Brainteaser

D9.4 Shared data package for the evaluation challenge and integration with EOSC

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_D9.4_Shared data package for the evaluation challenge and integration with EOSC
Deliverable leader	UNIPD
Due date	30/06/22
Deliverable date	22/07/22
Dissemination level	PUBLIC

AUTHORS – CONTRIBUTORS

Name	Organization
Alessandro Guazzo	University of Padua, Italy
Isotta Trescato	University of Padua, Italy
Enrico Longato	University of Padua, Italy
Enidia Hazizaj	University of Padua, Italy
Dennis Dosso	University of Padua, Italy
Guglielmo Faggioli	University of Padua, Italy
Giorgio Maria Di Nunzio	University of Padua, Italy
Gianmaria Silvello	University of Padua, Italy
Martina Vettoretti	University of Padua, Italy
Erica Tavazzi	University of Padua, Italy
Chiara Roversi	University of Padua, Italy
Piero Fariselli	University of Turin, Italy
Sara C. Madeira	University of Lisbon, Portugal
Mamede de Carvalho	University of Lisbon, Portugal
Marta Gromicho	University of Lisbon, Portugal
Adriano Chio'	University of Turin, Italy
Umberto Manera	University of Turin, Italy
Arianna Dagliati	University of Pavia, Italy
Giovanni Birolo	University of Turin, Italy
Helena Aidos	University of Lisbon, Portugal
Barbara Di Camillo	University of Padua, Italy
Nicola Ferro	University of Padua, Italy

PEER – REVIEWERS

Name	Organization
Vladimir Urošević, Ognjen Milićević	BELIT, Serbia

DOCUMENT HISTORY

Version	Date	Author/Organization	Modifications	Status
0.1	02/05/2022	UNIPD	Initial outline	Draft
0.2	17/05/2022	UNIPD	Definition of the filtering processing	Draft
0.3	24/05/2022	UNIPD	Description of the results of the filtering for each center's data	Draft
0.4	30/05/2022	UNIPD	Preparation of the SPARQL queries	Draft
0.5	05/06/2022	UNIPD	Validation and description of the SPARQL queries	Draft
0.6	10/06/2022	UNIPD	Description of the splitting of the dataset into subdatasets	Draft

Version	Date	Author/Organization	Modifications	Status
0.7	20/06/2022	UNIPD	Description of the splitting of the dataset into training and test sets	Draft
0.8	25/06/2022	UNIPD	Introduction	Draft
0.9	28/06/2022	UNIPD	Conclusions, and executive summary	Draft
1.0	30/06/2022	UNIPD	First complete version of the deliverable submitted for internal review	Draft
1.1	21/07/2022	UNIPD	Final version of the deliverable, updated according to the internal review performed by BELIT	Draft
2.0	22/07/2022	UPM	Final review and final version	Final

Disclaimer

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
ESCO	European Skills, Competences, Qualifications and Occupations
FVC	Forced Vital Capacity
iDPP	intelligent Disease Progression Prediction
NIV	Non-Invasive Ventilation
PEG	Percutaneous Endoscopic Gastrostomy

EXECUTIVE SUMMARY

The main goal of this deliverable is to present the steps we undertook to ingest, process, analyse and prepare the training and testing datasets for the Intelligent Disease Progression Prediction (iDPP) lab at CLEF 2022.

The lab was organized into three tasks: ranking of patients based on the risk of impairment (task A); predict when specific impairments will occur (i.e., in the correct time-window) (task B); and discussion on how to make these prediction algorithms explainable, also in a visual way (task C).

Hence, this deliverable presents the data pipeline employed to ingest and process the raw hospital Amyotrophic Lateral Sclerosis (ASL) data provided by the IMM center, UNITO, and Sermas. We describe the medical data provided by the partners and highlight their heterogeneity and the challenges we tackled to harmonize the data to prepare the three datasets for the training and testing of predictive algorithms. We present the SPARQL queries to retrieve the relevant data for creating the challenge datasets. We explain how the challenge datasets were created and how we split the training, validation, and test datasets for each task of the challenge. We report detailed descriptive statistics and further insights, also supported by visual aids, into the distribution of data and what can be learned from the provided datasets. Finally, we report the procedure followed to produce state-of-the-art training and testing datasets for ASL-related prediction tasks. The datasets comprise, depending on the task, from 1804 to 2250 records with detailed information about the clinical history of the patients. These datasets are crucial resources to be employed in explainable AI and to analyse/predict the temporal evolution of ALS.

1 INTRODUCTION

ALS is a neurological disease that causes the progressive degeneration of the motor neurons that control voluntary muscles, resulting in an increasing impairment of motor and vital functions and leading to death usually within 4-5 years from the diagnosis.

Likely resulting from a complex interplay of genetic and environmental factors, ALS is characterized by high heterogeneity in both symptoms and disease progression, especially in the early stages of the disease.

This heterogeneity is partly responsible for the lack of effective prognostic tools in medical practice, as well as for the current absence of a therapy able to effectively slow down or reverse the disease course.

On the one hand, patients and caregivers need support for facing the psychological and economic burdens deriving from the uncertainty of how the disease will progress; on the other, clinicians require tools that may assist them throughout the patient's care, recommending tailored therapeutic decisions and providing alerts for urgently needed actions.

To improve the current diagnostic and prognostic situation, we need to design and develop AI algorithms able to: stratify patients according to their phenotype, assessed throughout the disease evolution; predict the progression of the disease in a probabilistic, time dependent fashion; better describe disease mechanisms. The Intelligent Disease Progression Prediction (iDPP) [GTL+22,GTL+22b] lab, organized in the context of the CLEF 2022 Conference on Labs of the Evaluation Forum [FFH+22], aims to design and develop an evaluation infrastructure for driving the development of such AI algorithms. By "evaluation infrastructure", we mean experimental collections, evaluation protocols, evaluation measures, ground-truth creation protocols, and so on.

In this this context, it is fundamental, even if not so common yet, to develop shared approaches, promote the use of common benchmarks, and foster the comparability and replicability of the experiments. Differently from previous challenges in the field, iDPP addresses in a systematic way some issues related to the application of AI in clinical practice in ALS. Therefore, in addition to defining the risk scores based on the probability that an event will occur in the short- or long-term period, iDPP also addresses the issue of providing information in a more structured and understandable way to clinicians.

iDPP 2022 is the first edition of the lab and consists of pilot activities aimed both at an initial exploration of ALS progression prediction and at understanding of the challenges and limitations to refine and tune the labs itself for future iterations.

This deliverable describes the data pipeline employed to ingest and process the raw medical data about ALS provided by the BRAINTEASER partners. Moreover, we report about the creation of the training, validation and testing datasets created for the iDPP lab.

In the "Prefiltering" Section we present the descriptive statistics about the raw medical data, and we explain how we selected the medical records to be used in the challenge and what records we needed to discard and why (e.g., missing specific data points, as specified in detail in the next Chapter).

The “SPARQL queries” Section presents the queries we used to retrieve the medical data from the BRAINTEASER RDF database. This section describes in detail what can be retrieved from the database, explains the access paths to the data and enables the reproduction of the challenge dataset starting from the data modelled according to the BRAINTEASER Ontology.

The “Challenge Dataset creation” Section presents the datasets created for the iDPP 2022 lab at CLEF and reports about the main characteristics of the datasets and how they can be employed for training and testing predictive algorithm for ASL.

The ALS Challenge Datasets are publicly available as open data (URL) to enable data sharing, re-use, and citation.

Finally, in the “Conclusion” Section we draw some conclusions.

2 PREFILTERING

The complete original retrospective ALS dataset contained approximately 4800 records linked to patients, with around 22000 ALSFRS-R [CSM+99] questionnaires in total and 5500 records concerning spirometry visits. The original data contain minor inconsistencies and typos. Therefore, we have firstly processed the data, removing records that are likely wrong or do not provide essential information to enable prediction. In terms of patient records, we removed those presenting an unordered sequence of events (i.e., onset after diagnosis or diagnosis after death). Such event sequences are likely due to typos and other human errors, which result in wrong records that might introduce noise and spurious information in the final dataset.

Furthermore, a patient record was dropped if one or more of the following pieces of information were absent: onset or diagnosis dates; death date in records associated with dead patients; at least six months between the ALSFRS-R questionnaire taken by the patient and the clinical event that needs to be predicted (NIV, PEG, (competing event) Death, or a censoring event) - this follows previous literature that highlight the importance of seed information to produce a sensible prediction [KZN+15]. Concerning the ALSFRS-R questionnaires, we removed those records that had one or more of the following problems: duplicate records; missing date; one or more of the ALSFRS-R item sub-scores missing; ALSFRS in their pre-revision version (items 10, 11 and 12, all these scores were originally associated to a single item).

The ALSFRS-R sub-scores and the total ALSFRS-R have been recalculated for verification.

Finally, regarding the spirometry visits, we removed duplicated records, records with a missing date, and records with missing FVC percentage value.

2.1 Analysis of the Unusable Records

We report here the results of our record filtering process. We describe, for each of the entities considered (i.e., patient records/profiles, ALSFRS-R questionnaires, and spirometry visits) the number of dropped records, the reason why we dropped them and the number of retained records which constitute the final dataset used for the iDPP@CLEF 2022 lab. Notice that the procedure has been executed sequentially - at each step, we remove from further filtering steps those records that do not satisfy the given condition, regardless of whether they satisfy subsequent conditions. Regarding the clinical events to be predicted, namely *NIV*, *PEG,* and *tracheostomy*, we remove records for which even one of the events either happened before the first visit (the date of the first ALSFRS-R questionnaire) or after the death. We drop the record even though the problem is linked to only one of the events (e.g., NIV), while the others (e.g., PEG) are correct. This is because a wrongful event likely indicates a corrupt record: since we do not know the kind of the problem (i.e., if it is related to the date of the event, the date of the first visit or the date of death), we remove the record entirely to avoid introducing noise into the dataset. The event *tracheostomy* is replaced after the filtering phase with the event *death*. In these cases, the later “death” event is removed to avoid patients presenting two death events.

In fact, tracheostomy in some countries is considered as therapeutic persistence and not allowed -furthermore, the tracheostomy event is considered so severe that can be, from a disease progression standpoint, be considered equivalent to death.

2.2 IMM Data

Concerning the 1562 records linked to patients followed at the IMM centre, we observed that 25 of them did not report the date of onset, while 7 did not report the date of diagnosis. Since both these pieces of information are vital to providing sensible predictions on the course of the disease, we remove such records from the final dataset. Similarly, we removed 153 patients from the challenge dataset who did not have at least one ALSFRS-R questionnaire. Finally, we have 672 patients with invalid events according to the previously defined criteria that need to be removed from the challenge dataset. In total, 705 out of 1562 patients' records have been considered suitable for the challenge, corresponding to a retention rate of 45.1%. Figure 1 provides a visual overview of the number of patients excluded and the reasons for the IMM data.

