSFRS-R, 301 self-assessed ALSFRS-R, 13946 days of sensor data.

4.1.3.2 Test Datasat

There is a separate dataset for each task:

- Task 1 (clinical ALSFRS-R): available in **./data/Task1/test/**

  29 Patients, 29 clinical ALSFRS-R, 6,345 rows of sensor data.
- Task 2 (app ALSFRS-R): available in **./data/Task2/test/**

  11 Patients, 11 self-assessed ALSFRS-R, 2,347 rows of sensor data.

4.1.3.3 Details on the Datasets

Each dataset contains the following files, where **split** correspond to either **train** or **test**:

- **[split]-static.csv**: contains the static variables about each patient. In the following document, you can find a description of each column found in the CSV file available at **./data/features-description/als-variables-description.txt**.
- **[split]-alsfrs.csv**: contains the twelve ALSFRS-R scores, together with the (relative) time of assessment and whether it was evaluated by a clinician or self assessed. For **test-alsfrs.csv**, the time of assessment relative to the diagnosis is called "first_alsfrs_days_from_diagnosis" and the relative time of the second assessment is "target_alsfrs_days_from_diagnosis" whose corresponding ALSFRS-R values need to be predicted. In the following document, you can find a description of each column found in the CSV file available at **./data/features-description/als-variables-description.txt**.
- **[split]-sensors.csv**: contains the daily summaries of the sensor data measured by the smartwatch. In the following document, you can find a description of each column found in the CSV file available at **./data/features-description/als-variables-description.txt**.
- **test-target.csv**: it contains the twelve predicted ALSFRS-R scores at the time "target_alsfrs_days_from_diagnosis".

4.2 MS Data

4.2.1 RDF Data

We share a privacy-preserving snapshot of the BRAINTEASER Knowledge Base (KB) which contains the information used to generate the datasets associated with task 3 of the iDPP@CLEF 2024 (http://brainteaser.dei.unipd.it/challenges/idpp2024/) challenge, focused on predicting the progression of Multiple Sclerosis (MS).

The BRAINTEASER KB has been produced following the BRAINTEASER Ontology (BTO) (https://w3id.org/brainteaser/ontology), developed by the BRAINTEASER project, which ensures the consistency of the represented data. Moreover, several checks have been performed to ensure that all the instances are clean, contain proper values in the expected ranges, and do not have contradictions.

4.2.1.1 Data Overview

In **./data/**, a set of RDF files can be found. In particular, **brtMS.ttl** contains all patients information, while **brtMSEnv.ttl** comprises environmental measurements.

brtMS.ttl file includes 6,118,377 statements and comprises 280 patients, 280 relapses, and 1,124 EDSS scores.

Starting from this RDF dataset, it is possible to generate the same CSV files associated with task 3 of the iDPP@CLEF 2024 (http://brainteaser.dei.unipd.it/challenges/idpp2024/) challenge. We provide the SPARQL queries which will produce the desired CSV file.

4.2.1.2 Details on the Queries

In **./queries/**, a set of documents describing each SPARQL query that generates the desired CSV file can be found.

In particular:

- **MS_query_static.txt** generates the file **static.csv**, containing the static variables about a patient.
- **MS_query_edss.txt** generates the file **edss.csv**, containing the (relative) date when EDSS scores were measured, together with the EDSS score evaluated by clinicians.
- **MS_query_MS_environmental.txt** generates the file **environmental_meas.csv**, containing the environmental measurements.
- **MS_query_outcome_relapses.txt** generates the file **outcome.csv**, containing the (relative)date of the relapses.

4.2.2 MS Task: Predicting Relapses from EDDS Sub-scores and Environmental Data

It focuses on predicting a relapse using environmental data and EDSS (Expanded Disability Status Scale) sub-scores. This task allows us to assess if exposure to different pollutants is a useful variable in predicting a relapse.

