The BRAINTEASER Datasets: Clinical, Wearable and Environmental Data for ALS & MS Progression Modeling

Guglielmo Faggioli[1,*,†], Laura Menotti[1,*,†], Stefano Marchesin[1,*,†], Isotta Trescato[1,†], Lara Ahmad[2], Helena Aidos[3], Anca Loredana Alungulese[4], Riccardo Bellazzi[5], Roberto Bergamaschi[2], Giovanni Birolo[7], Pietro Bosoni[5], Maria Fernanda Cabrera-Umpierrez[8], Paola Cavalla[9,10], Adriano Chiò[9,10,11], Arianna Dagliati[5], Mamede de Carvalho[12], Piero Fariselli[7], Jose Manuel García Dominguez[13], Sergio Gonzalez Martinez[8], Marta Gromicho[12], Alessandro Guazzo[1], Aleksandar Jovanović[14], Borko Kostić[14], Enrico Longato[1], Sara C. Madeira[3], Umberto Manera[9,10], Jose Luis Muñoz Blanco[4], Eleonora Tavazzi[2], Erica Tavazzi[1], Elena Trasobares Iglesias[15], Vladimir Urošević[14], Martina Vettoretti[1], Giorgio Maria Di Nunzio[1], Gianmaria Silvello[1], Barbara Di Camillo[1], and Nicola Ferro[1]

[1]Department of Information Engineering, University of Padova, Padova, Italy
[2]IRCCS Mondino Foundation, Pavia, Italy
[3]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[4]ALS and Neuromuscular Disorders Department. Gregorio Marañón University Hospital, Madrid, Spain
[5]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
[6]University of Pavia, Pavia, Italy
[7]University of Turin, Turin, Italy
[8]Life Supporting Technologies, Universidad Politecnica de Madrid, Spain
[9]Azienda Ospedaliero Universitaria Città della Salute e della Scienza, Turin, Italy
[10]Rita Levi Montalcini Department of Neuroscience, University of Turin, Turin, Italy
[11]Institute of Cognitive Sciences and Technologies, C.N.R, Rome, Italy
[12]Faculty of Medicine - University of Lisbon, Lisbon, Portugal
[13]Gregorio Marañon Hospital in Madrid, Madrid, Spain
[14]Belit, Serbia
[15]Gregorio Marañón Health Research Institute, Madrid, Spain
[†]these authors contributed equally to this work
[*]corresponding authors: Guglielmo Faggioli (guglielmo.faggioli@unipd.it), Laura Menotti (laura.menotti@unipd.it), Stefano Marchesin (stefano.marchesin@unipd.it)

ABSTRACT

Amyotrophic lateral sclerosis (ALS) and multiple sclerosis (MS) are debilitating diseases with unpredictable progression. Artificial Intelligence-based tools for modelling disease progression could significantly improve the quality of life for patients and caregivers while supporting clinicians in delivering more personalized and timely care. However, the limited availability of data hinders the development, testing, and reproducibility of such predictive tools. To address this challenge, we curated, in the context of the H2020 BRAINTEASER project, four datasets containing clinical data from a total of 2,290 ALS patients and 723 MS patients. These datasets also include environmental data and information collected through wearable devices. Unlike most existing resources, the BRAINTEASER datasets are gathered from clinical practice, offering a more accurate representation of the data that an AI progression prediction tool would encounter in real-world scenarios. In addition to manual and automated data quality checks, the research community has validated the datasets through three editions of the intelligent Disease Progression Prediction challenges held within the Conference and Labs of the Evaluation Forum (CLEF).

Background & Summary

Amyotrophic lateral sclerosis (ALS) and multiple sclerosis (MS) are severe neurological diseases, each with distinct mechanisms and disease progression. However, they share some common symptoms and features, such as multisystem involvement (motor, cerebellar, brainstem, sensory, sphincter, visual, and cognitive impairments) and the gradual accumulation of disability[1,2]. Both diseases significantly affect the quality of life of individuals living with them and their caregivers[3]. The rapid, progressive neurodegeneration caused by ALS typically results in a deterioration of movement, speech, breathing, and swallowing abilities[4]. In contrast, the chronic and variable course of MS is generally characterized by periods of relapse, where symptoms worsen or new ones appear, followed by remission periods with a reduction in symptom severity[2].

A common and major challenge in managing ALS and MS is the difficulty in predicting their progression. Patients with these diseases often require hospitalization and varying levels of home-based assistance. The unpredictable nature of the diseases also makes it challenging for clinicians to provide timely care at different stages of progression. In this context, automatic predictive tools could greatly benefit patients, caregivers, and clinicians by forecasting disease progression. These tools would also help clinicians better stratify patients based on their phenotype, ensuring that interventions are administered tailored to the subjects' characteristics. Furthermore, these tools could contribute to a better understanding of certain aspects and mechanisms of ALS and MS that remain only partially understood by the research community. Despite the importance of such tools, the absence of publicly available data hinders the development, training, and testing of reproducible automatic predictive tools[5]. The European project "Bringing Artificial Intelligence home for a better care of amyotrophic lateral sclerosis and multiple sclerosis" (BRAINTEASER)[6] addresses these challenges by fostering collaboration between clinicians and engineers to advance the current state-of-the-art in automatic disease progression modelling for ALS and MS.

In this work, we present the ALS and MS datasets we created within BRAINTEASER, designed to support the development of automatic disease progression modelling and prediction tools, improve their reproducibility, and help the research community better understand the underlying mechanisms of ALS and MS[7]. Specifically, we release four *full datasets*: two for ALS and two for MS. These datasets were collected from four medical institutions across Italy, Portugal, and Spain. Originally, the data was used in the iDPP@CLEF challenges[8–10], where we divided it into eight *task-specific datasets*: five for ALS and three for MS. Besides the *full datasets*, we release the *task-specific datasets* as they appeared in the iDPP@CLEF challenges, preserving the same task splits, training, and test sets. This will allow future practitioners, who did not participate in the challenges, to compare their results with those of the challenge participants, favouring reproducibility. Additionally, the release of the *full datasets* will enable their use in different settings and tasks beyond those originally designed for the iDPP@CLEF challenges.

The BRAINTEASER datasets offer real-world clinical, wearable, and environmental data for modelling ALS and MS progression. These datasets address the significant challenge of data scarcity in AI-based predictive tool development for these diseases. These datasets enable the development of improved disease progression models for personalized and timely patient care by predicting outcomes like the need for medical procedures or relapses. Its public availability also fosters collaborative research and accelerates knowledge transfer within the scientific community. The datasets are publicly available and made persistent through a Zenodo repository[7].

Background: ALS Datasets

Regarding ALS, two major datasets are available in the literature: the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials Database) dataset[11] and the Answer ALS dataset[12]. The PRO-ACT dataset contains data on treated and control patients from 30 Phase II/III clinical trials. However, it only includes patients whose characteristics met the inclusion criteria for clinical trials, which may not reflect the general population, and has a relatively short follow-up period. In contrast, the BRAINTEASER dataset includes data from patients in a real-world setting. Additionally, the PRO-ACT dataset does not provide information about genetic mutations, which are instead available in the BRAINTEASER data. Despite its limited representativeness of the general population and lack of gene mutation details, the PRO-ACT dataset remains one of the largest, with over 12,000 patients. Conversely, the Answer ALS dataset contains data from 861 ALS patients. Like the BRAINTEASER dataset, it includes real-world clinical and omics data. Answer-ALS provides a detailed omics profile, as blood-derived iPS motor neurons were generated from each patient and subjected to multi-omics analyses, including whole genome sequencing, RNA transcriptomics, ATAC-Seq, and proteomics. Conversely, it does not include environmental data or data collected through wearable devices, which are present in the BRAINTEASER dataset.

Background: MS Datasets

Regarding MS, four datasets are publicly available. These include the Brain MRI Dataset of Multiple Sclerosis[13], the MICCAI 2016 challenge dataset[14], the MSSEG-2 challenge dataset[15], and the Brain MRI Image DICOM Dataset[16]. The Brain MRI Dataset of Multiple Sclerosis[13] includes multi-sequence MRI data from 60 MS patients, along with manual lesion segmentation, Expanded Disability Status Scale (EDSS) scores, general patient information, and clinical details. The MICCAI 2016 challenge dataset[14] provides demographic and MRI data for 53 patients, while the MSSEG-2 dataset[15] offers MRI data for 100 patients.

Both challenges focused on automating MRI annotation, so there is no information on disease progression, such as patient relapses. The Brain MRI Image DICOM Dataset[16] contains MRI scans, demographic information, case descriptions, and preliminary diagnoses. However, it does not include relapse data and is only accessible for a fee, meaning it does not comply with the Findability, Accessibility, Interoperability, Reusability (FAIR) principles. While all these MS datasets focus on imaging, the BRAINTEASER data aims to describe disease progression, addressing a complementary task.

On a different line, researchers can access registries that contain regional or national-level data from various national health systems. An example is the registry maintained by North American Research Committee on Multiple Sclerosis (NARCOMS)[17]. However, these registries are often not anonymized, and access is typically restricted to medical teams, with data usage governed by privacy regulations like General Data Protection Regulation (GDPR), due to the sensitive nature of the information. In contrast, the BRAINTEASER datasets are anonymized, allowing for unrestricted use regarding purpose and secondary applications.