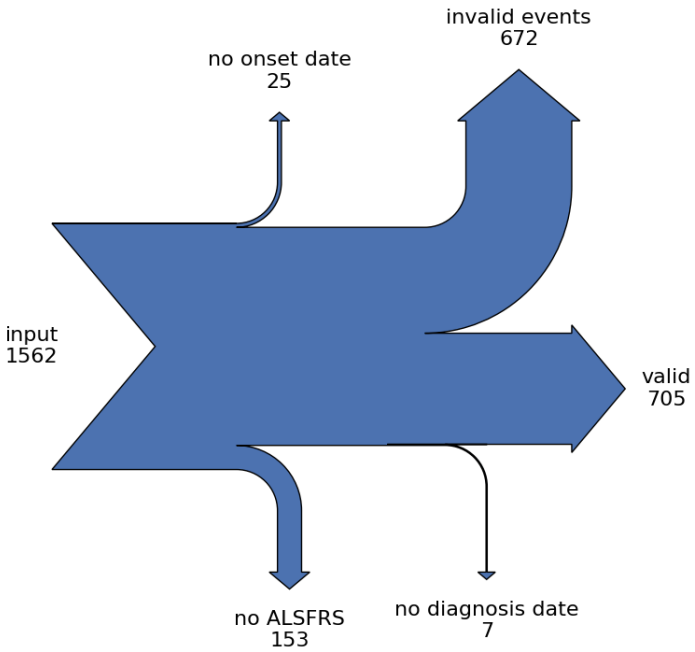

Figure 1 Number of dropped patients' records and associated reason, for the IMM research centre.

Concerning the ALSFRS-R records, among the 7446 questionnaires provided by the IMM centre, 396 did not contain the value for all the 12 individual scores. We treat equally records for which we do not have the date, those for which we only have the old item 10 (missing items revised 10, 11 and 12) and those for which we do not have the value for an arbitrary item. We also dropped 2352 records associated with patients that have not been included in the dataset due to previously described problems. The final version of the dataset contains 4598 records (retention rate of 61.8%). Figure 2 provides a visual overview of the number of ALSFRS questionnaires excluded and the reasons for the IMM data.

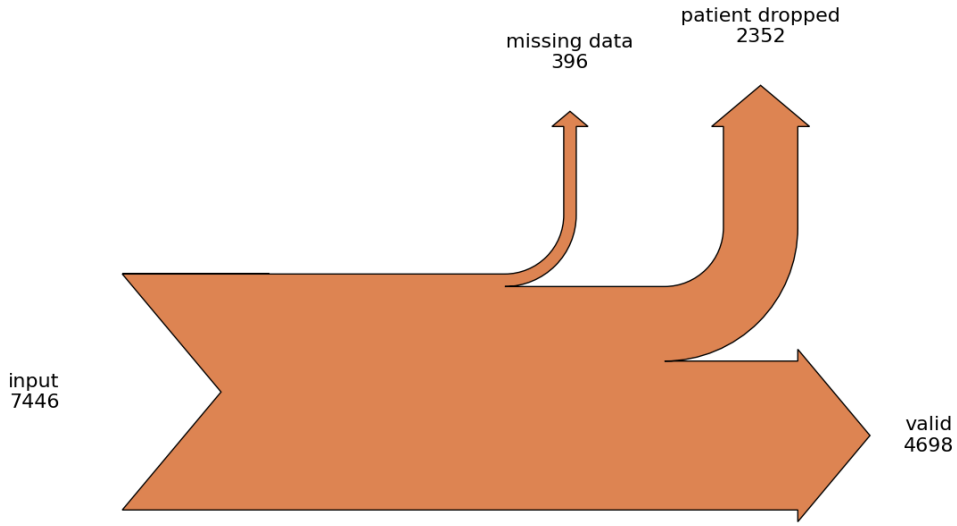

Figure 2 Number of dropped ALSFRS-R records and associated reason, for the IMM research centre.

Similarly to what was done for the ALSFRS-R questionnaires, the spirometry visits records provided by the IMM centre have also been filtered to retain only valid records. We removed 51 records for which we do not have the FVC percentage, and 928 records linked to previously removed patients. In total, 1652 spirometries were included in the final dataset, with a retention rate of 63.8%. Figure 3 provides a visual overview of the number of spirometry visits questionnaires excluded and the reasons for the IMM data.

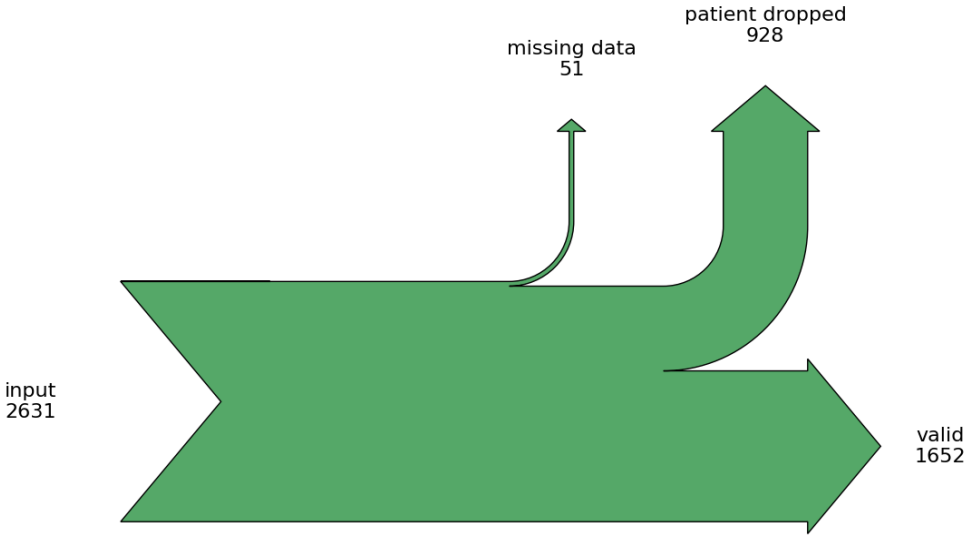

Figure 3 Number of dropped spirometry exam records and reason, for the IMM research centre.

2.3 UNITO Data

The UNITO centre provided records concerning 3257 patients, with 15006 ALSFRS-R questionnaires associated and 2890 spirometry records. Figure 4 reports the result of the filtering of the patients' records according to the criteria described in the previous section. As seen on the Figure, two out of 3257 patients were removed due to the discording reported dates of diagnosis and onset (i.e., the diagnosis is reported to have happened on a date before the onset). This indicates an inconsistency in the data and thus requires us

to drop the records. Similarly, we dropped 31 records because the date of death was reported to have happened before the last visit. Concerning the events, we removed 110 records linked to patients for which we have that the event either happened before the first visit or after the death, indicating a spurious record. Finally, we dropped 1260 patients for whom we do not have at least one valid ALSFRS-R questionnaire. We retain 1854 records linked to the patients, with a retention rate of 56.9%.

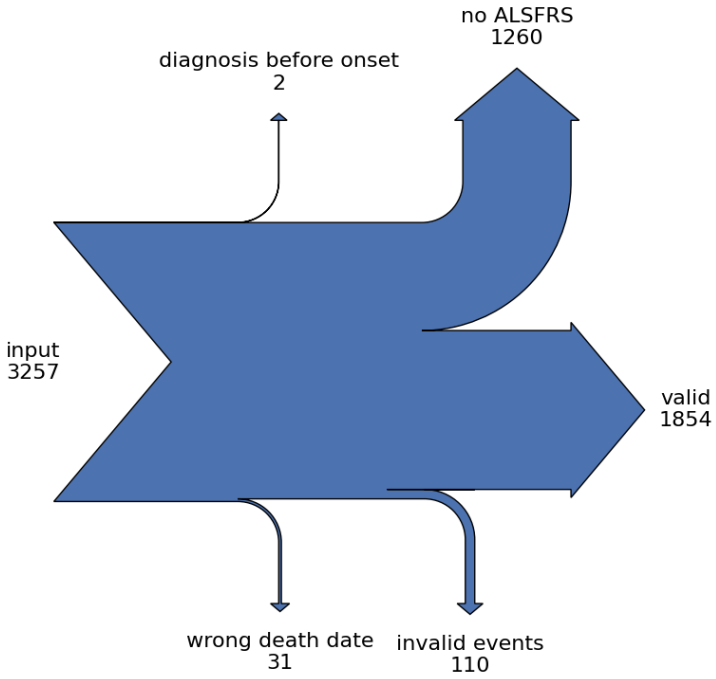

Figure 4 Number of dropped patients' records and associated reason, for the UNITO research centre.

Figure 5 illustrates the retention concerning the ALSFRS-R questionnaires. Out of the 15006 questionnaires provided by the UNITO centre, 27 were duplicated and thus dropped. Furthermore, 566 questionnaires were linked to patients that had been dropped in the previous phase. No ALSFRS-R record presented missing data. After this phase, we retained 14413 records, with a retention rate of 96.0%.

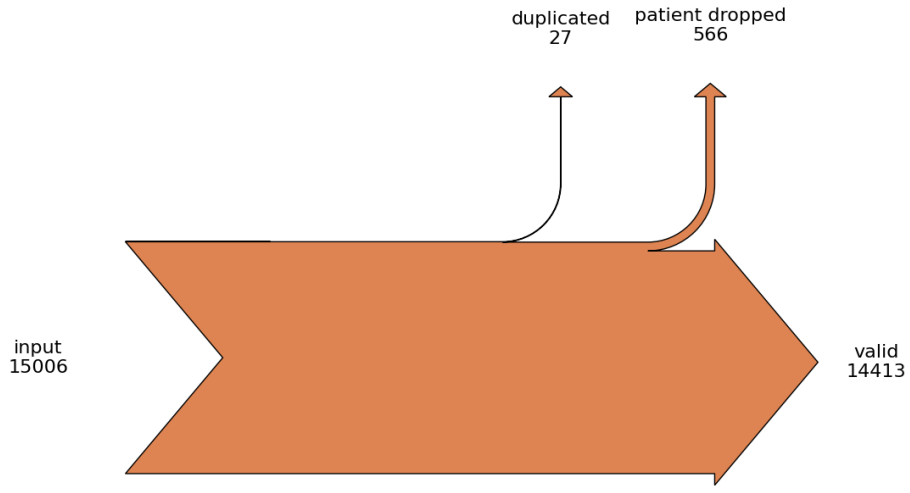

Figure 5 Number of dropped ALSFRS-R records and associated reason, for the UNITO research centre.

Finally, regarding the spirometry exams, Figure 6 reports the number of dropped records together with the reason why they were deleted. We started with 2890 records, of which 17 were duplicates, while 359 concerned previously dropped patients. In conclusion, we have 2514 valid records concerning the spirometry exams – they correspond to a retention rate of 87.0%.

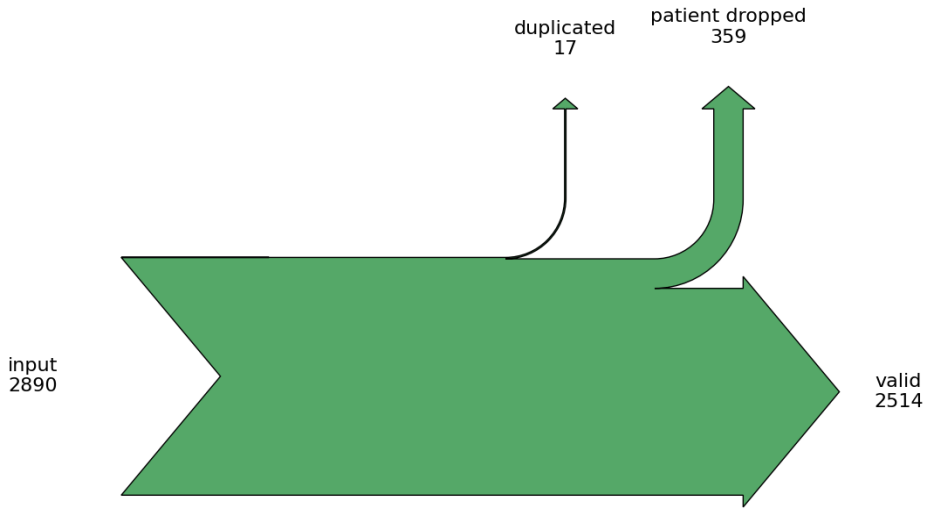

Figure 6 Number of dropped spirometry exam records and reason, for the UNITO research centre.

