ALS-TDI HACKATHON

Task:

1. **Understanding and associating Progression rate data**

Associate clinical progression metrics and rates of progression, including ALS-FRS-R, normalized accelerometry, and ML-based speech dysarthria scores with levels and changes in 7500 proteins measured in blood samples from people with ALS. Then, following that, apply gene ontology terms to reveal biological pathways most closely associated with slow vs fast disease progression.

1. **Identifying subsets**

Analyze the 7500 proteins measured in blood to cluster patient subsets following a principal component analysis. Following the PCA, then test for any clinical similarities using disease progression metrics.

Data:

The ALS Research Collaborative (ARC) contains data on blood samples, ALS-FRS progression, speech, and motion metrics for ALS patients. More about ARC can be found here; [ALS Research Collaborative | ALS Therapy Development Institute](https://www.als.net/arc/)

Participant data is voluntarily captured over time. As of 2023:4, each participant may have zero to eight blood tests, zero to x accelerometer data sets, zero to x FRS rating, and zero to x speech metric datasets.

Blood Draw:

The blood draw is a comprehensive metabolic panel (CMP) which includes the following protein and protein/enzyme related measures:

* albumin,
* globulin,
* total protein,
* alkaline phosphatase (ALP), an enzyme mostly found in the bones and liver,
* alanine aminotransferase (ALT), an enzyme found in the liver,
* aspartate aminotransferase (AST), an enzyme found in the liver+,
* bilirubin, which is waste resulting from the breakdown of red blood cells that the liver filters.

The total protein test is a blood test that assesses protein levels in the body. Age, diet, and other factors can affect protein levels, but high or low levels may also be a sign of inflammation, liver disease, and other health conditions. The normal range for protein levels in blood serum is 6 to 8 grams per deciliter (g/dl). Of this, albumin makes up 3.5 to 5.0 g/dl, and the rest is total globulins. The albumin to globulins ratio (A/G ratio) is typically between 1 .0 and 2.0.

There are other methods to measure specific proteins in blood, such as enzyme-linked immunosorbent assay (ELISA), western blot, and mass spectrometry. These methods are more sensitive and specific than the total protein test and can help identify the levels of specific proteins in the blood.

Levels of the TDP-43 protein in spinal fluid show promise as biomarkers for amyotrophic lateral sclerosis (ALS) and may aid in diagnosing this neurological disease, according to a new review of several published studies.

More information on blood proteins can be found in these links.

[Plasma Protein Tests: Purpose, Results, and More (healthline.com)](https://www.healthline.com/health/plasma-protein-tests)

[Total protein test: Normal levels, uses, results, and procedure (medicalnewstoday.com)](https://www.medicalnewstoday.com/articles/325320)

[Understanding Low or High A/G Ratio Results (healthline.com)](https://www.healthline.com/health/a-g-ratio-high#who-should-get-tested)

[Prognostic Effect of Albumin-to-Globulin Ratio in Patients with solid tumors: A Systematic Review and Meta-analysis - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5706002/)

[TDP-43 Protein's 'Critical' Impact in ALS, FTD Uncovered by Scientists (alsnewstoday.com)](https://alsnewstoday.com/news/scientists-find-how-tdp-43-affects-important-nerve-protein-als/)

FRS Data:

ALSFRS-R measures physical function in activities of daily living for patients with ALS. The measure consists of 12 questions, each scored out of four points, with a maximum score of 48 and a minimum score of 0. A score of 48 represents an absence of the measured symptoms of ALS. A score of 0 represents the worst performance across each of the measured dimensions. The score may be used to monitor ALS progression through serial measurements.

Voice Data:

Voice data scores generated by predictive machine learning model. Score is based on "I owe you a yo-yo today" phrases only. Values range from 0 (poorest) to 4 (best)

Accelerometer Data:

Need TDI input.

Characteristics of ARC Blood Draw data:

ARC blood draw data was accessed via Looker and downloaded to a csv dataset on November 13, 2023.

Both Excel and Jupyter notebooks were used to review and characterize the data. Selected data were:

* Blood Draw Data Total Protein (g/dL)
* Blood Draw Data Appointment Date
* Draw Data Participant ID
* Blood Draw Data A/G Ratio

The dataset contains 540 entries(rows) and 420 entries contained protein and A/G ratios. Based on expected ranges for Protein levels and A/G ratios, these entries were excluded.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Record | ID | Date | Protein | A/G Ratio |
| 182 | 1389 | 11/11/2022 | 68 | 1.7 |
| 234 | 1422 | 7/11/2022 | 71 | 1.3 |
| 331 | 1789 | 9/20/2021 | 734 | 1.5 |
| 233 | 1796 | 7/12/2022 | 68 | 1.6 |
| 181 | 1927 | 11/16/2022 | 76 | 1.8 |
| 136 | 1944 | 3/22/2023 | 69 | 1.8 |
| 83 | 1957 | 7/10/2023 | 71 | 1.8 |
| 119 | 1961 | 5/1/2023 | 67 | 1.2 |
| 41 | 2057 | 9/13/2023 | 66 | 1.9 |
|  | ID | Date | Protein | A/G Ratio |
| 43 | 571 | 9/13/2023 | 7.1 | 16 |
| 173 | 1937 | 11/30/2022 | 7.1 | 16 |
| 118 | 1974 | 5/2/2023 | 7.6 | 16 |
| 45 | 1983 | 9/11/2023 | 7 | 18 |

When grouped by ID, there are 176 participant IDs in the data. A histogram by number of blood draws is shown below. Most participants have three or fewer blood draws. After excluding aforementioned suspicious data, most protein levels are in the expected 6-7 g/dl level.

A graph of a number of different numbers

Description automatically generated with medium confidence

A histogram of the A/G ratios mimics the protein counts as expected and shows most measures in the 1.0 to 2.0 range.

A graph of a number of numbers

Description automatically generated with medium confidence

The final two charts show the trends over time for protein levels and A/G ratios. Observations were filtered by participants with six or more blood draws.

A chart of different types of graphs

Description automatically generated with medium confidence

A group of graphs with different colored lines

Description automatically generated