Background: iDPP@CLEF Challenges

The BRAINTEASER datasets were created as part of the intelligent Disease Progression Prediction at the Conference and Labs of the Evaluation Forum (iDPP@CLEF) challenges. These three challenges, held between 2022 and 2024, focused on developing comparable and reproducible artificial intelligence approaches to predict the progression of ALS and MS. The challenges were organized as "coopetitions": teams competed while collaborating to build a shared evaluation framework and advance knowledge on predictive algorithms for ALS and MS. The main objectives of the challenges were: i) to openly and publicly validate the data and prediction algorithms developed by the BRAINTEASER project; ii) to allow external researchers to build their predictive models using the BRAINTEASER datasets; iii) to provide a shared evaluation framework to ensure the comparability of experimental results; iv) to accelerate knowledge transfer to and from the BRAINTEASER project and facilitate the adoption of best practices; and v) to foster the growth of a research community through annual workshops to discuss challenge results.

We report here a brief overview of the iDPP@CLEF challenges:

- **iDPP@CLEF 2022**[8]: The first edition of the challenge focused exclusively on ALS. It included three tasks: predicting the need for Non-Invasive mechanical Ventilation (NIV) and Percutaneous Endoscopic Gastrostomy (PEG) and predicting the occurrence of death during the patient's follow-up.

- **iDPP@CLEF 2023**[9]: The second edition of the challenge expanded on the three ALS tasks from the first edition by incorporating environmental data to predict the same events. Additionally, two MS tasks were introduced, focused on predicting disease worsening based on changes in the EDSS score, using two different definitions of "worsening."

- **iDPP@CLEF 2024**[10]: The final edition introduced a completely new prospective dataset for ALS. The two ALS tasks for iDPP@CLEF 2024 aimed at predicting changes in the Revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) score, either assessed by clinicians or self-evaluated by patients. There was also a single MS task, which focused on predicting the occurrence of relapses, based on a subset of the MS *full dataset* used in iDPP@CLEF 2023, extended with environmental data.

Methods

In this section, we describe the construction of the BRAINTEASER datasets. Of the four *full datasets*, three are retrospective and one is prospective. The retrospective datasets contain patient data collected before the start of the project, during real-life clinical practice. The prospective dataset includes patient information collected during the project itself. The *full datasets* contain demographic and static clinical data, environmental information (including details about pollutants), which are obtained through public interfaces for retrospective patients or smart devices for prospective ones, and data collected through wearables. In more detail: The retrospective ALS dataset contains data from 2,204 ALS patients, including static clinical variables, ALSFRS-R questionnaires, spirometry tests, and environmental/pollution data. The prospective ALS dataset includes information from 86 patients, comprising static clinical variables, ALSFRS-R questionnaires (assessed either by clinicians and recorded with the clinical BRAINTEASER clinical tool, or by patients via the BRAINTEASER mobile application), and sensor data. Practically, the two sets of data – either collected by the clinicians or self-assessed by the patients – correspond to two different *task-specific datasets*, respectively Task 1 and Task 2, among the prospective data. A more detailed description of the collection of such data is available in the "*Data Collection*" Paragraph of the "*Prospective Full Dataset Curation*" Section. The two retrospective MS *full datasets* contain data on 723 and 280 patients, respectively. These datasets include static clinical variables, EDSS scores, evoked potentials, relapses, and numeric features derived from Magnetic Resonance Imagings (MRIs), with the latter dataset also including environmental and pollution data.

The dataset comprises over 250 variables, including static variables (e.g., patient demographics, previous surgeries and traumas, and health status), data collected during routine visits, disease onset details, diagnosis, and progression, as well as environmental and sensor-based observations. To maintain clarity, we highlight only the main variables and provide select examples in the remainder of the section. The complete list of variables is available at (https://docs.google.com/document/d/1KMadH9lMkFwGMMOFlN7uyjgxkbOyWdekO5Iw89CdaSk).

Retrospective Full Datasets Curation

Data Collection

We describe here the raw data sources exploited to build the retrospective *full datasets* and the filtering process followed to ensure the quality of the datasets. Figure 1 illustrates visually the steps the data underwent to obtain the BRAINTEASER datasets.

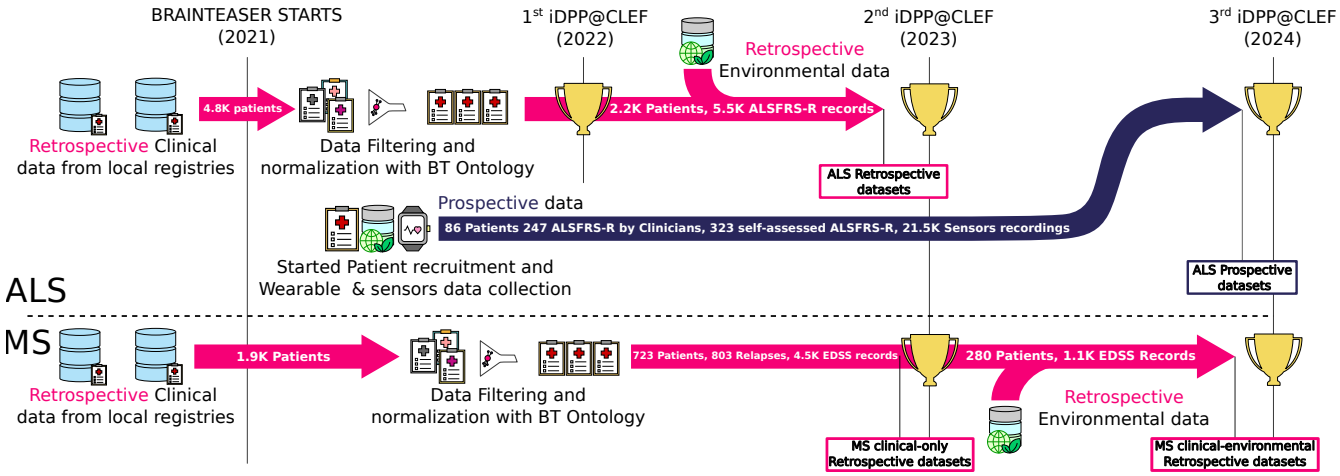


Figure 1. The modifications and processing applied to the raw data to obtain the final BRAINTEASER datasets.

Regarding the collection and storage of informed consent, all patients had provided informed consent for the use of their data for research purposes, following Article 89 of EU Regulation 2016/679 (GDPR), either through their contribution to national or regional registries or as part of other research studies. Every reasonable effort was made to obtain project-specific consent. Following the positive opinion of the relevant ethics committees, project-specific consent was not collected only in cases where it was organizationally impossible to contact the individuals, despite exhaustive attempts to do so. These efforts included verifying patients' vital status, reviewing clinical documentation, attempting contact via any available telephone numbers, and consulting relevant patient or population registries. In such cases, individuals were confirmed to be either deceased or uncontactable at the time of study enrolment.

ALS Clinical Data The ALS retrospective *full dataset* was available to the clinical and research centres participating in the BRAINTEASER project before its commencement, hence it represents the variables typically collected during the clinical practice. This dataset is provided by the Neurology Departments of the University of Turin (UNITO) (Turin, Italy) and of the Instituto De Medicina Molecular João Lobo Antune (iMM) (Lisbon, Portugal). The UNITO dataset is based on the Piemonte and Valle d'Aosta Register for Amyotrophic Lateral Sclerosis (PARALS)[18], a highly reliable dataset that covers two Italian regions, Piedmont and Valle d'Aosta. The register is anonymized, and the data is handled following the Italian Data Protection Code. Patients provided written informed consent to contribute to the registries and for their data to be used in anonymized form. The UNITO source dataset includes information about 3,257 ALS patients collected between January 1995 and December 2018. Most of these patients were monitored at two ALS centres in Torino and Novara. Patients in the UNITO dataset were diagnosed with ALS based on the El Escorial Criteria (EEC)/Revised El Escorial Criteria (EEC-R)[19]. The source dataset contains demographic and static clinical data for each patient, and the date of death was retrieved from municipal records. The data provided by iMM covers 1,562 ALS patients who attended an ALS clinic in Lisbon from 1995 to October 2021. Similar to the UNITO dataset, this database is highly reliable. Moreover, a single group of clinicians used standardized methods to assess all patients, reducing the risk of data bias. Both datasets include demographic information such as the year of birth, sex, ethnic origin, and habits like smoking. They also provide clinical and disease history, including past traumas, surgeries, comorbidities, onset details, and diagnosis date. Additionally, the datasets include details about patients' follow-up visits, conducted at intervals typical of clinical practice (about every 2–4 months) These records contain disease history, neurological

and laboratory findings (such as ALSFRS-R scores and respiratory test results), and treatment information. On average, each patient had five consultations.