2.4 SERMAS Data

Regardless of the high value of the dataset provided by the Servicio Madrileño de Salud (SERMAS) study centre, we could not include it in the dataset used for the iDPP@CLEF 2022 challenge.

SERMAS data does not contain individual items for the ALSFRS questionnaire, only the aggregated score. This presents two main challenges that could not be addressed in the current phase and will be addressed in future editions of the iDPP lab.

Our preliminary considerations concerned the fact that it is very likely that, given the type of data provided, most of the AI approaches would focus on the ALSFRS scores. Their absence is, therefore, a considerable limitation that will impair the AI models. Furthermore, records without the proper ALSFRS scores could not be treated equally to the others in the evaluation phases.

Since only SERMAS data do not have ALSFRS individual items scores, adding such records to the dataset would have represented a considerable risk in terms of privacy protection. A potential malicious user would have used the absence (or presence) of the ALSFRS individual items scores to recognize the centre where the record was produced. This would have, in turn, enabled dangerous linking attacks that could have exposed the patients, putting at risk their privacy and safety.

We plan to include these data in future editions of the iDPP lab by addressing the abovementioned limitations. First, we also plan to consider partial data to challenge machine learning approaches to learning with noisy or incomplete information. Secondly,

we plan to adopt the proper anonymization strategies (i.e., *k-anonymity* via cell suppression) to homogenize SERMAS data to those provided by other centres.

3 SPARQL QUERIES

This section contains the queries used to construct the overall dataset that later is further processed to extract the single datasets for each task. Notice that the queries are based on the ontology defined by [BDD+21].

3.1 Prefixes

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX BTO_schema: <https://w3id.org/brainteaser/ontology/schema/>
PREFIX NCIT: <http://purl.obolibrary.org/obo/NCIT_>
PREFIX BTO_ni: <https://w3id.org/brainteaser/ontology/named-individual/>
PREFIX MAXO: <http://purl.obolibrary.org/obo/MAXO_>
PREFIX OGG: <http://purl.obolibrary.org/obo/OGG_>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

3.2 Personal Data, Anamnesis and Statistical Variables

3.2.1 Personal Data

The subsequent query is used to extract personal details concerning patients from their records. Seven pieces of information are taken for each patient:

- **sex**: string - the biological sex of the patient.
- **ethnicity**: string - the ethnicity of the patient, as described in the SNOMEDCT ontology. Notice that, by being the variable highly coarse (Asian - ethnic group, Black – ethnic group, Caucasian) and thus represents a minimal privacy risk.
- **height**: float - the height of the patient.
- **weight**: float - the weight of the patient at the first visit.
- **alive**: binary value (0, 1) - if the patient is alive.
- **occupation**: string - the ESCO occupation of the patients.

```
SELECT ?id ?sex ?ethnicity ?height ?weight ?alive ?occupation  WHERE {
  ?idURI a NCIT:C16960 . bind( substr( (str(?idURI)), 48) as ?id)
  # personal info
  optional{ ?idURI BTO_schema:sex ?sex . }
  optional{ ?idURI BTO_schema:alive ?alive . }
  optional{ ?idURI BTO_schema:ethnicity  ?ethnicityURI . }
  optional{ ?idURI BTO_schema:hasOccupation ?occupationURI . }

   optional{ ?idURI BTO_schema:undergo ?eventCA .
     ?eventCA a NCIT:C39564 ;
        BTO_schema:consists ?CAfirstvisit .
     ?CAfirstvisit a MAXO:0000487 ;
        BTO_schema:Height ?height . }
  optional{ ?idURI BTO_schema:undergo ?eventCA .
     ?eventCA a NCIT:C39564 ;
```

```
        BTO_schema:consists ?CAfirstvisit .
     ?CAfirstvisit a MAXO:0000487 ;
        BTO_schema:Weight ?weight . }

  bind( substr( (str(?occupationURI)), 56) as ?occupation)
  bind( substr( (str(?ethnicityURI)), 56) as ?ethnicity)
}ORDER BY ?id
```

3.2.2 Onset And Diagnosis

The subsequent query allows extracting information from our knowledge base concerning the onset event and the diagnosis. We are interested in retrieving pieces of information about the onset's characteristics, the patient's state at that moment, and events that happened before that. Regarding the diagnosis, we consider only the "working status" at that time.

The variables that we extract with the subsequent query are the following:

- **onsetDate**: string reporting the date of the onset. We expect to have this date for all the patients.
- **age_onset**: float indicating the age of the patient (in years) when the first symptoms occurred.
- **prevalentLMN**: binary variable describing whether the onset mainly affected the lower motor neurons.
- **prevalentUMN**: binary variable describing whether the onset affected mostly the upper motor neurons.
- **mixedMN**: binary variable describing whether the onset affected both lower and upper motor neurons.
- **onset_bulbar**: binary variable indicating if the onset was of type *bulbar*.
- **onset_axial**: binary variable indicating if the onset was of type *axial*.
- **onset_generalized**: binary variable indicating if the onset was *generalized*.
- **onset_limbs**: binary variable indicating if the onset was on the limbs.
- **onset_limb_type**: variable indicating the type of onset when the onset is localized on the limbs. In encloses the information concerning whether the onset is localized on the upper or lower part of the body, if it involves left of right limbs, and if it is proximal or distal.
- **weight_before_onset**: float variable indicating the weight before the onset.
- **moreThan10PercentWeightloss**: binary variable that indicates whether the patient has lost more than 10% of their weight after the onset. Notice that this variable is set only if the weight before onset and the patient's weight are not null.
- **major_trauma_before_onset**: binary variable describing whether the patient has undergone major traumas before the onset. We later describe the query used to extract further details about such traumas (and possible surgical interventions) when such pieces of information are available.
- **surgical_intervention_before_onset**: binary variable describing whether the patient underwent surgical interventions before the disease onset.
- **diagnosisDate**: string containing the date of the diagnosis.
- **retired_at_diagnosis**: binary variable illustrating whether the patient was retired at diagnosis.
- **ALS_familiar_history**: string used to describe whether the patient has a familiar history of ALS. Notice that, at the current time, we treat it as a binary variable,

simply saying whether the patient has ALS familiar history. We plan to extend it with a more precise description of the family relation types and levels.

Notice that, even though we do not extract further information concerning the diagnosis with the subsequent query, other anamnestic details such as the blood test results, and smoking status refer to when the diagnosis was carried out.

```
SELECT ?id ?onsetDate ?age_onset ?prevalentLMN ?prevalentUMN ?mixedMN
?onset_bulbar ?onset_axial ?onset_generalized ?onset_limbs ?onset_limb_type
?weight_before_onset ?moreThan10PercentWeightloss ?major_trauma_before_onset
?surgical_interventions_before_onset ?diagnosisDate ?retired_at_diagnosis
?ALS_familiar_history WHERE {
  ?idURI a NCIT:C16960 . bind( substr( (str(?idURI)), 48) as ?id)
  optional{ ?idURI BTO_schema:hasRelative ?relativeURI . }
  optional{ ?idURI BTO_schema:retiredAtDiagnosis ?retired_at_diagnosis . }
  bind( substr( (str(?relativeURI)), 56) as ?ALS_familiar_history)
  # onset
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:startDate ?onsetDate . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:age_onset ?age_onset . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:axial ?onset_axial . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:bulbar ?onset_bulbar . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:generalized ?onset_generalized . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:limbs ?onset_limbs . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:site ?limbsite . }
  optional{ ?idURI BTO_schema:undergo ?eventOnsetURI . ?eventOnsetURI a
NCIT:C25279 ;
     BTO_schema:consists ?CAURI .
     ?CAURI a MAXO:0000487 ;
        BTO_schema:mixedMN ?mixedMN ;
        BTO_schema:prevalentLMN ?prevalentLMN ;
        BTO_schema:prevalentUMN ?prevalentUMN . }
  bind( substr( (str(?limbsite)), 56) as ?onset_limb_type)
  # diagnosis
  optional{ ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220 .
     ?eventDiagnosisURI BTO_schema:startDate ?diagnosisDate . }
```

```
  # before onset
  optional{ ?idURI BTO_schema:undergo ?eventBOURI .
     ?eventBOURI a BTO_schema:Before_Onset ;
        BTO_schema:consists ?CABOURI .
     ?CABOURI a MAXO:0000487 ;
        BTO_schema:Weight ?weight_before_onset . }
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventBOURI . ?eventBOURI a
BTO_schema:Before_Onset ;
        BTO_schema:hasTrauma BTO_ni:Major_Truma_Before_Onset .},
"true"^^xsd:boolean, ?x = 0) as ?major_trauma_before_onset)
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventBOURI . ?eventBOURI a
BTO_schema:Before_Onset ;
        BTO_schema:consists BTO_ni:Surgical_Interventions_Before_Onset .},
"true"^^xsd:boolean, ?x = 0) as ?surgical_interventions_before_onset)
  #weightloss
  optional{ ?idURI BTO_schema:undergo ?eventCA .
     ?eventCA a NCIT:C39564 ;
        BTO_schema:consists ?CAfirstvisit .
     ?CAfirstvisit a MAXO:0000487 ;
        BTO_schema:moreThan10PercentWeightloss ?moreThan10PercentWeightloss . }
}ORDER BY ?id
```

3.2.3 Traumas Before Onset and Diseases

The subsequent query is used to interrogate the knowledge base and identify the traumas that happened to the patient before the onset and possible comorbidities affecting them.

Unless differently specified, all the variables described below are binary variables reporting whether the specific occurrence (either a trauma or a disease) happened to the patient. We use the value "*true*" in case the clinicians explicitly indicated the presence of the event. Otherwise, we do not report any value. This is because, in the original data, the empty value represented both the absence of the event and the fact that the clinician filling in the historical data did not know whether the event occurred to the patient. In this sense, it might be possible that, in some cases, the empty value is "*true*"; nevertheless, such occurrences are sporadic and therefore can be safely ignored.

Finally, traumas and surgeries are divided into two main categories: those that happened within a five-year range before the onset (associated variables are marked with the *_last_5_years* suffix in the name) and more than five years before the onset (variables marked with *_more_than_5_years* suffix).

