**Cognitive Digital Biomarkers from Automated Transcription of Spoken Language - Supplementary Materials**

**Supplementary Table 1**. Codes used to represent each neuropsychological subtest of interest.

|  |  |  |
| --- | --- | --- |
| Test | Code | Description |
| Logical Memory | S | Anna Thompson Immediate Recall |
|  | SA | Anna Thompson Questions |
|  | SR | Anna Thompson Delayed Recall |
| Verbal Pair Associates | 1R | Paired Recall 1 |
|  | 2R | Paired Recall 2 |
|  | 3R | Paired Recall 3 |
| Digit Span | BN | Backward Numbers |
|  | N | Forward Numbers |
|  | Total | Total transcription (participant and proctor) |

**Supplementary Table 2**. Paralinguistic variables derived from automated transcription.

|  |  |
| --- | --- |
| count\_common\_words | Number of common words |
| frac\_common\_words | Fraction of common words |
| frac\_stop\_words | Fraction of stop words |
| n\_words | Number of words |
| pause\_time | Time spent in pauses |
| pause\_fraction | Fraction of pauses |
| section\_time | Time spent in task section |
| speaking\_time | Speaking time |
| speaking\_fraction | Speaking fraction |
| word\_mean\_certainty | Mean word transcription confidence |
| word\_std\_certainty | Standard deviation of word transcription confidence |
| word\_mean\_time | Mean word utterance time |
| word\_std\_time | Standard deviation of word utterance time |

**Supplementary Table 3.** Top 10 production rules for linguistic data.

|  |  |
| --- | --- |
|  Rule | Frequency (%) |
| NP → PRP | 9.1 |
| S → NP VP | 7.6 |
| S1 → NP | 5.5 |
| S1 → S | 5.1 |
| NP → NN | 3.8 |
| NP → DT NN | 3.1 |
| INTJ → UH | 2.9 |
| PP → IN NP | 2.8 |
| NP → NNP | 2.3 |
| ADVP → RB | 2.3 |

**Supplementary Table 4.** Demographic data for recordings from participants that were automatically transcribed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | CN (n = 35) | MCI (n = 55) | P vs CN\* | Dementia (n = 110) | P vs CN\* | P vs MCI\* |
| Age, mean (SD) | 82.6 (7.2) | 81.7 (7.9) | 0.65 | 83.3 (6.4) | 0.58 | 0.22 |
| Sex, n (%) |  |  | 0.83 |  | 0.69 | 0.40 |
| Female | 21 (60) | 31 (56) |  | 70 (64) |  |  |
| Male | 14 (40) | 24 (44) |  | 40 (36) |  |  |

\*Fisher’s Exact test for categorical data and the *t*-test for continuous data. CN = cognitively normal, MCI = mild cognitive impairment.

**Supplementary Table 5.**  Demographic data for recordings from participants with manually transcribed recordings.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CN (n = 86) | MCI or dementia (n = 41) | P*\** |
| Age, mean (SD) | 67.1 (16.6) | 82.2 (7.7) | < 0.001 |
| Sex, n (%) |  |  | 0.012 |
| Female | 42 (49) | 30 (73) |  |
| Male | 44 (51) | 11 (27) |  |

\*Fisher’s Exact test for categorical data and the *t*-test for continuous data. CN = cognitively normal, MCI = mild cognitive impairment.

**Supplementary Table 6.**  Demographic data at the time of recordings for those with manual transcriptions, excluding those age 67 or younger.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CN (n = 44) | MCI or dementia (n = 41) | P*\** |
| Age, mean (SD) | 80.4 (7.6) | 82.2 (7.7) | 0.28 |
| Sex, n (%) |  |  | 0.04 |
| Female | 22 (50) | 30 (73) |  |
| Male | 22 (50) | 11 (27) |  |

\*Fisher’s Exact test for categorical data and the *t*-test for continuous data. CN = cognitively normal, MCI = mild cognitive impairment.

**Supplementary Table 7.** Demographic data from participants at the time of recordings that were both manually and automatically transcribed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CN (n = 22) | MCI or dementia (n = 37) | P\* |
| Age, mean (SD) | 82.0 (7.7) | 81.9 (7.5) | 1.0 |
| Sex, n (%) |  |  | 0.17 |
| Female | 12 (55) | 27 (73) |  |
| Male | 10 (45) | 10 (27) |  |

\*Fisher’s Exact test for categorical data and the *t*-test for continuous data. CN = cognitively normal, MCI = mild cognitive impairment.

**Supplementary Fig. 1.** Predicted curve for recordings from participants with dementia, using the adjusted “Acoustic Watson” logistic regression model discriminating cognitively normal controls from those with mild cognitive impairment. The Y-axis represents the histogram count of subjects in the test set.

****

**Supplementary Fig. 2.** Predicted curve for recordings from cognitively normal controls, using the adjusted “Acoustic Watson” logistic regression model discriminating participants with mild cognitive impairment from those with dementia. The Y-axis represents the histogram count of subjects in the test set.