MS Clinical Data The two retrospective MS *full datasets* come from two different clinical and research centres: Fondazione Mondino IRCCS (FM), Pavia, Italy, and UNITO. The data provided by FM consists of 1,103 patients diagnosed with MS according to the diagnostic criteria valid at the time of the onset of the disease, while UNITO includes 750 MS patients. Note that diagnostic criteria, including clinical and paraclinical evidence, have changed over the years. Regarding FM, the dataset covers most of the disease phenotypes. The included patients originate from various Italian provinces. All patients are followed regularly at the outpatient of the FM center, and the frequency of the planned visits can vary from one to four per year, with the first year of visit ranging from 1957 to 2021. The dataset also includes, when available, the clinical information derived form each visit and the assessment of the clinical disability, measured via EDSS. Together with this information, demographic data, clinical and instrumental features for disease onset and progression, comorbidities, patients' family history, patients' medical history, and data about pharmacological treatments and other significant life events (e.g., pregnancy, vaccination) were also collected. As with UNITO, the patients are originally from the Piedmont region and are regularly followed at the outpatient of the UNITO center, with a frequency of one to four (planned) visits per year. The dataset contains first visits ranging from 1996 to 2021 Together with the visits, when available, the dataset also provides the derived clinical information and the assessment of the patient's clinical disability via EDSS. Again, demographic data, clinical, and instrumental features for disease onset and progression, comorbidities, patients' family history, patients' medical history, and data about pharmacological treatments and other significant life events were also collected. For both datasets, patients provided written informed consent to release their data, which has been completely anonymized.

Environmental Data For both diseases, environmental data consists of patients' exposure to air pollutants classified as significant public health risks in the World Health Organization (WHO) global air quality guidelines[20]. Such pollutants includes Particulate Matter (PM) with an aerodynamic diameter of 10 μm or less (i.e., PM_{10}) or of 2.5 or less (i.e., $PM_{2.5}$), Ozone (O_3), Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), and Carbon Monoxide (CO). In addition, environmental data include several weather factors, including wind speed, relative humidity, sea level pressure, global radiation, precipitation, and average, minimum, and maximum temperatures. Air pollutant data are gathered daily by the European Air Quality Portal using the DiscoMap tool (https://discomap.eea.europa.eu/Index/). Patients are linked to pollutant exposure by identifying the nearest public monitoring station to their residence. On the other hand, weather data were collected daily by 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 (https://www.ecad.eu/download/ensembles/download.php). Each grid is matched with the nearest monitoring station using Euclidean distance based on geographical coordinates to ensure that all environmental measurements are aligned with the same spatial and temporal granularity.

More in detail, to obtain the environmental data, we downloaded time series of air quality measurements from official monitoring stations, specifically focusing on E1a/validated data. In particular, the Air Quality e-Reporting platform provides detailed information on air quality assessment regimes and lists all individual sampling points and/or models applied for each pollutant within specific zones and agglomerations, as described on the official documentation: https://www.eea.europa.eu/data-and-maps/data/aqereporting-9/air-quality-assessment-regimes. Weather data were collected from the Copernicus E-OBS ensemble dataset, which was generated using a conditional simulation procedure, as documented on the official website: https://surfobs.climate.copernicus.eu/dataaccess/access_eobs.php. According to the description:

> For each ensemble member, a spatially correlated random field is generated based on a pre-calculated spatial correlation function. The mean across all ensemble members is then computed and provided as the "best-guess" field. The spread, representing the 90% uncertainty range, is calculated as the difference between the 5th and 95th percentiles across the ensemble. Starting with version E-OBSv24.0e, all elements now consist of a 20-member ensemble.

Data Filtering

We apply a filtering phase for ALS and MS clinical data to check if each patient's record provides useful and coherent information. More in detail, since the data sources were collected through manual data curation labour, they sometimes contain typographic errors, minor inconsistencies, duplicated information, and missing values. To maximize the quality of the final data release, we adopt a conservative approach, entirely removing patient records with incomplete or potentially inconsistent information. Specifically, only records satisfying the following criteria were included in the final dataset:

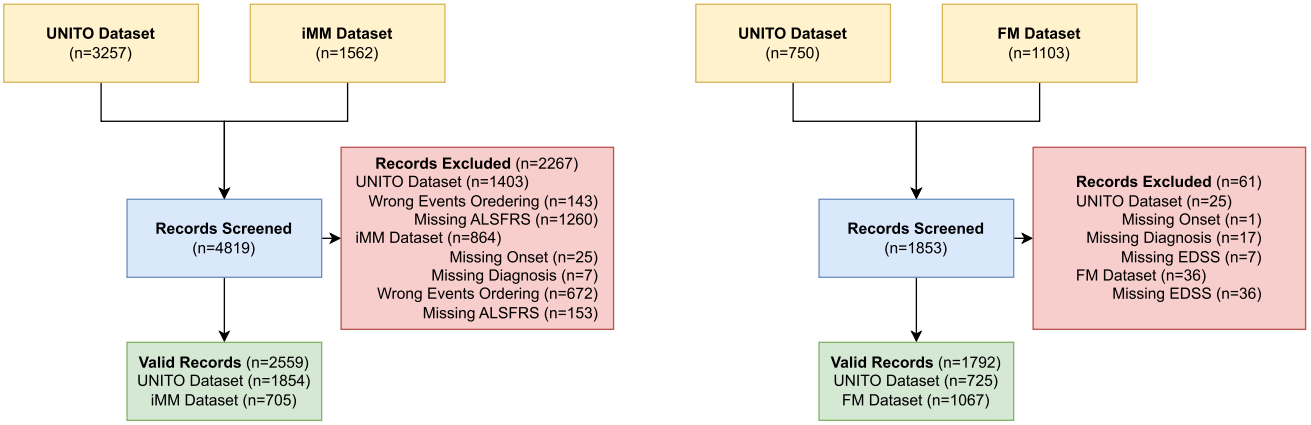
- All the relevant dates are available (e.g., onset date, diagnosis date, date of death for deceased people);
- The dates are in the correct order (e.g., onset before diagnosis, first visit after onset, death after any other event);

- The patient has at least one ALSFRS-R record associated (*applicable only to ALS patients*);
- The patient has at least one EDSS record associated (*applicable only to MS patients*);
- Relevant medical procedures for the ALS iDPP@CLEF tasks (i.e., NIV and PEG) occurred after the first visit (*applicable only to ALS patients*).

If a patient fails the quality control, their record and corresponding visits are dropped. We remove duplicate visits, visits without a date, or with invalid or non-registered values. For instance, we discard ALSFRS-R records without a registered date or where no ALSFRS-R score was registered or the scores are invalid.

About ALS retrospective data, UNITO originally provided 3,257 ALS patients, 15,006 visits with a recorded ALSFRS-R, and 2,890 spirometries, while iMM provided 1,562 ALS patients, 7,446 visits with a recorded ALSFRS-R, and 2,631 spirometries. On the other hand, for MS retrospective data, FM originally provided 1,103 MS patients, 17,529 visits with a recorded EDSS, 4,571 relapses, 5,407 Evoked Potentials (EPs), 8,429 MRI records, and 2,250 records about MS clinical course. UNITO instead provided 750 MS patients, 11,133 visits with a recorded EDSS, 2,164 relapses, 853 EPs, 1,941 MRI records, and 1,510 records about MS clinical course.

The PRISMA diagrams in Figure 2 show the number of removed patients for retrospective data, divided by the medical centre. Concerning ALS retrospective data, from UNITO data, we remove 1,260 patients due to visits without recorded ALSFRS-R and 143 due to issues related to event ordering, leaving us with 1,854 valid patients (57% of the raw dataset). Consequently, we remove 566 ALSFRS-R records due to dropped patients and 27 due to duplicated rows, resulting in 14,413 valid visits (96%). About spirometries, we remove 359 records due to dropped patients and 17 duplicated rows, resulting in 2,514 valid spirometries (87%). On the other hand, from iMM data, we remove 25 patients due to missing onset, seven due to missing diagnosis, 153 due to missing ALSFRS-R, and 672 due to issues in the order of events, leaving us with 705 valid patients (45%). Consequently, we remove 2,352 ALSFRS-R recordings due to dropped patients and 396 due to missing data, resulting in 4,698 valid visits (63%). About spirometries, we remove 928 records due to dropped patients and 51 missing values, leaving us with 1,652 valid spirometries (63%). We consider only patients with six months of visits from the ALS retrospective clinical data and linked them with environmental data to generate the ALS retrospective *full dataset*.



(a) PRISMA Flow Diagram for ALS Data

(b) PRISMA Flow Diagram for MS Data

Figure 2. PRISMA diagrams describing the data filtering for retrospective data. Patients are dropped if they satisfy at least one of the conditions for removal. For simplicity, we consider the following macro-category: “Missing ALSFRS-R" (applicable only to ALS patients), “Missing EDSS" (applicable only to MS patients), “Missing Diagnosis", “Missing Onset", “Event Ordering". All conditions based on the order of two events fall into the last category.

About MS retrospective data, from UNITO data, we remove one patient due to missing onset, 17 due to missing diagnosis date, and seven due to missing EDSS visits, leaving us with 725 valid patients (97%). As a result, we are left with 10,810 valid EDSS visits (97%), 2,028 relapses (94%), 843 EPs (99%), 1,769 MRIs (91%), and 1,466 records related to MS clinical course (97%). From FM data, we remove 36 patients due to missing EDSS, resulting in 1,067 valid patients (97%). Consequently, we are left with 14,479 valid EDSS visits (83%), 4,169 relapses (91%), 5,369 EPs (99%), 5,938 MRIs (70%), and 2,180 records related to MS clinical course (97%). MS retrospective clinical data is further processed to generate the MS retrospective *full dataset* with clinical data only. In particular, visits are restricted to a 2.5-year window, and patients are excluded if they had an EDSS score higher than or equal to three but no other recorded EDSS within one year after the last visit. The MS retrospective

full dataset with environmental data instead is generated by considering a subset of 280 patients for which environmental data were available, i.e., with the first visit after 2013.