The query is used to extract the following information concerning different diseases, traumas and surgeries occurred before the onset:

- **hypertension**
- **diabetes**
- **dyslipidemia**
- **thyroid_disorder**
- **autoimmune_disease**
- **stroke**

- **cardiac_disease**
- **primary_neoplasm**
- **head_trauma_last_5_years**
- **head_trauma_more_than_5_years**
- **neck_trauma_last_5_years**
- **neck_trauma_more_than_5_years**
- **cervical_trauma_last_5_years**
- **cervical_trauma_more_than_5_years**
- **thoracic_trauma_last_5_years**
- **thoracic_trauma_more_than_5_years**
- **lumbo_sacral_trauma_last_5_years**
- **lumbo_sacral_trauma_more_than_5_years**
- **cervical_spine_surgery_last_5_years**
- **cervical_spine_surgery_more_than_5_years**
- **thoracic_spine_surgery_last_5_years**
- **thoracic_spine_surgery_more_than_5_years**
- **lumbo_sacral_spine_surgery_last_5_years**
- **lumbo_sacral_spine_surgery_more_than_5_years**
- **upper_limb_surgery_last_5_years**
- **upper_limb_surgery_more_than_5_years**
- **lower_limb_surgery_last_5_years**
- **lower_limb_surgery_more_than_5_years**
- **abdominal_surgery_last_5_years**
- **abdominal_surgery_more_than_5_years**
- **thoracic_surgery_last_5_years**
- **thoracic_surgery_more_than_5_years**
- **pelvic_surgery_last_5_years**
- **pelvic_surgery_more_than_5_years**
- **head_neck_surgery_last_5_years**
- **head_neck_surgery_more_than_5_years**

```
SELECT ?id ?hypertension ?diabetes ?dyslipidemia ?thyroid_disorder
?autoimmune_disease ?stroke ?cardiac_disease ?primary_neoplasm
?head_trauma_last_5_years ?head_trauma_more_than_5_years
?neck_trauma_last_5_years ?neck_trauma_more_than_5_years
?cervical_trauma_last_5_years ?cervical_trauma_more_than_5_years
?thoracic_trauma_last_5_years ?thoracic_trauma_more_than_5_years
?lumbo_sacral_trauma_last_5_years ?lumbo_sacral_trauma_more_than_5_years
?cervical_spine_surgery_last_5_years ?cervical_spine_surgery_more_than_5_years
?thoracic_spine_surgery_last_5_years ?thoracic_spine_surgery_more_than_5_years
?lumbo_sacral_spine_surgery_last_5_years
?lumbo_sacral_spine_surgery_more_than_5_years ?upper_limb_surgery_last_5_years
?upper_limb_surgery_more_than_5_years ?lower_limb_surgery_last_5_years
?lower_limb_surgery_more_than_5_years ?abdominal_surgery_last_5_years
?abdominal_surgery_more_than_5_years ?thoracic_surgery_last_5_years
?thoracic_surgery_more_than_5_years ?pelvic_surgery_last_5_years
?pelvic_surgery_more_than_5_years ?head_neck_surgery_last_5_years
?head_neck_surgery_more_than_5_years WHERE {
  ?idURI a NCIT:C16960 . bind( substr( (str(?idURI)), 48) as ?id)
  # diseases
```

```
  bind(IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:Hypertension . },
"true"^^xsd:boolean, ?x = 0) as ?hypertension)
  bind(IF(exists{ ?idURI BTO_schema:hasDisease BTO_ni:Diabetes_Mellitus . },
"true"^^xsd:boolean, ?x = 0) as ?diabetes)
  bind( IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:Dyslipidemia . },
"true"^^xsd:boolean, ?x = 0) as ?dyslipidemia)
  bind( IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:Thyroid_Gland_Disorder . },
"true"^^xsd:boolean, ?x = 0) as ?thyroid_disorder)
  bind( IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:Autoimmune_Disease . },
"true"^^xsd:boolean, ?x = 0) as ?autoimmune_disease)
  bind( IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:stroke . }, "true"^^xsd:boolean,
?x = 0) as ?stroke)
  bind( IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:Cardiovascular_Disorder . },
"true"^^xsd:boolean, ?x = 0) as ?cardiac_disease)
  bind( IF( exists{ ?idURI BTO_schema:hasDisease BTO_ni:Primary_Neoplasm . },
"true"^^xsd:boolean, ?x = 0) as ?primary_neoplasm)
  # trauma
  # head
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
     BTO_schema:howLong "in the last 5 years"^^xsd:string ;
     BTO_schema:hasTrauma ?traumaURI .
     ?traumaURI a NCIT:C3671 ;
        BTO_schema:traumaArea BTO_ni:head . }, "true"^^xsd:boolean, ?x = 0) as
?head_trauma_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
     BTO_schema:howLong "> 5 years"^^xsd:string ;
     BTO_schema:hasTrauma ?traumaURI.
     ?traumaURI a NCIT:C3671 ;
        BTO_schema:traumaArea BTO_ni:head . }, "true"^^xsd:boolean, ?x = 0) as
?head_trauma_more_than_5_years )
  # neck
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
     BTO_schema:howLong "in the last 5 years"^^xsd:string ;
     BTO_schema:hasTrauma ?traumaURI.
     ?traumaURI a NCIT:C3671 ;
        BTO_schema:traumaArea BTO_ni:neck . }, "true"^^xsd:boolean, ?x = 0) as
?neck_trauma_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
     BTO_schema:howLong "> 5 years"^^xsd:string ;
     BTO_schema:hasTrauma ?traumaURI.
     ?traumaURI a NCIT:C3671 ;
        BTO_schema:traumaArea BTO_ni:neck . }, "true"^^xsd:boolean, ?x = 0) as
?neck_trauma_more_than_5_years )
  # cervical
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
```

```
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:hasTrauma ?traumaURI.
      ?traumaURI a NCIT:C3671 ;
         BTO_schema:traumaArea BTO_ni:cervical_region_of_vertebral_column . },
"true"^^xsd:boolean, ?x = 0) as ?cervical_trauma_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:hasTrauma ?traumaURI.
      ?traumaURI a NCIT:C3671 ;
         BTO_schema:traumaArea BTO_ni:cervical_region_of_vertebral_column . },
"true"^^xsd:boolean, ?x = 0) as ?cervical_trauma_more_than_5_years )
  # thoracic
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:hasTrauma ?traumaURI.
      ?traumaURI a NCIT:C3671 ;
         BTO_schema:traumaArea BTO_ni:thoracic_skeleton .}, "true"^^xsd:boolean, ?x = 0)
as ?thoracic_trauma_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:hasTrauma ?traumaURI.
      ?traumaURI a NCIT:C3671 ;
         BTO_schema:traumaArea BTO_ni:thoracic_skeleton . }, "true"^^xsd:boolean, ?x = 0)
as ?thoracic_trauma_more_than_5_years )
  # lumbo
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:hasTrauma ?traumaURI.
      ?traumaURI a NCIT:C3671 ;
         BTO_schema:traumaArea BTO_ni:lumbosacral_nerve_plexus . },
"true"^^xsd:boolean, ?x = 0) as ?lumbo_sacral_trauma_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:hasTrauma ?traumaURI.
      ?traumaURI a NCIT:C3671 ;
         BTO_schema:traumaArea BTO_ni:lumbosacral_nerve_plexus . },
"true"^^xsd:boolean, ?x = 0) as ?lumbo_sacral_trauma_more_than_5_years )
  # surgery
  # cervical
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
```

```
        BTO_schema:surgicalArea BTO_ni:cervical_region_of_vertebral_column . },
"true"^^xsd:boolean, ?x = 0) as ?cervical_spine_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
        BTO_schema:surgicalArea BTO_ni:cervical_region_of_vertebral_column .},
"true"^^xsd:boolean, ?x = 0) as ?cervical_spine_surgery_more_than_5_years )
  # thoracic spine
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
        BTO_schema:surgicalArea BTO_ni:thoracic_spinal_cord . }, "true"^^xsd:boolean, ?x
= 0) as ?thoracic_spine_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
        BTO_schema:surgicalArea BTO_ni:thoracic_spinal_cord . }, "true"^^xsd:boolean, ?x
= 0) as ?thoracic_spine_surgery_more_than_5_years )
  # lumbo
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
        BTO_schema:surgicalArea BTO_ni:lumbosacral_nerve_plexus . },
"true"^^xsd:boolean, ?x = 0) as ?lumbo_sacral_spine_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
        BTO_schema:surgicalArea BTO_ni:lumbosacral_nerve_plexus . },
"true"^^xsd:boolean, ?x = 0) as ?lumbo_sacral_spine_surgery_more_than_5_years )
  # upper limb
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
        BTO_schema:surgicalArea BTO_ni:upper_limb . }, "true"^^xsd:boolean, ?x = 0) as
?upper_limb_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
```

```
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:upper_limb . }, "true"^^xsd:boolean, ?x = 0) as
?upper_limb_surgery_more_than_5_years )
  # lower limb
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:lower_limb . }, "true"^^xsd:boolean, ?x = 0) as
?lower_limb_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:lower_limb . }, "true"^^xsd:boolean, ?x = 0) as
?lower_limb_surgery_more_than_5_years )
  # abdominal
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:abdominal_fascia . }, "true"^^xsd:boolean, ?x = 0)
as ?abdominal_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:abdominal_fascia . }, "true"^^xsd:boolean, ?x = 0)
as ?abdominal_surgery_more_than_5_years )
  # thoracic
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:thoracic_skeleton . }, "true"^^xsd:boolean, ?x = 0)
as ?thoracic_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:thoracic_skeleton .}, "true"^^xsd:boolean, ?x = 0)
as ?thoracic_surgery_more_than_5_years )
  # pelvic
```

```
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:pelvic_complex . }, "true"^^xsd:boolean, ?x = 0)
as ?pelvic_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:pelvic_complex . }, "true"^^xsd:boolean, ?x = 0)
as ?pelvic_surgery_more_than_5_years )
  # head neck
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "in the last 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:head-neck . }, "true"^^xsd:boolean, ?x = 0) as
?head_neck_surgery_last_5_years )
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a
BTO_schema:Before_Onset ;
      BTO_schema:howLong "> 5 years"^^xsd:string ;
      BTO_schema:consists ?surgeryURI.
      ?surgeryURI a NCIT:C15329 ;
         BTO_schema:surgicalArea BTO_ni:head-neck . }, "true"^^xsd:boolean, ?x = 0) as
?head_neck_surgery_more_than_5_years )
}ORDER BY ?id
```

3.2.4 Smoking

The smoking habit is another important behavior that might affect the progression of the ALS concerns. We extract from our knowledge base the following variables concerning the smoking behavior of the patient:

- **smoking**: binary variable indicating whether the patient was smoking at the diagnosis time.
- **smoking_startYear**: the year when patients began smoking (if available/concerning the patient).
- **smoking_endYear**: when the patient stopped smoking (if available/concerning the patient).
- **dailyCigarettes**: float value indicating the average number of cigarettes smoked by the patient in a day (either currently or when they were smoking).
- **packYear**: the packYear ((number of cigarettes smoked per day/20 × number of years smoked).

```
SELECT ?id ?smoking ?smoking_startYear ?smoking_endYear ?dailyCigarettes
?packYear WHERE {
```

```
  ?idURI a NCIT:C16960 . bind( substr( (str(?idURI)), 48) as ?id)
  # smoking
  bind( IF( exists{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a NCIT:C25499 ;
     BTO_schema:hasRegisteredBehaviour ?smokingURI . }, "true"^^xsd:boolean, ?x = 0)
as ?smoking)
  optional{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a NCIT:C25499 ;
     BTO_schema:hasRegisteredBehaviour ?smokingURI .
     ?smokingURI BTO_schema:startYear ?smoking_startYear . }
  optional{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a NCIT:C25499 ;
     BTO_schema:hasRegisteredBehaviour ?smokingURI .
     ?smokingURI BTO_schema:endYear ?smoking_endYear . }
  optional{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a NCIT:C25499 ;
     BTO_schema:hasRegisteredBehaviour ?smokingURI .
     ?smokingURI BTO_schema:packYear ?packYear . }
  optional{ ?idURI BTO_schema:undergo ?eventURI . ?eventURI a NCIT:C25499 ;
     BTO_schema:hasRegisteredBehaviour ?smokingURI .
     ?smokingURI BTO_schema:dailyCigarettes ?dailyCigarettes . }
}ORDER BY ?id
```

3.2.5 Blood And Genetic Tests

Finally, concerning the patient's clinical history, we report pieces of information concerning their blood tests. We also report, when available, details about possible genetic mutations that affect the patient and correlate with the progression of the disease.

