Darmiyan®

#00314

Cognitive assessment and clinical trials

Inter-Site Variability and Standardization of AD and MCI Diagnoses

Nicolas Pannetier, Thomas Liebmann, Elham Khosravi, Pavan Krishnamurthy, Padideh Kamali-Zare, & **Kaveh Vejdani***

Darmiyan, Inc - San Francisco (United States)

kvejdani@darmiyan.com

BACKGROUND

Grading the stages of Alzheimer's Disease (AD) from Mild Cognitive Impairment (MCI) to dementia is of critical importance for understanding and modeling the disease course, providing accurate clinical diagnosis, and enabling therapeutic effect monitoring. Coming to a reliable and accurate diagnosis is challenging, especially in multicentric studies and in cases where there is clinical uncertainty about the presence of dementia. Guidelines have been put in place (e.g. ADNI) to help standardize assessments based on cognitive tests, but the final diagnosis is still subject to a clinician's interpretation and subjectivity. In a multicentric context, this variability introduces site-specific dependency that reduces the statistical reliability of subsequent analyses. In addition, when combining clinical trials, the inconsistency in diagnostic labels leads to uncertainty in the quantification of disease progression.

OBJECTIVES

1) To characterize the inter-site variability in diagnosing MCI and AD.
2) To propose a data-driven approach to data standardization.

METHODS

Cognitive scores from ADNI1/GO/2/3 databases were first aggregated and a multi-step analysis was performed:

1) To quantify the clinician assessment variability, z-scores for MCI and AD cases were computed for each cognitive summary score at subject baseline, then averaged to obtain a mean z-score for each subject.
The cognitive tests included summary scores from CDR, MMSE, ADAS, MOCA, FAQ, CCI and subcategories of the Neuropsychological Battery. The mean z-score serves as a summary measure of all cognitive testing scores available to the clinician. Comparison was made across 69 different sites and was adjusted for age, gender, education and APOE using analysis of covariance (ANCOVA). The optimal mean z-score cutoff between MCI and AD classes was also computed.

2) A novel classifier (dx-labeler) was built on ⅔ of the data (2242 CN, 2925 MCI and 1521 AD) using Darmiyan's proprietary algorithm to determine CN, MCI and AD class probabilities. The other ⅓ of the data was reserved for blind testing (1059 CN, 1476 MCI and 704 AD).
The algorithm inputs were all available cognitive scores, as well as age, gender and years of education, equivalent to what the clinician's judgement is mostly based on. Performance of the classifier was evaluated through both nested cross-validation and blind testing. Inter-site variability was re-calculated using dx-labeler's decision and compared with the variability based on clinician's decision.

RESULTS

Class	CV recall	Blind recall
Balanced	**89.8±1.2%**	**90.8%**
AD	89.7 ± 2.8 %	90
CN	95.2 ± 1.2 %	96
MCI	84.4 ± 2.0 %	87

- Highly significant difference was found for MCI between sites on mean z-score ($p<1e-10$) but was not significant for AD.
- The standard-deviation on the BA per site was 9.2%, demonstrating variability in the agreement between Darmiyan's dx-labeler and clinician assessment per site.
- Most of the mismatch cases were found in the MCI class (63%) with 57% predicted as AD and 43% predicted as CN.
- When using the ADNI diagnostic label, the distribution of the optimal mean z-score cutoffs between MCI and AD spanned over max-min=0.96 and did not follow a normal distribution (Shapiro-Wilk, $p<0.02$), demonstrating high variability in the clinician assessment of uncertain cases.
- When using the class predicted by Darmiyan's dx-labeler instead, the min/max span in the cutoff distribution was reduced by 30% and the test for normality was not rejected.

CONCLUSIONS

1) We demonstrated that, even under a well controlled protocol, diagnosis of MCI and AD can vary depending on clinician subjectivity.
2) Darmiyan's proprietary classifier algorithm (dx-labeler) shows high accuracy (90-91% BA) in the diagnostic task compared to the clinician. Interestingly, the 9-10% drop compared to ground truth (defined collectively by each clinician) was in the same range as the inter-site standard-deviation of BA, suggesting that the inter-clinician agreement and the agreement between the dx-labeler and the ground truth are alike.
3) We showed that using the dx-labeler algorithm reduces the variability in defining MCI and AD for borderline cases, helping with standardization and consistency, and is a solution to help clinicians with diagnosing uncertain cases. Finally, Darmiyan's dx-labeler can also be used to classify subjects from various cohorts and/or clinical trials, so that data from multiple data sources can be merged and mined in a consistent way and proper statistical conclusions can be drawn accordingly.