Participants will be asked to predict the week of the first relapse after the baseline considering environmental data based on a weekly granularity, given the status of the patient at the baseline, which is the first visit available in the considered time span. For each patient, the date of the baseline will be week 0 and all the other weeks will be relative to it.

Participants will be given all the observations and environmental data about a patient, i.e. also observations which may happen after the relapse to be predicted. All the patients are guaranteed to experience, at least, one relapse after the baseline.

4.2.3 MS Tabular Data

We share the datasets associated with task 3 of the iDPP@CLEF 2024 (http://brainteaser.dei.unipd.it/challenges/idpp2024/) challenge, focused on predicting the relapses of Multiple Sclerosis (MS) patients.

4.2.3.1 Data Overview

Task 3 encompasses the following data:

- Static patients data, comprising demographic and clinical information;
- EDSS scores assessed by clinicians, alongside their corresponding sub-scores, each annotated with the week of the EDSS visit;
- Environmental exposure measurements aggregated weekly;

- The week of the first relapse occurrence after the baseline for each patient (all patients are guaranteed to experience at least one relapse after the baseline).

All time intervals are expressed in weeks from the baseline. EDSS visits occur between the baseline and the occurrence of the first relapse, while environmental measurements span from January 1st, 2013, to 2023. It's important to note that environmental measurements may not be available for every week and contain missing data.

For more information on the EDSS score, please refer to https://www.neurology.org/doi/10.1212/WNL.33.11.1444 .

Training and test datasets are divided in a 70-30% proportion. The test data will include the same static, EDSS, and environmental data for patients not included in the training set.

- Training Dataset available in **./data/train/**
- 199 patients, 834 EDSS scores, 113,923 environmental measurements.
- Test Dataset available in **./data/test/**
- 81 patients, 290 EDSS scores, 46,354 environmental measurements.

4.2.3.2 Details on the Datasets

Each dataset contains the following files, where **split** correspond to either **train** or **test**:

- **[split]_static.csv**: contains the static variables about a patient.
- **[split]_edss.csv**: contains the (relative) week when EDSS scores were measured, together with the EDSS scores evaluated by clinicians.
- **[split]_environmental_meas.csv**: contains the environmental measurements aggregated with a weekly granularity.
- **[split]_outcome.csv**: contains the week of the first relapse occurrence after the baseline for each patient.

In **./features-description/**, a document describing each column of the CSV files can be found.

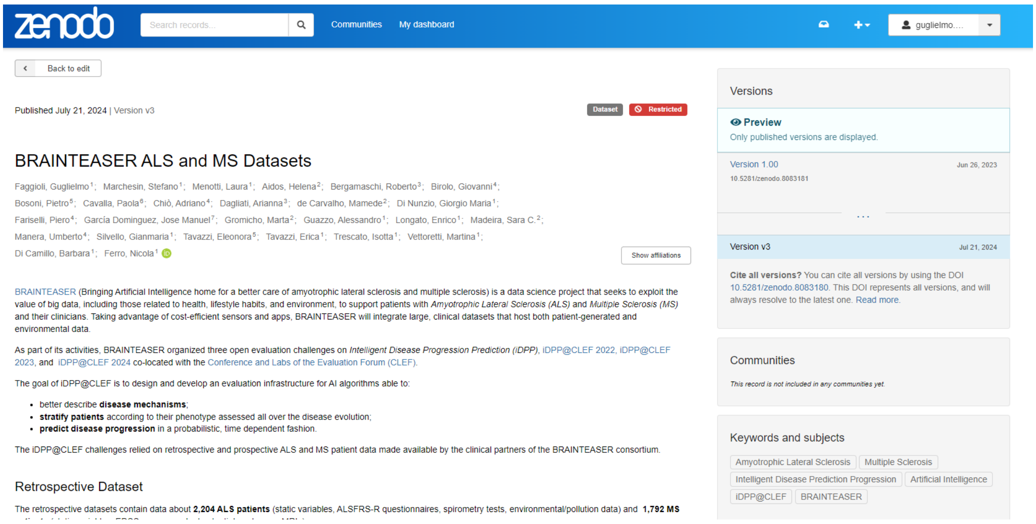
More in detail, the **outcomes.csv** file adopts the following format:

```
67396654612589370083623092407810766693,190
267505643482073532971907590162817517438,96
292211325266845962706770530341466085902,30
272084949450943266236977822325902693890,216
291361686090369065397020169907588173057,13
...
```

where:

- Columns are separated by a comma;
- The first column is the patient ID, a hashed version of the original patient ID (should be considered just as a string);

- The second column is the week of the first relapse occurrence after the baseline for each patient.