Since the two considered clinics report two different types of mutations for the same genes, we need two sets of variables describing such genetic mutations.

About Turin, the presence of mutations on the FUS, SOD1 and TARDBP genes is described with a set of binary variables, while, for what concerns mutations on the C9orf72 gene, the clinic reports a categorical classification of such mutation. On the other hand, Lisbon reports a textual description of the mutation and thus we report it in the dataset.

The variables extracted from the knowledge base are the following:

- **turin_FUS**
- **turin_SOD1**
- **turin_TARDBP**
- **turin_C9orf72_kind**
- **lisbon_FUS_openbox**
- **lisbon_SOD1_openbox**
- **lisbon_TARDBP_openbox**
- **lisbon_C9orf72**
- **hypertension**
- **diabetes**
- **dyslipidemia**
- **thyroid_disorder**
- **CK_level**
- **CK_lower_range**
- **CK_upper_range**

- **Albumin_level**
- **Albumin_lower_range**
- **Albumin_upper_range**
- **Creatinine_level**
- **Creatinine_lower_range**
- **Creatinine_upper_range**
- **Total_Cholesterol_level**
- **Total_Cholesterol_lower_range**
- **Total_Cholesterol_upper_range**
- **HDL_Cholesterol_level**
- **HDL_Cholesterol_lower_range**
- **HDL_Cholesterol_upper_range**
- **LDL_Cholesterol_level**
- **LDL_Cholesterol_lower_range**
- **LDL_Cholesterol_upper_range**
- **Triglycerides_level**
- **Triglycerides_lower_range**
- **Triglycerides_upper_range**

```
SELECT ?id ?turin_FUS ?turin_SOD1 ?turin_TARDBP ?turin_C9orf72_kind
?lisbon_FUS_openbox ?lisbon_SOD1_openbox ?lisbon_TARDBP_openbox
?lisbon_C9orf72 ?hypertension ?diabetes ?dyslipidemia ?thyroid_disorder ?CK_level
?CK_lower_range ?CK_upper_range ?Albumin_level ?Albumin_lower_range
?Albumin_upper_range ?Creatinine_level ?Creatinine_lower_range
?Creatinine_upper_range ?Total_Cholesterol_level ?Total_Cholesterol_lower_range
?Total_Cholesterol_upper_range ?HDL_Cholesterol_level ?HDL_Cholesterol_lower_range
?HDL_Cholesterol_upper_range ?LDL_Cholesterol_level ?LDL_Cholesterol_lower_range
?LDL_Cholesterol_upper_range ?Triglycerides_level ?Triglycerides_lower_range
?Triglycerides_upper_range WHERE {
  ?idURI a NCIT:C16960 . bind( substr( (str(?idURI)), 48) as ?id)
  # gene turin
  optional{ ?idURI BTO_schema:hasGene ?geneURI . ?geneURI a OGG:3000203228 ;
    BTO_schema:kind ?turin_C9orf72_kind . }
  bind( IF( exists{ ?idURI BTO_schema:hasGene BTO_ni:FUS .}, "true"^^xsd:boolean, ?x =
0) as ?turin_FUS)
  bind( IF( exists{ ?idURI BTO_schema:hasGene BTO_ni:SOD1 .}, "true"^^xsd:boolean, ?x =
0) as ?turin_SOD1)
  bind( IF( exists{ ?idURI BTO_schema:hasGene BTO_ni:TARDBP .}, "true"^^xsd:boolean,
?x = 0) as ?turin_TARDBP)
  # gene lisbon
  optional{ ?idURI BTO_schema:hasGene ?geneURI . ?geneURI a OGG:3000002521 ;
    BTO_schema:open_box ?lisbon_FUS_openbox . }
  optional{ ?idURI BTO_schema:hasGene ?geneURI . ?geneURI a OGG:3000006647 ;
    BTO_schema:open_box ?lisbon_SOD1_openbox . }
  optional{ ?idURI BTO_schema:hasGene ?geneURI . ?geneURI a OGG:3000023435 ;
    BTO_schema:open_box ?lisbon_TARDBP_openbox . }
  bind( IF( exists{ ?idURI BTO_schema:hasGene BTO_ni:C9orf72 .}, "true"^^xsd:boolean,
?x = 0) as ?lisbon_C9orf72)
  #blood test
```

```
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:CK_level ?CK_level . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:CK_lower_range ?CK_lower_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:CK_upper_range ?CK_upper_range .}
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Albumin_level ?Albumin_level . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Albumin_lower_range ?Albumin_lower_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Albumin_upper_range ?Albumin_upper_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Creatinine_level ?Creatinine_level . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Creatinine_lower_range ?Creatinine_lower_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Creatinine_upper_range ?Creatinine_upper_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Total_Cholesterol_level ?Total_Cholesterol_level . }
```

```
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Total_Cholesterol_lower_range ?Total_Cholesterol_lower_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Total_Cholesterol_upper_range ?Total_Cholesterol_upper_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:HDL_Cholesterol_level ?HDL_Cholesterol_level . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:HDL_Cholesterol_lower_range ?HDL_Cholesterol_lower_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:HDL_Cholesterol_upper_range ?HDL_Cholesterol_upper_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:LDL_Cholesterol_level ?LDL_Cholesterol_level . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:LDL_Cholesterol_lower_range ?LDL_Cholesterol_lower_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:LDL_Cholesterol_upper_range ?LDL_Cholesterol_upper_range . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Triglycerides_level ?Triglycerides_level . }
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Triglycerides_lower_range ?Triglycerides_lower_range . }
```

```
  optional { ?idURI BTO_schema:undergo ?eventDiagnosisURI . ?eventDiagnosisURI a
NCIT:C15220;
     BTO_schema:consists ?BloodURI .
     ?BloodURI a NCIT:C49286 ;
        BTO_schema:Triglycerides_upper_range ?Triglycerides_upper_range . }
}ORDER BY ?id
```

3.3 Visits

Besides providing personal information about the patients, to allow proper prediction of the events, we need to gather details concerning the clinical progression of the disease. Therefore, we include in our dataset data concerning patient's spirometries and the value of the ALSFRS-R items observed in different moments in time.

We are interested in the following variables:

- **date_spiro**: date of a spirometric visit that the patient underwent.
- **fvcValue**: percentual value of the FVC measured during the spirometry.
- **date_alsfrs_r**: date of the administration of the ALSFRS-R questionnaire to the patient.
- **alsfrs_r_tot_score**: total sum of the ALSFRS-R questionnaire items.
- **bulbar_subscore**: sum of the scores relative to the bulbar system. This variable is computed as q1+q2+q3.
- **motor_subscore**: sum of the scores relative to the motor system. This variable is computed as q4+q5+q6+q7+q8+q9.
- **respiratory_subscore**: sum of the scores relative to the respiratory system. This variable is computed as q10+q11+q12.
- We also report the scores for the single ALSFRS items with the following names:
    - **q1**
    - **q2**
    - **q3**
    - **q4**
    - **q5**
    - **q6**
    - **q7**
    - **q8**
    - **q9**
    - **q10**
    - **q11**
    - **q12**

Each row corresponds to only one type of record, either a spirometry or an ALSFRS-R questionnaire. Therefore, for each row, only a subset of the variables is set (either all those related to the spirometry, or all those related to the ALSFRS-R). The following query extracts such information.

```
SELECT ?id ?date_spiro ?fvcValue ?date_alsfrs_r ?alsfrs_r_tot_score ?bulbar_subscore
?motor_subscore ?respiratory_subscore ?q1 ?q2 ?q3 ?q4 ?q5 ?q6 ?q7 ?q8 ?q9 ?q10 ?q11
?q12 WHERE {
{
```

```
  # Spiro
  SELECT ?id ?date_spiro ?fvcValue WHERE {
     ?idURI a NCIT:C16960 ;
        BTO_schema:undergo ?eventURI .
     ?eventURI a NCIT:C74589 ;
        BTO_schema:startDate ?date_spiro ;
        BTO_schema:consists ?testURI .
     ?testURI a NCIT:C38081 ;
        BTO_schema:FVCrelative ?fvcValue .
     bind( substr( (str(?idURI)), 48) as ?id)
  } ORDER BY ?id ?date_spiro
}
UNION
{
  # ALSFRS-R
  SELECT ?id ?date_alsfrs_r ?alsfrs_r_tot_score ?bulbar_subscore ?motor_subscore
?respiratory_subscore ?q1 ?q2 ?q3 ?q4 ?q5 ?q6 ?q7 ?q8 ?q9 ?q10 ?q11 ?q12 WHERE {
     ?idURI a NCIT:C16960 ;
        BTO_schema:undergo ?eventURI .
     ?eventURI a NCIT:C74589 ;
        BTO_schema:startDate ?date_alsfrs_r ;
        BTO_schema:consists ?qRURI .
     ?qRURI a BTO_schema:ALSFRS-R ;
        BTO_schema:alsfrs-r-tot ?alsfrs_r_tot_score ;
        BTO_schema:bulbar_subscore ?bulbar_subscore ;
        BTO_schema:motor_subscore ?motor_subscore ;
        BTO_schema:respiratory_subscore ?respiratory_subscore ;
        BTO_schema:alsfrs_1 ?q1 ;
        BTO_schema:alsfrs_2 ?q2 ;
        BTO_schema:alsfrs_3 ?q3 ;
        BTO_schema:alsfrs_4 ?q4 ;
        BTO_schema:alsfrs_5 ?q5 ;
        BTO_schema:alsfrs_6 ?q6 ;
        BTO_schema:alsfrs_7 ?q7 ;
        BTO_schema:alsfrs_8 ?q8 ;
        BTO_schema:alsfrs_9 ?q9 ;
        BTO_schema:alsfrs_10 ?q10 ;
        BTO_schema:alsfrs_11 ?q11 ;
        BTO_schema:alsfrs_12 ?q12 .
     bind( substr( (str(?idURI)), 48) as ?id)
  } ORDER BY ?id ?date_alsfrs
}
}ORDER BY ?id ?date_alsfrs_r ?date_spiro
```

3.4 Clinical Interventions

The query described here is used to retrieve the dates concerning the four main clinical events that might affect a patient: NIV, PEG, Tracheostomy, and Death. Notice that, in post processing, the tracheostomy date will be considered as the death date.