Prospective Full Dataset Curation

Data Collection

The prospective ALS *full dataset* is based on observations of patients recruited within the BRAINTEASER project. The patients affected by ALS were recruited and followed by three medical centres in Lisbon (iMM), Turin (UNITO), and Madrid (Servicio Madrileño de Salud (SERMAS)). During the recruitment, patients were given a commercial fitness tracker (the Garmin VivoActive 4 smartwatch), a commercial personal air monitor that tracks exposure (Atmotubo PRO), and an app created within the project (BRAINTEASER Patient App[21]). During the follow-up (median duration of 270 days), the fitness tracker was used to record the vital parameters of the patients. The patients were encouraged to wear the tracker all the time, as long as they felt comfortable. 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. Additionally, every three months, the clinicians recorded the results of the ALSFRS-R questionnaire to monitor the progression of the disease using the BRAINTEASER clinical tool[21]. Furthermore, once a month, the patients self-assessed their progression through the ALSFRS-R questionnaire using the BRAINTEASER app on their smartphone. This allows the researchers to compare the observations made by a professional with the subjective patient experience. Both the BRAINTEASER clinical tool and the BRAINTEASER app were developed in the context of the BRAINTEASER project and are specifically tailored to handle the collection of data concerning patients affected by ALS and MS[21]. Cossu et al. (2024)[22,23] provide a complete description of the collection and processing of the wearable devices' data.

Informed consent was consistently collected from all patients who contributed to the prospective dataset and who were recruited during the BRAINTEASER project.

Data Filtering

The final BRAINTEASER ALS prospective *full dataset* is filtered to remove patients with insufficient information. In particular, we remove the records associated with patients with less than 3 months of follow-up data, patients' records for which more than 50% of the sensor data were missing across the entire monitored period, and those for which there are less than two clinical and self-reported ALSFRS-R. After applying these criteria, a dataset of 86 patients is obtained, with a median of 254 days of sensor data per patient.

Task-specific datasets

From the four BRAINTEASER *full dataset*, we derive eight *task-specific datasets*, which correspond to those used in the iDPP@CLEF challenges. Each *task-specific datasets* is split into training and test sets. Specifically, 80% of the data is allocated to the training set, and the remaining 20% to the test set. We randomly split the data except for the retrospective ALS *task-specific datasets*, where we stratify on the outcome time. In both cases, we ensure that the patient distribution in the training and test sets is similar in demographics (e.g., sex, ethnic origin) and relevant static clinical attributes (e.g., age at onset, onset site).

Table 1. ALS Retrospective Dataset. For each subtask, we report the number of patients, the number of ALSFRS-R questionnaires, the number of spirometries performed, and the number of environmental measurements. Since a patient could be eligible for more than one task, we also report the number of unique entries in row "Total Unique".

Dataset		Patients	ALSFRS-R	Spirometry	Environmental
Subtask A	Training	1,432	3,636	1,185	2,549,665
	Test	346	868	273	419,229
Subtask B	Training	1,679	4,214	1,490	3,471,137
	Test	422	1,039	358	769,209
Subtask C	Training	1,716	4,306	1,518	3,566,689
	Test	486	1,210	411	885,407
Total Unique	- - -	2,204	5,519	1,930	4,455,326

We derive three *task-specific datasets* from the ALS retrospective dataset. These datasets focus on predicting when the patient required a NIV, a PEG, or when the patient died, respectively. The features used for prediction include demographic and static clinical data, environmental data, and six months of visit records (i.e., ALSFRS-R and spirometry results). Compared to the original *full dataset*, the *task-specific datasets* are refined to ensure that every patient has at least six months of visits after their first ALSFRS-R is recorded, and that the event being predicted either did not occur or occurred after more than six

Table 2. Comparison between training and test populations of the retrospective ALS *task-specific datasets*. Continuous variables are presented as medians (interquartile range); categorical variables as count (percentage on available data), for each level.

	Subtask A (Outcome target = NIV)		Subtask B (Outcome target = PEG)		Subtask C (Outcome target = Death)	
	Train	Test	Train	Test	Train	Test
Male	733 (51.19%)	185 (53.47%)	901 (53.66%)	236 (55.92%)	908 (52.91%)	268 (55.14%)
Female	699 (48.81%)	161 (46.53%)	778 (46.34%)	186 (44.08%)	808 (47.09%)	218 (44.86%)
Age at onset	64.87 [56.60-71.64]	64.70 [55.46-70.76]	65.04 [56.76-71.82]	64.76 [55.66-70.42]	65.35 [57.20-72.10]	65.01 [56.95-70.86]
Onset Axial	3 (0.21%)	3 (0.87%)	29 (1.73%)	11 (2.60%)	30 (1.75%)	12 (2.57%)
Onset Bulbar	442 (30.86%)	105 (30.35%)	489 (29.12%)	123 (29.15%)	543 (31.64%)	147 (30.25%)
Onset Generalized	4 (0.28%)	0 (0.00%)	7 (0.42%)	1 (0.24%)	7 (0.41%)	1 (0.20%)
Onset Limbs	983 (68.65%)	238 (68.79%)	1154 (68.73%)	287 (68.01%)	1136 (66.20%)	326 (67.08%)
C9orf72 normal	896 (62.57%)	229 (66.19%)	1013 (60.33)	274 (64.93)	1025 (69.74)	311 (63.99)
C9orf72 expansion	72 (5.03%)	23 (6.65%)	76 (4.53)	24 (5.69)	81 (4.71)	28 (5.76)
C9orf72 NA	464 (32.40%)	94 (27.16%)	590 (35.14)	124 (29.38)	610 (35.55)	147 (40.25)
ALSFRS-R slope	0.43 [0.24-0.79]	0.41 [0.23-0.80]	0.47 [0.25-0.84]	0.44 [0.24-0.85]	0.49 [0.26-0.88]	0.45 [0.24-0.85]
ALSFRS-R recorded	3.0 [2.0-3.0]	3.0 [2.0-3.0]	2.0 [2.0-3.0]	2.0 [2.0-3.0]	2.0 [2.0-3.0]	2.0 [2.0-3.0]
Outcome time	18.07 [11.43-31.33]	20.82 [11.47-36.94]	20.43 [12.87-37.12]	22.47 [13.30-38.59]	23.07 [14.16-39.13]	25.20 [14.62-42.47]
Outcome censoring	121 (8.45%)	32 (9.25%)	209 (12.45%)	51 (12.09%)	230 (13.40%)	69 (14.20%)
Outcome death	636 (44.41%)	152 (43.93%)	969 (57.71%)	251 (59.48%)	1486 (86.60%)	417 (85.80%)
Outcome target	675 (47.14%)	162 (46.82%)	501 (29.84%)	120 (28.44%)		

months from the first ALSFRS-R. The size of the retrospective ALS *task-specific datasets* (including the sizes of the training and test sets) is reported in Table 1. Table 2 reports the distribution of a sample of variables available in the ALS retrospective *task-specific datasets*. We report the comparison between training and test variables, showing comparable distributions. In detail, we report the median and the interquartile range for continuous variables (e.g., "age at onset"). On the other hand, we report the count and the percentage over the entire population for categorical data (e.g., "onset location").

Similarly, from the MS retrospective dataset containing only clinical data, we derive two *task-specific datasets*. In this case, the task aims to predict the risk of worsening in the condition, considering two different definitions of worsening. The first definition of "worsening" occurs if the patient's EDSS exceeds the threshold of 3 twice within a year. The second definition considers the change in EDSS from the baseline (the values of the first EDSS recorded after 2.5 years from the first EDSS in absolute). A worsening is defined as an increase in EDSS by 1.5, 1, or 0.5, depending on the value of the baseline EDSS (i.e., ≤ 1, between 1 and 5.5, or >5, respectively). Table 3 presents the sizes of these two MS clinical-only retrospective *task-specific datasets*. Table 4 reports the comparison between training and test populations for a subset of variables available in the clinical-only retrospective *task-specific datasets*. As before, we report the count and percentage for each level of the categorical variables. For continuous variables, on the other hand, we report the mean and the standard deviation.

Table 3. MS Retrospective Dataset with only clinical data. For each subtask, we report the number of patients, the number of relapses, EDSS scores, EPs, MRIs, and MS Course. Since a patient could be eligible for more than one task, we also report the number of unique entries in row "Total Unique".

Dataset		Patients	Relapses	EDSS Scores	EPs	MRIs	MS Course
Subtask A	Training	441	481	2,661	1,211	960	310
	Test	111	95	675	278	236	68
Subtask B	Training	511	553	3,069	1,522	966	325
	Test	129	125	813	299	266	75
Total Unique	- - -	723	803	4,457	2,056	1,439	463

For the MS retrospective dataset that includes both clinical and environmental data, we derive a single *task-specific dataset*. The task is to predict the week of the first relapse, given the patient's status at baseline. The size of this dataset is reported in Table 5, while Table 6 reports the comparison between the training and test populations within the *task-specific dataset*.