Published July 21, 2024 | Version v3

BRAINTEASER ALS and MS Datasets

Faggioli, Guglielmo[1]; Marchesin, Stefano[1]; Menotti, Laura[1]; Aidos, Helena[2]; Bergamaschi, Roberto[3]; Birolo, Giovanni[4]; Bosoni, Pietro[5]; Cavalla, Paola[6]; Chiò, Adriano[4]; Dagliati, Arianna[3]; de Carvalho, Mamede[2]; Di Nunzio, Giorgio Maria[1]; Fariselli, Piero[4]; García Dominguez, Jose Manuel[7]; Gromicho, Marta[2]; Guazzo, Alessandro[1]; Longato, Enrico[1]; Madeira, Sara C.[2]; Manera, Umberto[4]; Silvello, Gianmaria[1]; Tavazzi, Eleonora[5]; Tavazzi, Erica[1]; Trescato, Isotta[1]; Vettoretti, Martina[1]; Di Camillo, Barbara[1]; Ferro, Nicola[1]

BRAINTEASER (Bringing Artificial Intelligence home for a better care of amyotrophic lateral sclerosis and multiple sclerosis) is a data science project that seeks to exploit the value of big data, including those related to health, lifestyle habits, and environment, to support patients with *Amyotrophic Lateral Sclerosis (ALS)* and *Multiple Sclerosis (MS)* and their clinicians. Taking advantage of cost-efficient sensors and apps, BRAINTEASER will integrate large, clinical datasets that host both patient-generated and environmental data.

As part of its activities, BRAINTEASER organized three open evaluation challenges on *Intelligent Disease Progression Prediction (iDPP)*, iDPP@CLEF 2022, iDPP@CLEF 2023, and iDPP@CLEF 2024 co-located with the Conference and Labs of the Evaluation Forum (CLEF).

The goal of iDPP@CLEF is to design and develop an evaluation infrastructure for AI algorithms able to:

- better describe **disease mechanisms**;
- **stratify patients** according to their phenotype assessed all over the disease evolution;
- **predict disease progression** in a probabilistic, time dependent fashion.

The iDPP@CLEF challenges relied on retrospective and prospective ALS and MS patient data made available by the clinical partners of the BRAINTEASER consortium.

Retrospective Dataset

The retrospective datasets contain data about **2,204 ALS patients** (static variables, ALSFRS-R questionnaires, spirometry tests, environmental/pollution data) and **1,792 MS**

Versions

Preview
Only published versions are displayed.

Version 1.00 Jun 26, 2023
10.5281/zenodo.8083181

Version v3 Jul 21, 2024

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.8083180. This DOI represents all versions, and will always resolve to the latest one. Read more.