More in details, we retrieve from our knowledge base the following three variables:

- **NIV_date**: string describing the date of the first NIV.
- **PEG_date**: string describing the date of the first PEG.
- **Tracheostomy_date**: string containing the date of the tracheostomy.
- **deathDate**: string reporting the date of death for patients whose status alive is set to false.

All these values may be empty in case the patient has not undergone a certain event.

```
SELECT ?id ?NIV_date ?PEG_date ?Tracheotomy_date ?deathDate WHERE {
  ?idURI a NCIT:C16960 . bind( substr( (str(?idURI)), 48) as ?id)
  # niv - peg - tracheo
  optional{ ?idURI BTO_schema:undergo ?eventNIVURI . ?eventNIVURI a NCIT:C74589 ;
     BTO_schema:consists BTO_ni:Non Invasive_Mechanical_Ventilation ;
     BTO_schema:startDate ?NIV_date . }
  optional{ ?idURI BTO_schema:undergo ?eventPEGURI . ?eventPEGURI a NCIT:C74589 ;
     BTO_schema:consists BTO_ni:Percutaneous_Endoscopic_Gastrostomy ;
     BTO_schema:startDate ?PEG_date . }
  optional{ ?idURI BTO_schema:undergo ?eventTracheURI . ?eventTracheURI a
NCIT:C74589 ;
     BTO_schema:consists BTO_ni:Tracheotomy ;
     BTO_schema:startDate ?Tracheotomy_date . }
  optional{ ?idURI BTO_schema:dateOfDeath ?deathDate . }
}ORDER BY ?id
```

4 CHALLENGE DATASETS CREATION

The iDPP@CLEF 2022 lab was divided into three subtasks:

- Subtask A: this subtask involves the prediction of either an NIV outcome or the patient's death.
- Subtask B: this subtask involves the prediction of either a PEG outcome or the patient's death.
- Subtask C: this subtask involves the prediction of the patient's death.

We divide our set of (valid) patients – filtered according to the rules specified in Section 2 – into three overlapping datasets, each of which is used for one of the subtasks. Each patient record is then associated with a label describing the outcome – NIV, PEG, or DEATH. Notice that if a patient undergoes an NIV or PEG event and later dies, then we label the record for such patient with either NIV or PEG, and we require participants to the iDPP challenge to predict such event since it happened before death.

Besides including patients that either undergo the specific outcome focus of the subtask or death, we also include into the dataset patients that do not undergo any event – we refer to this case as "censoring event". Patients with whom we associate the censoring event are labelled as NONE in the iDPP dataset.

4.1 Filtering Data on Information Availability Constraints

Besides removing records with missing attributes or spurious information, as described in Section 2, we exclude from the challenge datasets records that do not have at least six months of information before the outcome event. This filtering process grants that every record in the final dataset contains enough information to allow proper predictions.

Furthermore, we remove from the dataset all the information concerning visits happened six months after the first ALSFRS to grant that all the patients have the same quantity of information.

Figure 7 illustrates a set of - synthetic - patients and their clinical history, describing whether they satisfy the conditions to be inserted into the dataset.

By construction, the first ALSFRS visit (blue bullets) is considered at time 0, while the moment of the previous spirometries (yellow bullets) and subsequent visits are indicated as the difference in months concerning the reference ALSFRS.

- Patient 1 is inserted into the dataset, having a proper sequence of visits, questionnaires, and events (at least six months of information before the first event).
- Patient 2, however, cannot be included in the dataset since they do not have enough information.
- For Patient 3, we observe that only four months passed between the first ALSFRS and the first event. Thus, we cannot retain the record even though we have 6 months of global information (first spirometry to event).

- Although patient 4 has a single ALSFRS, they can be included in the dataset since the distance between the first ALSFRS and the event is above six months.
- Patients 5 presents overall more than 6 months of information, but not before the first event (X at month -4) and therefore is excluded.
- Patients 6 need to be excluded from further analyses not having enough history, regardless of the spirometry taken before the first ALSFRS.
- Patients 7 and 8, on the other hand, can be considered: the former has a good clinical history, while the latter, even though they have a "censored" event, has more than six months of history.

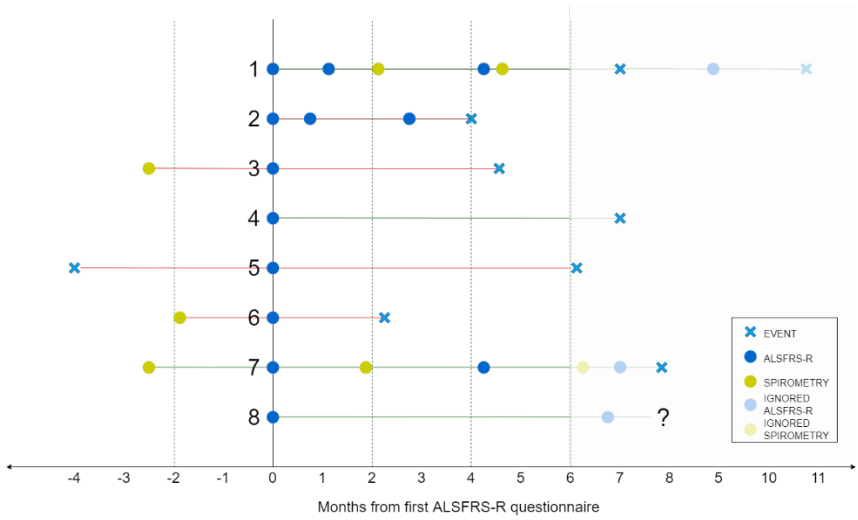

Figure 7 Sequences of events that allow (or forbid) a patient to be considered as suitable to belong to the dataset. Notice that, we remove from the dataset all events happened after 6 months from the first visit (such events are greyed out in the plot).

Notice that in several scenarios, a record does not have enough information to enter the dataset for a specific subtask, but it has accumulated it for another. For example, if the patient undergoes an NIV outcome at month 4 after the first ALSFRS-R questionnaire, we do not consider such patient suitable for the dataset used in subtask A.

Nevertheless, after the NIV event, the patient might take additional ALSFRS-R questionnaires and spirometry visits and undergo a PEG event at month 7. In this case, the patient has enough information to predict a PEG event, which will be included in the dataset used in subtask B.

4.2 Splitting into Sub-Datasets

Table 1 reports the number of valid patients present in each dataset, with the count of the labels that characterize them.

Dataset A is the least populated: an NIV procedure is often required quite early in the course of the disease, and thus, for a large group of patients (755), it happened before six months of follow-up. Thus, we were forced to exclude them since not enough information was available in those cases.

We observe that dataset B contains 2145 individuals (414 were discarded). We observe an increase in the number of individuals considered suitable for this dataset: it is more likely that a PEG will be necessary much later in the progression of the disease, and thus more patients accrue more than six months of data before the outcome. Notice that patients that were labelled as NIV in dataset A, if they are suitable to enter dataset B (meaning that has at least six months of information before the PEG), can be labelled as either PEG, DEATH or NONE in the case they did not undergo any events other than the NIV for which they entered dataset A. Conversely, patients labelled as DEATH can either be labelled with PEG, in case they received a PEG after six months from the first ALSFRS-R and before their death, or DEATH in case they did not, but they cannot switch to class NONE. The converse is also true if we consider the relationship between datasets B and A.

Finally, dataset C contains 2250 patients, with only 305 records discarded. Given the criteria used to construct Dataset C, it contains patients present in datasets A and B. Patients that received an NIV or a PEG after six months from the first ALSFRS-R have at least six months of information before their death. We also include new patients: those that had an NIV before six months from the first ALSFRS-R and those that had a PEG before six months but survived (or died) more than six months after the first ALSFRS-R. Patients that were not included in dataset C are those that survived less than six months after the first ALSFRS-R.

Table 1 Result of the splitting of the patients among different Datasets.

Variable	Dataset A	Dataset B	Dataset C
Number of subjects	1804	2145	2250
Outcome type	NIV: 837 Death: 788 Censoring: 169	PEG: 621 Death: 1220 Censoring: 304	Death: 1903 Censoring: 347

4.2.1 Temporal Analysis of the Data

We now report the main statistics concerning the three released datasets.

4.2.1.1 NIV and Death

Concerning dataset A (NIV, Death or Censoring Events), we observe that 1804 patients satisfy the conditions to be included in the dataset (6 months of data between the first ALSFRS-R and the event).

Figure 8 reports the distribution of patients with respect to the distance from the event. We notice a very steep distribution, suggesting that the events often happen before month 6. Compared to Figure 12 and Figure 16, we argue that the NIV is the most likely event in the first six months. Both Figure 12 and Figure 16 present a lower peak in the first part of the distribution.

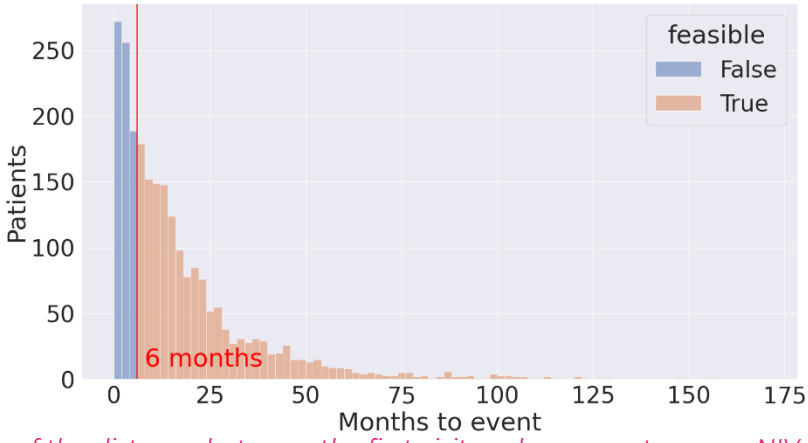

Figure 8 Distribution of the distance between the first visit and one event among NIV, Death, or Censoring event (the one happening first) over the patient set.

Figure 9 reports the number of visits, according to their type, with respect to the time from the reference ALSFRS-R questionnaire. We notice a predictable increase in the number of visits with respect to the months passed since the diagnosis.

Figure 10 illustrates the number of visits associated with each feasible patient that was included in the dataset. Interestingly, we notice that the mean number of visits in the considered six months is 3.3, with a slightly lower median number.

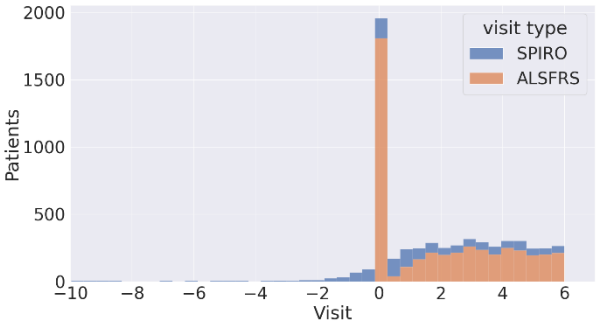

Figure 9 Distribution of the visits included in dataset A. - Number of visits by time span.