Finally, from the prospective ALS *full dataset*, we derive two *task-specific datasets*. The first task involves predicting the ALSFRS-R score assessed by the clinician during the second visit, given the ALSFRS-R score from the first visit, along with demographic and static clinical data, environmental observations, and wearable sensor recordings. The second task focuses on predicting the self-assessed ALSFRS-R score at the second visit, given the first ALSFRS-R score and the same set of features available in the first task. Table 7 details the numerosity of the two prospective ALS *task-specific datasets* while Table 8 reports

Table 4. Training and test populations of the clinical-only MS *task-specific datasets*. Continuous variables are presented as means (standard deviation); categorical variables as count (percentage on available data), for each level.

			Subtask A		Subtask B	
			Train	Test	Train	Test
Static Variables	Sex	Female	305 (69.32%)	76 (69.09%)	355 (69.61%)	85 (66.41%)
		Male	135 (30.68%)	34 (30.91%)	155 (30.39%)	43 (33.59%)
	Residence classification	Cities	120 (27.27%)	32 (29.09%)	152 (29.8%)	37 (28.91%)
		Rural Area	100 (22.73%)	18 (16.36%)	106 (20.78%)	28 (21.88%)
		Towns	208 (47.27%)	54 (49.09%)	236 (46.27%)	56 (43.75%)
		NA	12 (2.73%)	6 (5.45%)	16 (3.14%)	7 (5.47%)
	Ethnicity	Caucasian	424 (96.36%)	99 (90.00%)	491 (96.27%)	122 (95.31%)
		Hispanic	0 (0.00%)	4 (3.64%)	0 (0.00%)	2 (1.56%)
		Black African	0 (0.00%)	2 (1.82%)	0 (0.00%)	3 (2.34%)
		NA	16 (3.64%)	5 (4.55%)	19 (3.73%)	1 (0.78%)
	MS in pediatric age	False	410 (93.18%)	103 (93.64%)	483 (94.71%)	116 (90.62%)
		True	30 (6.82%)	7 (6.36%)	27 (5.29%)	12 (9.38%)
	age at onset	Mean (sd)	31 (9.427)	30 (8.775)	31 (9.816)	31 (10.642)
	Diagnostic delay	Mean (sd)	1029 (1727.8)	967 (1447.6)	1094 (1809.46)	1332 (2092.90)
		NA	12 (2.73%)	1 (0.91%)	9 (1.76%)	5 (3.91%)
	Spinal cord symptoms	False	348 (79.09%)	83 (75.45%)	389 (76.27%)	95 (74.22%)
		True	92 (20.91%)	27 (24.55%)	121 (23.73%)	33 (25.78%)
	Brainstem symptoms	False	305 (69.32%)	79 (71.82%)	367 (71.96%)	85 (66.41%)
		True	135 (30.68%)	31 (28.18%)	143 (28.04%)	43 (33.59%)
	Eye symptoms	False	318 (72.27%)	82 (74.55%)	370 (72.55%)	95 (74.22%)
		True	122 (27.73%)	28 (25.45%)	140 (27.45%)	33 (25.78%)
	Supratentorial Symptom	False	301 (68.41%)	74 (67.27%)	355 (69.61%)	91 (71.09%)
		True	139 (31.59%)	36 (32.73%)	155 (30.39%)	37 (28.91%)
	Other symptoms	Epilepsy	2 (0.45%)	0 (0.00%)	2 (0.39%)	0 (0.00%)
		Sensory	4 (0.91%)	1 (0.91%)	5 (0.98%)	0 (0.00%)
		RM+	3 (0.68%)	2 (1.82%)	5 (0.98%)	2 (1.56%)
		None	431 (97.95%)	107 (97.27%)	498 (97.65%)	126 (98.44%)
	Time since onset	mean (sd)	2524 (2448.3)	2446 (2235.9)	2871 (2775.14)	3773 (3595.14)
MS type	MS type	CIS	99 (32.04%)	18 (26.87%)	108 (33.33%)	22 (29.73%)
		RR	210 (67.96%)	49 (73.13%)	216 (66.67%)	48 (64.86%)
		PR	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.35%)
		SP	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (4.05%)
	delta observation T0	Mean (sd)	-718 (210.2)	-715 (237.6)	-726 (193.54)	-726 (226.50)
EDSS	EDSS as evaluated by clinician	Mean (sd)	2 (0.716)	2 (0.655)	2 (1.2)	3 (1.7)
		NA	37 (1.39%)	3 (0.45%)	39 (1.27%)	7 (0.86%)
	delta EDSS T0	Mean (sd)	-499 (251.6)	-499 (254.4)	-501 (248.58)	-494 (253.84)
Evoked Potentials	Altered potential	Auditory	280 (23.14%)	58 (20.94%)	341 (22.42%)	68 (22.82%)
		Motor	101 (8.35%)	19 (6.86%)	130 (8.55%)	22 (7.38%)
		Somatosensory	482 (39.83%)	111 (40.07%)	625 (41.09%)	130 (43.62%)
		Visual	347 (28.68%)	89 (32.13%)	425 (27.94%)	78 (26.17%)
	Location	Left	311 (25.70%)	73 (26.35%)	379 (24.92%)	73 (24.5%)
		Lower left	126 (10.41%)	29 (10.47%)	167 (10.98%)	37 (12.42%)
		Lower right	136 (11.24%)	31 (11.19%)	177 (11.64%)	36 (12.08%)
		Right	316 (26.12%)	74 (26.71%)	387 (25.44%)	73 (24.5%)
		Upper left	156 (12.89%)	34 (12.27%)	201 (13.21%)	40 (13.42%)
		Upper right	165 (13.64%)	36 (13.00%)	210 (13.81%)	39 (13.09%)
	Delta evoked potential T0	Mean (sd)	-712 (206.3)	-731 (213.3)	-714 (196.78)	-656 (252.93)
Relapses	Delta relapse T0	Mean (sd)	-561 (286.1)	-551 (286.5)	-561 (280.915)	-595 (279.73)
MRI	MRI area	Brain Stem	681 (71.01%)	164 (69.79%)	688 (71.3%)	188 (70.94%)
		Cervical Spinal Cord	62 (6.47%)	25 (10.64%)	67 (6.94%)	15 (5.66%)
		Spinal Cord	201 (20.96%)	36 (15.32%)	191 (19.79%)	57 (21.51%)
		Thoracic Spinal Cord	15 (1.56%)	10 (4.26%)	19 (1.97%)	5 (1.89%)
	Lesions T1	False	175 (18.25%)	45 (19.15%)	155 (16.06%)	37 (13.96%)
		True	149 (15.54%)	29 (12.34%)	164 (16.99%)	56 (21.13%)
		NA	635 (66.21%)	161 (68.51%)	646 (66.94%)	172 (64.91%)
	Lesions T1 gadolinium	False	575 (59.96%)	145 (61.70%)	566 (58.65%)	162 (61.13%)
		True	247 (25.76%)	51 (21.70%)	243 (25.18%)	57 (21.51%)
		NA	137 (14.29%)	39 (16.1%)	156 (16.17%)	46 (17.36%)
	# Lesions T1 gadolinium	Mean (sd)	0 (1.0)	0 (1.0)	0 (1.049)	0 (0.772)
		NA	187 (19.5%)	48 (20.43%)	222 (23.01%)	57 (21.51%)
	New or enlarged lesions T2	False	377 (39.31%)	107 (45.53%)	363 (37.62%)	116 (43.77%)
		True	240 (25.03%)	52 (22.13%)	222 (23.01%)	55 (20.75%)
		NA	342 (35.66%)	76 (32.34%)	383 (39.69%)	94 (35.47%)
	# New or enlarged lesions T2	Mean (sd)	1 (1.486)	1 (1.401)	1 (1.54)	1 (1.32)
	Lesions T2	False	55 (5.74%)	10 (4.26%)	61 (6.32%)	12 (4.53%)
		True	275 (28.68%)	62 (26.38%)	256 (26.53%)	65 (24.53%)
		NA	629 (65.59%)	163 (69.36%)	648 (67.15%)	188 (70.94%)
	Delta MRI T0	Mean (sd)	-512 (282.0)	-534 (275.5)	-526 (280.304)	-525 (280.263)
Outcomes	Outcome occurred	False	367 (83.41%)	93 (84.55%)	384 (75.29%)	97 (75.78%)
		True	73 (16.59%)	17 (15.45%)	126 (24.71%)	31 (24.22%)
	Outcome time	Mean (sd)	5 (4.4)	5 (4.1)	5 (4.396)	5 (4.396)

Table 5. MS Retrospective Dataset with environmental data. For the training and test dataset, we report the number of patients, EDSS scores, and environmental data. We also report the total number of entries in row "Total".

Dataset	Patients	EDSS Scores	Environmental
Training	199	834	113,923
Test	81	290	46,354
Total	280	1,124	160,277

Table 6. Comparison between training and test populations for MS clinical and environmental *task-specific dataset*. Continuous variables are shown as medians (interquartile range); categorical variables as count (percentage on available data).