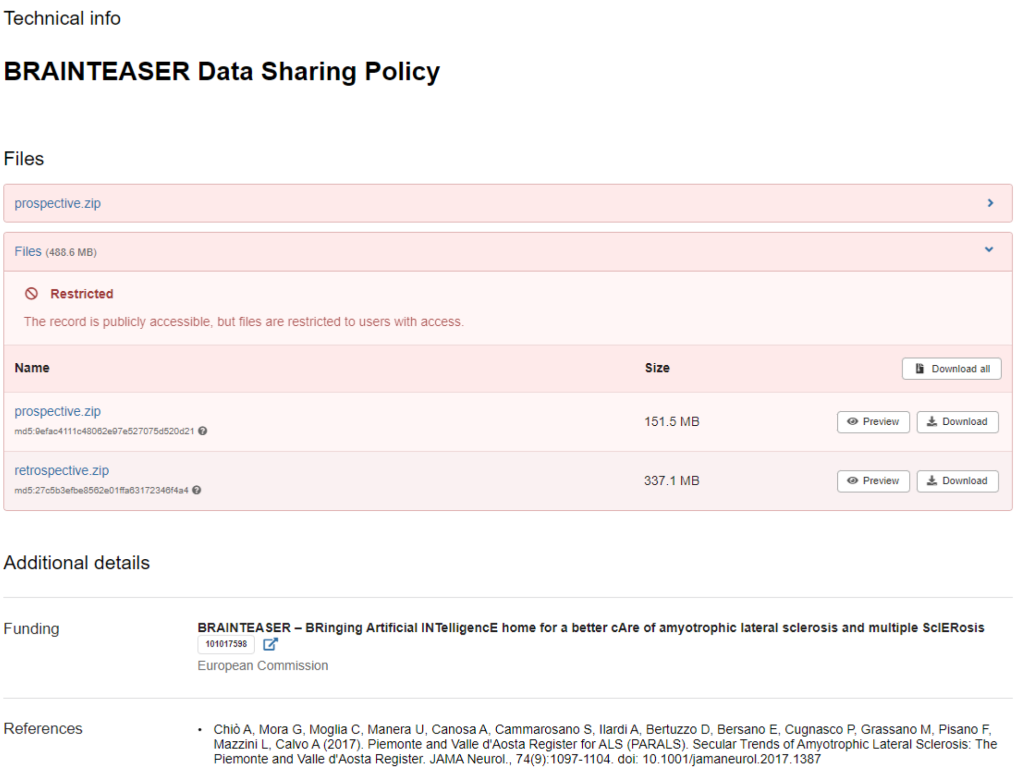
Communities

This record is not included in any communities yet.

Keywords and subjects

Amyotrophic Lateral Sclerosis · Multiple Sclerosis · Intelligent Disease Prediction Progression · Artificial Intelligence · iDPP@CLEF · BRAINTEASER

Figure 3. A screenshot of the BRAINTEASER data Zenodo page



Technical info

BRAINTEASER Data Sharing Policy

Files

prospective.zip

Files (488.6 MB)

Restricted

The record is publicly accessible, but files are restricted to users with access.

Name	Size	Download all
prospective.zip md5:9efac4111c48062e97e527075d520d21	151.5 MB	Preview Download
retrospective.zip md5:27c5b3efbe8562e01ffa63172346f4a4	337.1 MB	Preview Download

Additional details

Funding

BRAINTEASER – BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis
101017598
European Commission

References

- Chiò A, Mora G, Moglia C, Manera U, Canosa A, Cammarosano S, Ilardi A, Bertuzzo D, Bersano E, Cugnasco P, Grassano M, Pisano F, Mazzini L, Calvo A (2017). Piemonte and Valle d'Aosta Register for ALS (PARALS). Secular Trends of Amyotrophic Lateral Sclerosis: The Piemonte and Valle d'Aosta Register. JAMA Neurol., 74(9):1097-1104. doi: 10.1001/jamaneurol.2017.1387

Figure 4. A screenshot of the BRAINTEASER data Zenodo page

5 CONCLUSIONS

This deliverable describes the datasets released for the iDPP@CLEF 2024 challenge that are part of the data publicly released for within the BRAINTERASER project. In this deliverable, we described the datasets, how they have been collected ad processed, and we provided some statistics of their content. In total, we released two new datasets and we extended a previously existing one with environmental. More in details, the two new datasets concern ALS and they contain static data rearding 83 patients and ALSFRS-R scores, both annotated by the clinicians and self-assessed by the patients. Rearding the third dataset, it constructs upon the iDPP@CLEF 2023 (D9.5) MS dataset, by extending it with environmental observations.