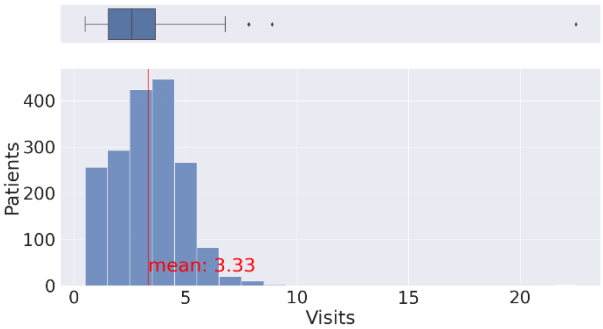

Figure 10 Distribution of the visits included in dataset A. - Number of visits by patient.

Finally, Figure 11 reports the number of patients incurring in a certain event for dataset A. Among patients included in dataset A, 788 are labelled with the event DEATH, 837 with the event NIV and 179 patients are labelled with NONE, which indicates the censoring event. If a patient incurred both NIV and Death, the event associated with that patient that needs to be predicted is only the NIV.

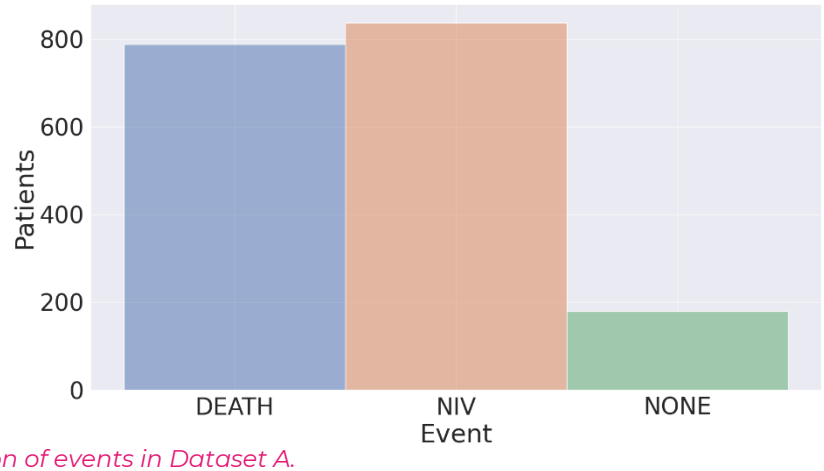

Figure 11 Distribution of events in Dataset A.

4.2.1.2 PEG and Death

Figure 12 reports the distribution of the distance to the event if we consider as feasible events the PEG, Death, or the censoring event. As we noticed with the number of subjects that belong to each dataset, for what concerns dataset B, we have fewer subjects that have less than 6 months between their first ALSFRS-R and the event. We can accrue 2145 suitable subjects for this dataset.

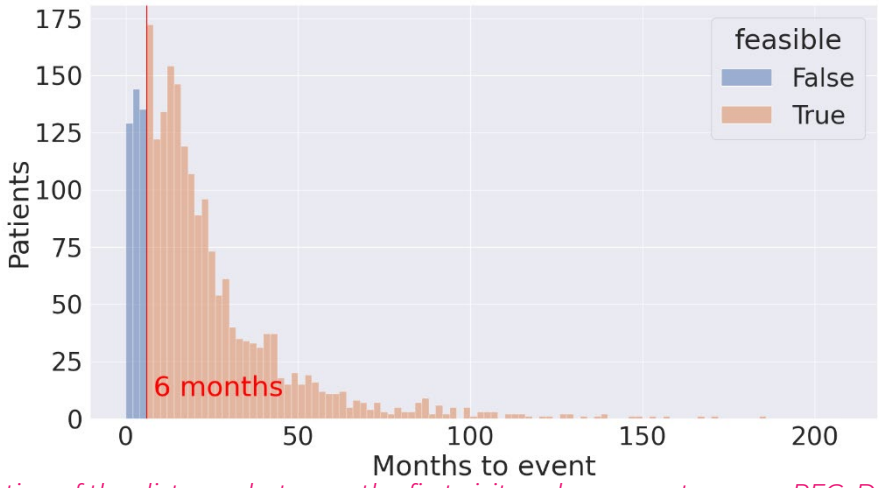

Figure 12 Distribution of the distance between the first visit and one event among PEG, Death or Censoring event (the one happening first) over the patient set.

Following what we observed for dataset A, we report the number of visits according to their distance from the first ALSFRS-R (Figure 13) as before, we notice an increase in the number of visits with the months passing.

In this case, as Figure 14 illustrates, we have on average 3.35 visits for each patient, with most patients having between 2 and 3 visits.

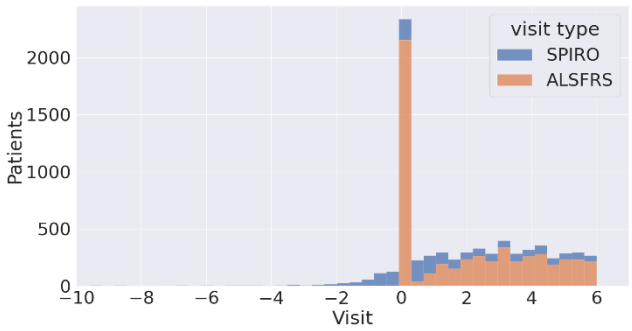

Figure 13 Number of visits by time span.

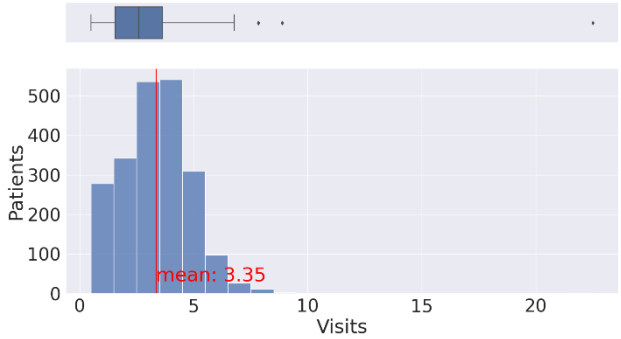

Figure 14 Number of visits by patient.

Figure 15 exhibits the distribution of events in dataset B. Differently from dataset A, where the most common event was the NIV (the focus of the subtask) in this case, the most common event is DEATH. This indicates a generally increased probability of incurring death before receiving a PEG. From the perspective of an AI practitioner, it also indicates that models that were suited to task A are likely not usable trivially on this second dataset but would require a specific tuning to be applied on this second scenario.

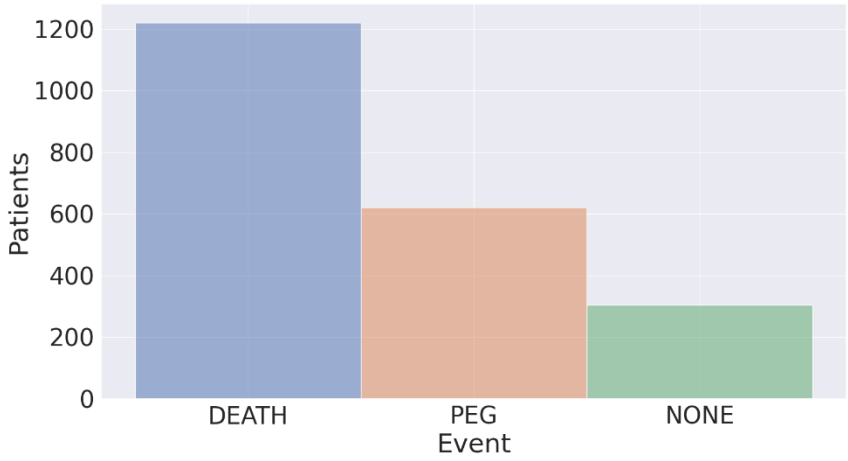

Figure 15 Distribution of outcomes in Dataset B.

4.2.1.3 Death and Censoring events

The final dataset, dataset C, concerns the death and censoring events and it is used in the tasks that require predicting such events. As shown by Figure 16 for what concerns the number of patients included in this dataset, we overcome both dataset A and dataset B: several patients that did not have the right features to be considered feasible before (at least 6 months of information from the first ALSFRS-R) can be included in this dataset. Dataset C counts 2250 records associated with as many patients.

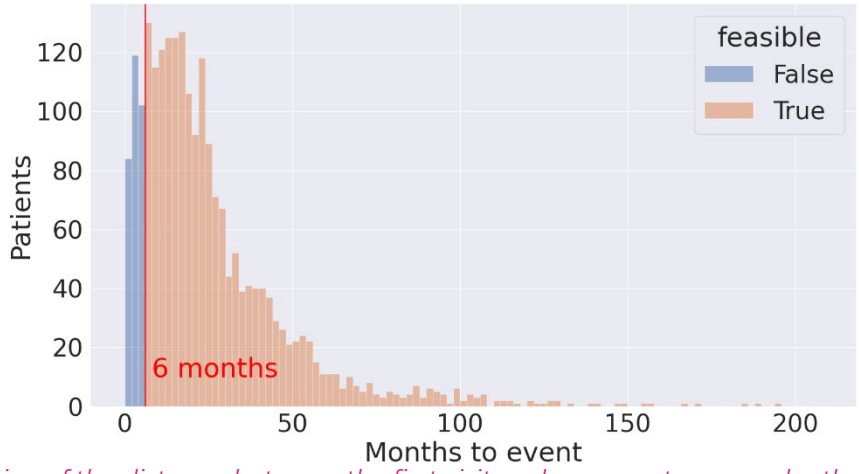

Figure 16 Distribution of the distance between the first visit and one event among death or Censoring event over the patient set.

Similarly, to the previous datasets, Figure 17 and Figure 18 illustrate the distribution of the visits, either with respect to the timespan or to the patients. As for the previous case, we have 3.35 visits per patient on average, with the mode on 3 visits and approximately 50% of the patients having between 2 and 3 visits.

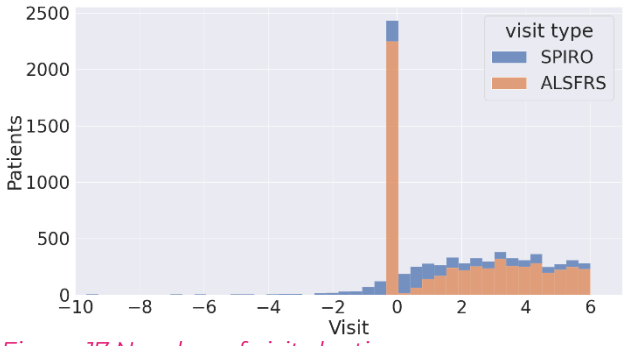

Figure 17 Number of visits by time span.

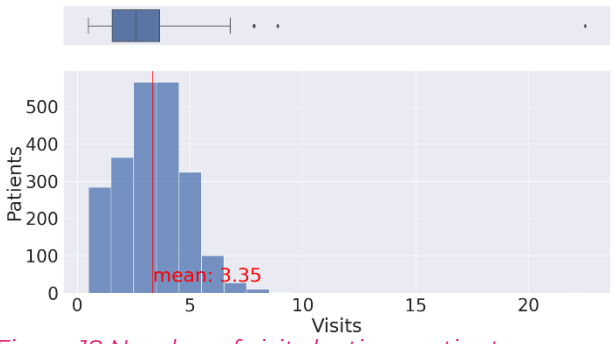

Figure 18 Number of visits by time patient.

Figure 19 illustrates how events distribute in dataset C. We have more deaths than censoring events. This is expected: historical data regards patients with ages up to more than 90 years followed sometimes for several years. As mentioned before, 1903 patients overwent the death event, while 347 incurred into the censoring event. Notice that, in this case, we have more censoring events than for what concerns previous datasets, since patients that incurred NIV and/or PEG but survived are considered suitable for the censoring event in this case.