		Train	Test
Sex	Female	148 (74.37%)	54 (66.67%)
	Male	51 (25.63%)	27 (33.33%)
Ethnicity	Caucasian	181 (90.96%)	77 (95.06%)
	Hispanic	2 (1.00%)	0 (0.00%)
	Black African	2 (1.00%)	0 (0.00%)
	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%)
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 symptoms	False	143 (71.86%)	54 (66.67%)
	True	56 (28.14%)	27 (33.33%)
Brainstem symptoms	False	146 (73.37%)	57 (70.37%)
	True	53 (26.63%)	24 (29.63%)
Eye symptoms	False	148 (74.37%)	59 (72.84%)
	True	51 (25.63%)	22 (27.16%)
Supratentorial symptoms	False	140 (70.35%)	50 (61.73%)
	True	59 (29.65%)	31 (38.27%)
Other symptoms	Sensory	1 (0.50%)	1 (1.23%)
	Epilepsy	1 (0.50%)	0 (0.00%)
	False	197 (99.00%)	80 (98.77%)
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)

some statistics and the comparison between training and test populations for a subset of variables.

Table 7. ALS Prospective Dataset. The two tasks share the same training dataset. Thus, we report the number of patients, the number of ALSFRS-R questionnaires evaluated by clinicians (Column "Clinical ALSFRS-R") or self-assessed by patients (Column "Self ALSFRS-R"), and the number of sensor data rows, for the training and both test datasets. Row "Totale unique" reports the number of unique entries inside the Resource Description Framework (RDF) dataset.

Dataset	Patients	Clinical ALSFRS-R	Self ALSFRS-R	Sensor Data
Training	52	189	301	13,946
Test-Task1	29	58	- - -	6,345
Test-Task2	11	- - -	22	2,347
Total Unique	86	247	323	21,456

Table 8. Comparison between training and test populations of the Prospective ALS *task-specific datasets*. Continuous variables are presented as medians (interquartile range); categorical variables as count (percentage on available data), for each level.

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

For more detailed information on each iDPP@CLEF challenge, the tasks, the expected outcomes, how these were computed, and the available patient attributes for each task, interested readers can refer to the iDPP@CLEF overview papers[8–10] and the challenges' websites (https://brainteaser.dei.unipd.it/challenges/idpp2022/, https://brainteaser.dei.unipd.it/challenges/idpp2023/, https://brainteaser.dei.unipd.it/challenges/idpp2024/).

Data Anonymization

The institutional ethics boards of each medical centre involved in the BRAINTEASER project approved the data collection and the study. Specifically, this study was conducted in compliance with the Declaration of Helsinki, and the BRAINTEASER project was approved in July 2021 by the Ethics Committees of the Lisbon Medical Academic Center, Portugal (Protocol number 162-2021), AOU Città della Salute e della Scienza di Torino, Italy (Protocol number 0079511), and IRCCS Mondino Foundation, Pavia, Italy (Protocol number 20210065554 and 20210080126), and on the 20th of September 2021 by the Ethics Committee of Gregorio Marañon Hospital in Madrid, Spain (Protocol id BRAINTEASER_01).

Although the original data contains sensitive information, the released data has been fully anonymised, and multiple safeguards have been implemented to ensure that individuals cannot be identified, even through indirect identifiers.

The dataset does not contain any direct identifiers or biometric data. It includes information on genetic mutations on relevant genes (e.g., C9orf72), which has been generalised to a binary value indicating the presence or absence of a mutation. In cases where a patient has not been tested, this value is recorded as null.

The original dataset included certain personal quasi-identifiers, such as location, biological sex, ethnic group, and visit dates. To mitigate the risk of re-identification, substantial transformations were applied. Residency location was categorised into three general groups—city, town, and rural areas—using the European Nomenclature of Territorial Units for Statistics (NUTS) classification. Biological sex was encoded as a binary value, and the "ethnic origin" field was generalised into broad categories: "Black African," "Caucasian," and "Hispanic." As for visit and environmental observation dates, these were made relative to a relevant date in the patient's history, which is not disclosed.

The relevant dates, referred to as Time 0, serve as the reference for all other dates:

- The date of the first recorded ALSFRS-R score for retrospective ALS data.
- The first visit following recruitment for prospective ALS data.
- The date of the first EDSS score after the patient received a confirmed diagnosis of MS for clinical-only MS data.
- The date of the first recorded EDSS after January 1, 2013, for clinical and environmental MS data.

Given the sensitive nature of the data, since all subjects are diagnosed with rare diseases (either ALS or MS), and most attributes concern health data, such as symptoms and disease onset, we have implemented multiple measures to minimise the risks associated with potential re-identification. Primarily, we followed the principle of data minimisation, releasing only those attributes that have been scientifically validated as relevant to disease progression. Moreover, the released data pertains to a specific time window before the event that needs to be predicted. We also applied extensive generalisation, with most fields containing either binary values or values from small, controlled dictionaries that represent highly generalised categories.

Finally, any applicant wishing to access the data must sign a Data Usage Agreement (DUA), which ensures that the data will only be used for ethical and legal purposes. The agreement stipulates that users will not attempt to de-anonymise the data and must immediately notify us if any accidental breach occurs.

Data Records

Access to the datasets can be requested through Zenodo[7] (`https://zenodo.org/records/14857741`), following the BRAINTEASER Data Sharing Policy. Figure 3 visually illustrates the structure of the repository. The repository contains two main directories: one for retrospective data and one for prospective data. The retrospective data directory is divided into two subdirectories: "*ALS*" and "*MS*". The "*MS*" retrospective data directory is further split into two subdirectories: "*clinical only*" and "*clinical and environmental*". The prospective data directory contains a single subdirectory, named "*ALS*", which holds the ALS prospective data. Each of these four directories corresponds to a *full dataset* and includes an "*RDF*" subdirectory, which contains the *full dataset* in RDF format and the queries used to generate the *task-specific datasets*. Additionally, each *full dataset* directory includes a "*CSV*" subdirectory, containing the *task-specific datasets* derived from the dataset. Each *task-specific dataset* directory contains two subdirectories, one for training and one for test data. Each *task-specific dataset* directory also includes a `.txt` file that contains the description of all the fields in the associated CSV files.

The BRAINTEASER Ontology

The *Full datasets* are ingested into an RDF graph compliant with the BRAINTEASER Ontology (BTO)[24] (`https://w3id.org/brainteaser/ontology/`). The BTO is an ontology specifically designed to represent clinical data related to ALS and MS in a comprehensive and modular way. It was developed through collaboration between engineers, medical professionals, and domain experts. The ontology unifies data schemas from multiple centres into a single, common structure, ensuring that it meets the diverse requirements of real-world clinical scenarios. This design approach enhances the portability and interoperability of the data. The BTO design is centred around patients and the clinical events that occur throughout each patient's lifecycle. The BTO adheres to several design principles that ensure compliance with Open Biological and Biomedical Ontology Foundry (OBO) (`https://obofoundry.org/principles/fp-000-summary.html`) and the FAIR principles[25] (`https://www.go-fair.org/fair-principles/`), promoting its adoption in diverse contexts. Additionally, BTO incorporates classification schemes that reference abstract concepts in other semantic resources, following the Simple Knowledge Organization System (SKOS) data mode (`https://www.w3.org/TR/skos-reference/`). This approach helps BTO avoid classism[26], reducing the number of required URIs and simplifying query complexity. For in-depth technical details and the validation of BTO, we resort the interested reader to[24].

Technical Validation

The data published here are collected from real-life clinical practice or obtained through wearable sensors, making it intrinsically of high quality. In the remainder of this section, we describe how such data have been validated using Shapes Constraint Language (SHACL) and by the research community when used in practice within the iDPP@CLEF challenges (held from 2022 to 2024).