The data have been also ingested in a machine-readable format (turtle), according to the Brainteaser ontology, to make them publicly available according to the FAIR principles in Zenodo.

6 REFERENCES

BGT+21	Bettin, M., Guazzo, A., Trescato, I., Longato, E., Hazizaj, E., Dosso, D., Faggioli, G., Di Nunzio, G.M., Silvello, G., Vettoretti, M., Tavazzi, E., Roversi, C., Fariselli, P., Madeira, S.C., de Carvalho, M., Gromicho, M., Chio', A., Manera, U., Dagliati, A., Birolo, G., Aidos, H., Di Camillo, B., Ferro, N.: Deliverable 9.4 – Shared Data Package for the Evaluation Challenge and Integration with EOSC. BRAINTEASER, EU Horizon 2020, Contract N. GA101017598. https://brainteaser.health/ (June 2022)
GTL+22	Guazzo, A., Trescato, I., Longato, E., Hazizaj, E., Dosso, D., Faggioli, G., Di Nunzio, G.M., Silvello, G., Vettoretti, M., Tavazzi, E., Roversi, C., Fariselli, P., Madeira, S.C., de Carvalho, M., Gromicho, M., Chiò, A., Manera, U., Dagliati, A., Birolo, G., Aidos, H., Di Camillo, B., Ferro, N.: Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022.In: Barrón-Cedeño, A., Da San Martino, G., Degli Esposti, M., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany (2022)
GTL+22b	Guazzo, A., Trescato, I., Longato, E., Hazizaj, E., Dosso, D., Faggioli, G., Di Nunzio, G.M., Silvello, G., Vettoretti, M., Tavazzi, E., Roversi, C., Fariselli, P., Madeira, S.C., de Carvalho, M., Gromicho, M., Chiò, A., Manera, U., Dagliati, A., Birolo, G., Aidos, H., Di Camillo, B., Ferro, N.: Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge.
BDD+21	Bettin, M., Di Nunzio, G.M., Dosso, D., Faggioli, G., Ferro, N., Marchetti, N., Silvello, G.: Deliverable 9.1 – Project ontology and

	terminology, including data mapper and RDF graph builder. BRAINTEASER, EU Horizon 2020, Contract N. GA101017598. https://brainteaser.health/ (December 2021)
GDPR	General Data Protection Regulation (GDPR, EU Regulation 2016/679)
BMT+21	Bergamaschi R, Monti MC, Trivelli L, Mallucci G, Gerosa L, Pisoni E, Montomoli C.PM2.5 exposure as a risk factor for multiple sclerosis. An ecological study with a Bayesian mapping approach. Environ Sci Pollut Res Int. 2021 Jan;28(3):2804-2809.
CMM+17	Chiò A, Mora G, Moglia C, Manera U, Canosa A, Cammarosano S, Ilardi A, Bertuzzo D, Bersano E, Cugnasco P, Grassano M, Pisano F, Mazzini L, Calvo A; Piemonte and Valle d'Aosta Register for ALS (PARALS). Secular Trends of Amyotrophic Lateral Sclerosis: The Piemonte and Valle d'Aosta Register. JAMA Neurol. 2017 Sep 1;74(9):1097-1104. doi: 10.1001/jamaneurol.2017.1387. PMID: 28692730; PMCID: PMC5710181.
FGM+23	Faggioli, Guglielmo, Guazzo, Alessandro, Marchesin, Stefano, Menotti, Laura, Trescato, Isotta, Aidos, Helena, Bergamaschi, Roberto, Birolo, Giovanni, Cavalla, Paola, Chiò, Adriano, Dagliati, Arianna, de Carvalho, Mamede, Di Nunzio, Giorgio Maria, Fariselli, Piero, García Dominguez, Jose Manuel, Gromicho, Marta, Longato, Enrico, Madeira, Sara C., Manera, Umberto, … Ferro, Nicola. (2023). BRAINTEASER ALS and MS Datasets (1.00) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8083181