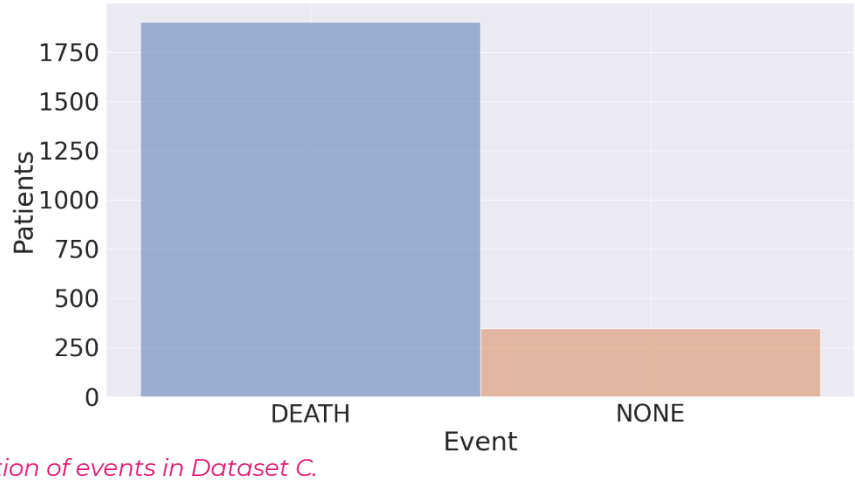

Figure 19 Distribution of events in Dataset C.

4.2.2 Dataset Splitting into Training and Test Sets

We now report the procedure followed to split the available datasets into training and test sets.

4.2.2.1 Splitting procedure

Each of the three available datasets (dataset A, B, and C) was separately split into a training set and a test set, with proportions 80% and 20%, respectively. The data were split stratifying the subjects according to: the clinical event experienced, namely "outcome type", i.e., death/NIV/none for dataset A, death/PEG/none for dataset B, and death/none for dataset C; the time between the first available visit and the clinical event occurrence, namely "outcome time".

Stratifying by these two variables is instrumental to the fairness of the challenge as it forces an equal distribution of their levels across the two subsets.

4.2.2.2 Assessment of the splitting effectiveness

The simplest method to verify whether stratification has been performed correctly is to compare the distribution of the stratification variables (outcome time and outcome type) in each training/test pair. From the literature, certain variables are known to be particularly relevant in predicting events related to ALS progression [CLH+09], therefore, even though they were not included in the stratification criteria, we verified that sex, age at onset, onset site, ALSFRS-F slope, and the number of available visits in the first 6 months were also equally represented in the training and test sets. Table, Table 3 and Table 4 report the variables' distributions for datasets A, B, and C, respectively: the second column report the distribution in the training sets and the third column in the test sets.

Table 2 Dataset A, comparison between training and test populations. Continuous variables are presented as median [1st - 3rd quartiles]; discrete variables as count (percentage on sample total), for each level.

Variable	Training set	Test set
Number of subjects	1454	350
Outcome type	Death: 636 (44%) NIV: 675 (46%) Censoring: 143 (10%)	Death: 152 (43%) NIV: 162 (46%) Censoring: 36 (10%)
Outcome time (Months)	17.75 [11.14-30.99]	20.72 [11.25-36.76]

Variable	Training set	Test set
Sex	M: 743 (51%) F: 711 (49%)	M: 188 (54%) F: 16 (46%)
Age at onset (Years)	64.89 [55.66-70.76]	64.76 [56.66-71.58]
Onset site	Bulbar: 449 (31%) Axial: 3 (0.002%) Generalized: 4 (0.003%) Limbs: 998 (68%)	Bulbar: 105 (30%) Axial: 0 (0%) Generalized: 0 (0%) Limbs: 242 (70%)
ALSFRS-R slope	0.43 [0.24-0.79]	0.41 [0.23-0.80]
Number of available visits	2.00 [2.00-3.00]	3.00 [2.00-3.00]

Table 3 Dataset B, comparison between training and test populations. Continuous variables are presented as median [1st - 3rd quartiles]; discrete variables as count (percentage on sample total), for each level.

Variable	Training set	Test set
Number of subjects	1715	430
Outcome type	Death: 969 (57%) PEG: 501 (29%) Censoring: 245 (14%)	Death: 251 (58%) PEG: 120 (28%) Censoring: 59 (14%)
Outcome time (Months)	19.97 [12.57-36.53]	21.82 [12.70-38.30]
Sex	M: 923(54%) F: 792 (46%)	M: 241 (56%) F: 189 (44%)
Age at onset (Years)	65.14 [56.86-71.88]	64.83 [55.99-70.42]
Onset site	Bulbar: 499 (29%) Axial: 31 (2%) Generalized: 8 (0.5%) Limbs: 1177 (68.5%)	Bulbar: 125 (29%) Axial: 12 (3%) Generalized: 1 (0.2%) Limbs: 292 (68%)
ALSFRS-R slope	0.47 [0.25-0.84]	0.44 [0.24-0.85]
Number of available visits	2.00 [2.00-3.00]	2.00 [2.00-3.00]

Table 4 Dataset C, comparison between training and test populations. Continuous variables are presented as median [1st - 3rd quartiles]; discrete variables as count (percentage on sample total), for each level.

Variable	Training set	Test set
Number of subjects	1756	494
Outcome type	Death: 1486(85%) Censoring: 270 (15%)	Death: 417(84%) Censoring: 77 (16%)
Outcome time (Months)	24.68 [14.42-41.84]	22.48 [13.72-38.91]
Sex	M: 930 (53%) F: 826 (47%)	M: 273 (55%) F: 221 (45%)
Age at onset (Years)	65.38 [58.27-72.18]	65.03 [57.02-70.86]
Onset site	Bulbar: 554 (31.5%)	Bulbar: 149 (30%)

Variable	Training set	Test set
	Axial: 32 (2%) Generalized: 8 (0.5%) Limbs: 1162 (66%)	Axial: 13 (3%) Generalized: 1 (0.2%) Limbs: 331 (67%)
ALSFRS-R slope	0.49 [0.26-0.88]	0.45 [0.24-0.85]
Number of available visits	2.00 [2.00-3.00]	2.00 [2.00-3.00]

The distributions of the two variables used to stratify are also represented in Figures 20, 21, and 22. Bar plots are employed for categorical variables, e.g. outcome type, and density plots for continuous variables, i.e. outcome time.

Figure 20 Comparison of the distributions of stratification variables for dataset A: outcome type on the left and outcome time on the right. The distribution on the training set is in blue while that of the test set in orange.

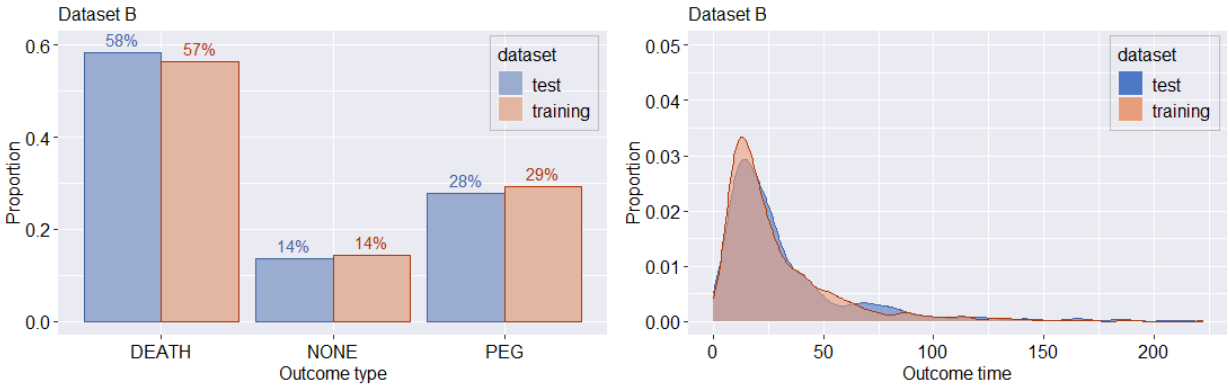

Figure 21 Comparison of the distributions of stratification variables for dataset B: outcome type on the left and outcome time on the right. The distribution on the training set is in blue while that of the test set in orange.

Figure 22 Comparison of the distributions of stratification variables for dataset C: outcome type on the left and outcome time on the right. The distribution on the training set is in blue while that of the test set in orange.

Since the distributions are similar, we concluded that the training/test split provided to the participants met best-practice quality standards.

5 CONCLUSIONS

In this deliverable, we detailed our endeavours in developing the datasets used in the context of the iDPP@CLEF 2022 Lab. The lab focused on three subtasks concerning the progression of the ALS disease in patients enrolled in two European countries, Italy and Portugal. In particular, the tasks consisted in predicting the risk (or time) that an ALS patient will need medical intervention under the form of NIV or PEG or the risk of patient's death. To obtain the dataset, we started from the knowledge base constructed by organizing the BRAINTEASER data according to the ontology defined in the deliverable 9.1 [BDD+21]. The knowledge base mentioned above contains the whole BRAINTEASER data, including those not suited to the prediction task. In this deliverable thus, we first describe the steps needed to filter BRAINTEASER data: we first removed likely noisy and wrong records, maintaining only reliable information. More in detail, we removed records containing unordered sequences of events (e.g.., onset after diagnosis or visits, death before other events).

Similarly, we filtered data concerning ALSFRS-R questionnaires and spirometry visits that did not provide enough information to be used during the iDPP@CLEF lab. We dropped ALSFRS-R questionnaires that did not report a value for each of the 12 (revised) questionnaire items. Similarly, we removed spirometries not indicating the FVC percentage, usually considered an essential piece of information in predicting the progression of ALS. Then, we created three datasets for each of the iDPP@CLEF2022 subtasks. The datasets differ for the type of outcome that needs to be predicted and for the individuals that are considered feasible to be included in each dataset. In conclusion, our datasets contain respectively 1804 records (Dataset A: predict NIV, Death or Censoring event), 2145 records (Dataset B: predict PEG, Death or Censoring event) and 2250 records (Dataset C: predict Death or Censoring event).

According to the expected development of the BRAINTEASER project, future steps involve introducing environmental and biometric data, including data from other research centres, and developing similar datasets dedicated to Multiple Sclerosis.

6 REFERENCES

REFERENCES	
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)
CLH+09	Chio, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., Traynor, B.G., Consortium, E., et al.: Prognostic factors in als: a critical review. Amyotrophic lateral sclerosis 10(5-6), 310–323 (2009)
CSM+99	Cedarbaum, J.M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., Nakanishi, A.: The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. Journal of the Neurological Sciences 169 (1–2), 13–21 (October 1999)
FFH+22	Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.): CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073 (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.
KZN+15	Küffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L., Li, G., Fang, L., Mackey, L., Hardiman, O., Cudkowicz, M., Sherman, A., Ertaylan, G., Grosse-Wentrup, M., Hothorn, T., van Ligtenberg, J., Macke, J.H., Meyer, T., Schölkopf, B., Tran, L., Vaughan, R., Stolovitzky, G., Leitner, M.L.: Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nature Biotechnology 33(1), 51–57 (January 2015)