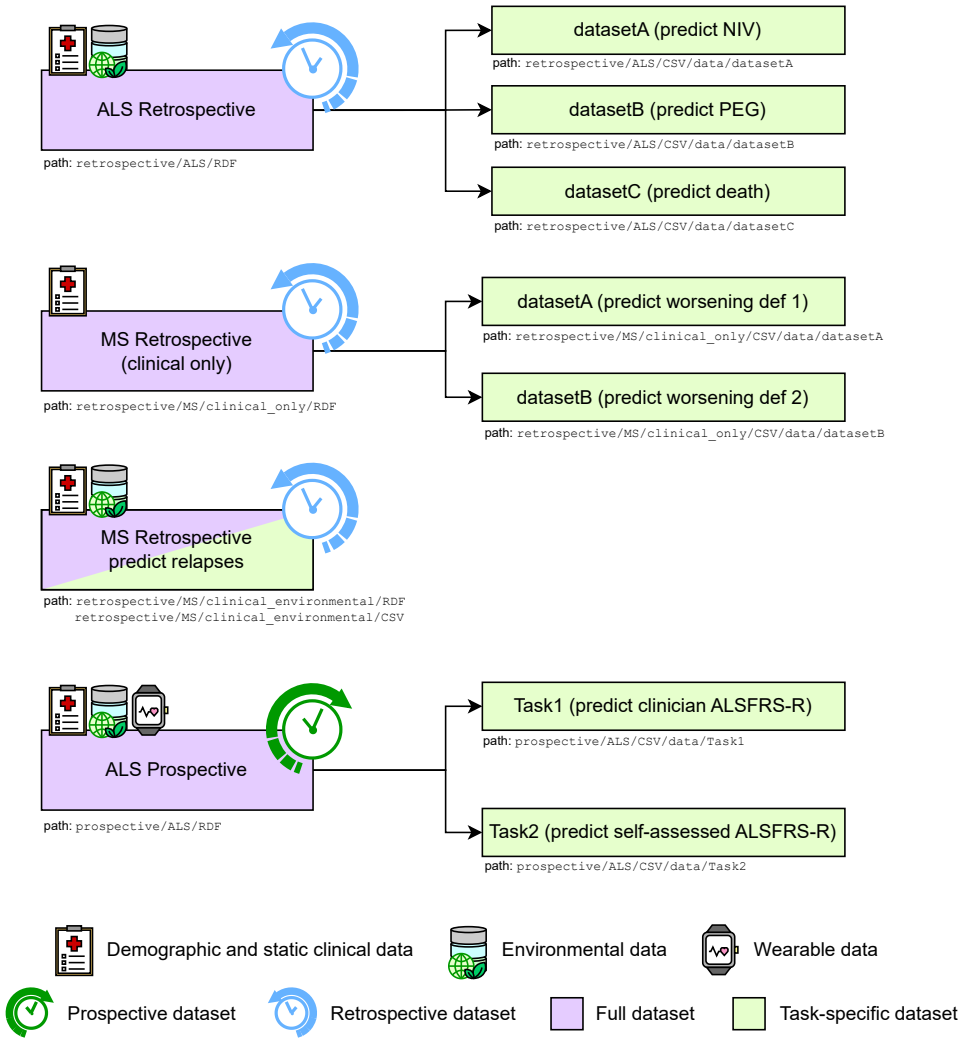


Figure 3. Structure of the Zenodo repository for the BRAINTEASER datasets. From each *full dataset* (in pink) we derive the *task-specific datasets* (in green) corresponding to different tasks of iDPP@CLEF. The only exception is MS clinical and environmental data, where the *full dataset* and *task-specific dataset* coincide. The icons indicate if the data are retrospective (i.e., available before the beginning of the BRAINTEASER project) or prospective (i.e., collected during the BRAINTEASER project) and which data are included in the dataset.

SHACL Validation

To ensure data consistency, the RDF graph resulting from the ingestion of ALS and MS patients from different medical centers, namely the BRAINTEASER Knowledge Base (KB), is validated using SHACL (https://www.w3.org/TR/shacl/). A W3C recommendation language for validating RDF data, SHACL validation is based on *shapes*, graph patterns establishing specific constraints and determining the nodes in a graph that should be evaluated against them. Validating data against a set of constraints is especially important for RDF graphs, as it helps identify issues within a dataset and validates synthactic data quality, facilitating reliable data exchange and interoperability[27]. To this end, we represent each quality condition of the retrospective *full datasets* as a SHACL shape, that is then applied to the BRAINTEASER KB to check for inconsistencies. To provide an example of such an approach, we report the shape to verify all patients are linked to a diagnosis in Listing 1.

```
@prefix sh:    <http://www.w3.org/ns/shacl#> .
@prefix bto:  <https://w3id.org/brainteaser/ontology/schema/> .

bto:Diagnosis4All
    a sh:NodeShape ;
    sh:targetClass bto:Patient ;
    sh:property [
        sh:path bto:undergo ;
        sh:qualifiedValueShape [
            sh:class bto:Diagnosis
        ] ;
        sh:qualifiedMinCount 1 ;
        sh:qualfiedMaxCount 1 ;
    ] ;.
```

Listing 1. SHACL Shape verifying all patients are linked to a diagnosis.

Community Validation of the ALS BRAINTEASER Data

For the ALS track in iDPP@CLEF 2022 and 2023, we used the ALS retrospective *full dataset* split into three *task-specific datasets*.

iDPP@CLEF 2022. The first iteration of iDPP@CLEF focused exclusively on predicting the progression of ALS. The challenge was organized into three tasks (and thus three *task-specific datasets*): i) predicting the patient's need for NIV; ii) predicting the need for PEG; iii) predicting the occurrence of death. Each task was further divided into two subtasks: ranking patients by the risk of event occurrence or predicting the time window in which the event would occur. Participants could submit two types of answers: either using all available data for up to 6 months after the first visit (i.e., the entire dataset), or using all data available up until the first ALSFRS-R questionnaire was recorded.

Five teams from four countries participated in the first edition of iDPP@CLEF: France, Greece, Italy, and Portugal. The first iteration allowed us to identify 48 patients lost to follow-up, with only a single ALSFRS-R recorded, for which the models failed to make predictions. This led to introducing a new filtering criterion: patients lost to follow-up must have at least two ALSFRS-R records. These patients were removed from the datasets released, as they were not used for evaluation in the first iDPP@CLEF.

iDPP@CLEF 2023. In the second iteration of iDPP@CLEF, the three tasks from iDPP@CLEF 2022 were extended by incorporating environmental data. The objective was to rank patients based on the risk of requiring medical treatment (NIV or PEG) or the occurrence of death, while also using environmental data. Participants could submit three types of predictions: i) ignoring environmental data; ii) using up to 6 months of environmental data before and after the first recorded ALSFRS-R; iii) using any arbitrary window of environmental data. Two teams from Portugal and Tunisia participated in the ALS-related track in iDPP@CLEF 2023.

iDPP@CLEF 2024. The final iteration of iDPP@CLEF introduced a new prospective dataset. The task was to predict ALSFRS-R scores, as assessed by clinicians and self-reported by patients enrolled in the BRAINTEASER project.

A total of 7 teams participated in the ALS track of iDPP@CLEF 2024, with teams from institutions in Botswana, Italy, Portugal, Romania, and the United States. Most of the participating teams observed better performance from predictive algorithms when predicting clinician-assessed data. This is likely due to clinician assessments being more consistent and objective, whereas self-reported ALSFRS-R scores tend to be more variable and less precise, as they were provided by non-expert patients directly affected by ALS progression.

Community Validation of the MS BRAINTEASER Data

The MS track was first introduced in iDPP@CLEF 2023 and continued in iDPP@CLEF 2024. There are two MS *full datasets*, one for each edition of the track. Both are retrospective, with one containing environmental data and the other not.

iDPP@CLEF 2023. The second iteration of iDPP@CLEF was the first to feature the MS track. Specifically, iDPP@CLEF 2023 introduced two tasks related to MS, both focused on predicting the risk of disease worsening, either in terms of probability or as a cumulative probability over time. Each task was further divided into two subtasks, with different definitions of worsening based on crossing EDSS threshold values. For each subtask, participants were given a dataset containing 2.5 years of visits, along with the occurrence and time of the worsening event. A total of 9 teams participated in at least one of the two MS tasks in iDPP@CLEF 2023, with 7 teams participating in both tasks and both subtasks. These teams came from 5 different countries: Bulgaria, Czech Republic, Italy, Portugal, and Spain. Overall, the performance on the MS tasks was promising, with participants achieving AUC scores as high as 92.4%.

iDPP@CLEF 2024. The final iteration of iDPP@CLEF focused on predicting relapses using environmental data and EDSS subscores, examining whether exposure to different pollutants was a relevant feature for relapse prediction. To this end, the MS dataset from the previous year was extended by adding environmental data and revised by filtering out EDSS data and patients not aligned with the environmental data. The environmental data included information on patients' exposure to various air pollutants identified as health risks, as well as weather factors such as wind speed, global radiation, and precipitation. For this task, participants were asked to predict the week of the first relapse after the baseline, using weekly-grained environmental data and the patient's status at baseline (the first available visit in the considered time period). Two teams from Italy participated in the task in iDPP@CLEF 2024, underscoring that, while environmental data can improve relapse prediction, better methods for extracting and managing pollution exposure patterns are necessary to effectively leverage such data.

Limitations

There are several aspects to consider when using the BRAINTEASER datasets. First, while some examinations ideally should take place on the same day, in certain cases, they are delayed by a few days in our dataset. This delay is likely due to the unavailability of patients or clinicians, resulting in the examination being postponed. Secondly, there are missing values in the data. In real-life clinical practice, it is sometimes impossible to collect all data, perform certain examinations, or ensure patients attend every consultation. To address this limitation, statistical methods such as imputation techniques should be employed to replace missing data with plausible values. Additionally, as dropout rates and survival times differ among patients, the number of consultations per patient varies, and the time between visits is inconsistent for all patients. Finally, concerning prospective data, sensor recordings may not be available every day. Gaps in the data may occur due to patient non-adherence or technical issues during data collection. Despite these challenges, it is important to note that they reflect the realities of real-world clinical practice. These issues are common in any AI-driven monitoring and progression modelling technique applied to clinical data. Therefore, our datasets provide realistic and practical training and assessment, ensuring that the results are reliable and applicable in real-world conditions, the ultimate target scenario for AI-driven monitoring models.

Usage Notes

Data is publicly available upon completion and return of the Data Usage Agreement. In particular, the data-sharing policy is developed under the General Data Protection Regulation (GDPR, EU Regulation 2016/679), which provides the basis for sharing personal information. To obtain these datasets, the researcher should send a request (The request should be sent to brainteaser-data@dei.unipd.it). for access to the data, together with a detailed and structured study proposal that the BRAINTEASER Project Data Committee will evaluate to understand the use purposes. The proposal will be evaluated to ensure that the objective aligns with the BRAINTEASER Project objectives (i.e., improving the knowledge on ALS and MS and the quality of life of people living with it), and to ensure that the applicant has the means and the interest to protect the privacy of the individuals who contributed to the dataset with their data. After the decision and authorisation, the requesting research group will receive all the information and data. The subsequent passage, following the analysis and the potential results, will be characterized by the revision and validation process made by the BRAINTEASER Project Data Committee. The RDF graphs can be imported into an RDF-compliant Database Management System, such as GraphDB (`https://graphdb.ontotext.com/`).

Code availability

The code and models developed by the participants of the iDPP@CLEF 2022[28], iDPP@CLEF 2023[29], and iDPP@CLEF 2024[30] challenges are publicly available and can be used to replicate the results.

Institutional Ethics Board Approval

The institutional ethics boards of each medical centre involved in the BRAINTEASER project approved the data collection and the study. Specifically, this study was conducted in compliance with the Declaration of Helsinki, and the BRAINTEASER project was approved in July 2021 by the Ethics Committees of the Lisbon Medical Academic Center, Portugal (Protocol number 162-2021), AOU Città della Salute e della Scienza di Torino, Italy (Protocol number 0079511), and IRCCS Mondino Foundation, Pavia, Italy (Protocol number 20210065554 and 20210080126), and on the 20th of September 2021 by the Ethics Committee of Gregorio Marañon Hospital in Madrid, Spain (Protocol id BRAINTEASER_01).

References

1. Grossman, M. Amyotrophic lateral sclerosis - a multisystem neurodegenerative disorder. *Nat. Rev. Neurol.* **15**, 5–6, 10.1038/s41582-018-0103-y (2019).
2. Jakimovski, D. *et al.* Multiple sclerosis. *Lancet* **403**, 183–202, 10.1016/S0140-6736(23)01473-3 (2024).
3. Marrie, R. A., Lancia, S., Cutter, G. R., Fox, R. J. & Salter, A. Access to care and health-related quality of life in multiple sclerosis. *Neurol. Clin. Pract.* **14**, e200338, 10.1212/CPJ.0000000000200338 (2024).
4. van Es, M. A. *et al.* Amyotrophic lateral sclerosis. *The Lancet* **390**, 2084–2098, https://doi.org/10.1016/S0140-6736(17)31287-4 (2017).
5. Tavazzi, E. *et al.* Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review. *Artif. Intell. Medicine* **142**, 102588, https://doi.org/10.1016/j.artmed.2023.102588 (2023).
6. BRAINTEASER Consortium. Bringing artificial intelligence home for a better care of amyotrophic lateral sclerosis and multiple sclerosis (brainteaser). https://cordis.europa.eu/project/id/101017598 (2021).
7. Faggioli, G. *et al.* Brainteaser als and ms datasets, 10.5281/zenodo.14857741 (2025).
8. Guazzo, A. *et al.* Intelligent disease progression prediction: Overview of idpp@clef 2022. In Barrón-Cedeño, A. *et al.* (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, vol. 13390 of *Lecture Notes in Computer Science*, 395–422, 10.1007/978-3-031-13643-6_25 (Springer, 2022).
9. Faggioli, G. *et al.* Intelligent disease progression prediction: Overview of idpp@clef 2023. In Arampatzis, A. *et al.* (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, vol. 14163 of *Lecture Notes in Computer Science*, 343–369, 10.1007/978-3-031-42448-9_24 (Springer, 2023).
10. Birolo, G. *et al.* Intelligent disease progression prediction: Overview of idpp@clef 2024. In Goeuriot, L. *et al.* (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II*, vol. 14959 of *Lecture Notes in Computer Science*, 118–139, 10.1007/978-3-031-71908-0_6 (Springer, 2024).
11. Atassi, N. *et al.* The pro-act database. *Neurology* **83**, 1719–1725, 10.1212/WNL.0000000000000951 (2014).
12. Baxi, E. G. *et al.* Answer als, a large-scale resource for sporadic and familial als combining clinical and multi-omics data from induced pluripotent cell lines. *Nat. Neurosci.* **25**, 226–237, 10.1038/s41593-021-01006-0 (2022).
13. Muslim, A. M. *et al.* Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. *Data Brief* **42**, 108139, 10.1016/j.dib.2022.108139 (2022).
14. Commowick, O. *et al.* Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset. *NeuroImage* **244**, 118589, 10.1016/j.neuroimage.2021.118589 (2021).
15. Commowick, O., Cervenansky, F., Cotton, F. & Dojat, M. Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In *MICCAI 2021-24th international conference on medical image computing and computer assisted intervention*, 126 (2021).
16. Training Data. The Brain MRI Image DICOM Dataset, https://trainingdata.pro/datasets/brain-mri?utm_source=kaggle&utm_medium=cpc&utm_campaign=multiple-sclerosis (2023).
17. Marrie, R. A. *et al.* Narcoms and other registries in multiple sclerosis: Issues and insights. *Int. J. MS Care* **23**, 276–284, 10.7224/1537-2073.2020-133 (2021).

18. Chiò, A. *et al.* Secular trends of amyotrophic lateral sclerosis: The piemonte and valle d'aosta register. *JAMA Neurol.* **74**, 1097–1104, 10.1001/jamaneurol.2017.1387 (2017).

19. Benjamin Rix Brooks, M. S., Robert G Miller & Munsat, T. L. El escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Other Mot. Neuron Disord.* **1**, 293–299, 10.1080/146608200300079536 (2000). PMID: 11464847.

20. World Health Organization. *WHO global air quality guidelines: Particulate matter (PM(2.5) and PM(10)), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide* (World Health Organization, Geneva, 2021). Review.

21. Gonzalez-Martinez, S. *et al.* Novel interactive brainteaser tools for amyotrophic lateral sclerosis (als) and multiple sclerosis (ms) management. In *Participative Urban Health and Healthy Aging in the Age of AI: 19th International Conference, ICOST 2022, Paris, France, June 27–30, 2022, Proceedings*, 302–310, 10.1007/978-3-031-09593-1_26 (Springer-Verlag, Berlin, Heidelberg, 2022).

22. Cossu, L., Cappon, G. & Facchinetti, A. Adaptive and self-learning bayesian filtering algorithm to statistically characterize and improve signal-to-noise ratio of heart-rate data in wearable devices. *J. Royal Soc. Interface* **21**, 10.1098/rsif.2024.0222 (2024).

23. Cossu, L., Cappon, G. & Facchinetti, A. Automated pipeline for denoising, missing data processing, and feature extraction for signals acquired via wearable devices in multiple sclerosis and amyotrophic lateral sclerosis applications. *Front. Digit. Heal.* **6**, 10.3389/fdgth.2024.1402943 (2024).

24. Faggioli, G. *et al.* An extensible and unifying approach to retrospective clinical data modeling: the BrainTeaser Ontology. *J. Biomed. Semant.* **15**, 16, 10.1186/S13326-024-00317-Y (2024).

25. Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, 10.1038/sdata.2016.18 (2016).

26. Allemang, D., Hendler, J. & Gandon, F. Good and bad modeling practices. In *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL*, 436–440, 10.1145/3382097.3382113 (Association for Computing Machinery, New York, NY, USA, 2020).

27. Pareti, P. & Konstantinidis, G. A Review of SHACL: From Data Validation to Schema Reasoning for RDF Graphs. In *Reasoning Web. Declarative Artificial Intelligence : 17th International Summer School 2021, Leuven, Belgium, September 8–15, 2021, Tutorial Lectures*, 115–144, 10.1007/978-3-030-95481-9_6 (Springer International Publishing, Cham, 2022).

28. Ferro. iDPP@CLEF 2022 - Participants' repositories for the Intelligent Disease Prediction Progression Challenge. Zenodo, 10.5281/zenodo.7477919 (2022).

29. Faggioli, G. *et al.* iDPP@CLEF 2023 - Participants' repositories for the Intelligent Disease Prediction Progression Challenge. Zenodo, 10.5281/zenodo.10210125 (2023).

30. Birolo, G. *et al.* iDPP@CLEF 2024 - Participants' repositories for the Intelligent Disease Prediction Progression Challenge. Zenodo, 10.5281/zenodo.14030410 (2024).

Acknowledgements

This work was supported by the BRAINTEASER Project, as part of the European Union Horizon 2020 Research and Innovation Program under Grant Agreement no. GA101017598.

Author contributions statement

G.F., L.M., and S.M. implemented the data filtering pipeline and wrote the paper, under the supervision of G.M.D.N., G.S., and N.F.. I.T. operated additional checks on the data quality and the split into training and test. L.A., A.A., R.B., P.C., A.C., A.D., M.D.C., P.F., J.M.G.D., S.G.M., M.G., A.G., A.J., B.K., E.L., S.C.M., U.M., J.L.M.B., E.T., E.T., E.T.I., V.U., M.V. and B.D.C. operated on different steps of the data collection, processing and filtering, and contributed to the organisation of the challenges to validate the data. All the authors reviewed the manuscript.

Competing interests

The authors declare no competing interests as defined by Nature, or other interests that might be perceived to influence the results and/or discussion reported in this paper.