Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/16/20240577

# Analysing factors among five categories in Alzheimer's disease by multiple logistic regression

**Jinxi Yang**

Shanghai University of International Business and Economics, Shanghai, 210620, China

21024077@suibe.edu.cn

**Abstract.** Alzheimer's disease (AD) is a widespread disease around the world with significant consequences and economic burdens in nearly all countries. This study aims to integrate three main types of indicators which doctors and researchers separately used to find out patients' cognition states belongs to which stage. The dataset involves 643 participants from recorded in the ADNI website, with 88 diagnosed with Alzheimer's disease. The study firstly tests the correlation degree among ten independent variables. After careful and numerous iterations of the model, the study remains PTAU_bl, AV45_bl, WholeBrain_bl, RAVLT_learning_bl, MMSE_bl, CDRSB_bl, and ADAS13_bl in the later multiple logistic regression, and then output four formulas for each state comparing to the baseline state, which is the cognition normal (CN). The model is of good performance especially among CN, early mild cognitive impairment (EMCI) and AD, with the accuracy beyond 89%. This research makes contribution to the early identification and proactive management of individuals experiencing cognitive difficulties.

**Keywords:** Alzheimer's Disease, Multiple Logistic Regression, Influencing Factors.

## 1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder featured by various factors such as genetics and environment [1]. Its exact etiologic is still uncertain, but it has gradually become one of the most prevalent diseases among the middle-aged and elderly population in recent years, which is precisely predicted to influence more than 100 million people around the word by 2050 [2]. Clinically, the disease primarily manifests as cognitive impairments, decline in memory, abnormal personality, and changes in behaviours [3]. The disease may be performed in several stages. Without a correct diagnose and treatment, those patients are likely to end in dementia. It is estimated that around 150 million individuals are for all types of dementia [4].

The pathogenic mechanism underlying Alzheimer's disease (AD) is regarded to initiate several years prior to the recognition of AD-related cognitive impairment. This extended period of "preclinical" AD offers a vital window for potential treatment interventions. After searching papers on websites, it is clear that most recent scientific reconceptualization of the affliction has substantiated the articulation of the fundamental state of individuals with Alzheimer's disease based on relative biomarkers [5]. They suggested that biomarkers may be helpful in choosing patients for clinical traits [6]. Besides, they have delved further into how the evolving changes in biomarkers linked to AD over time impact cognitive function in the human brain. Moreover, researches suggested that regarding the selection of biomarkers,

DOI: 10.54254/2753-8818/16/20240577

one's state of being can potentially be partially elucidated by means of an individual cognitive assessment [7]. However, it should be noted that these methodologies may be constrained by the tests' disparate levels of sensitivity across different phases of the ailment [8]. Some studies listed several typical biomarkers in research, including Quantitative biomarkers such as abeta, tau, BACE1, Ng, GAP-43 and in quantitative biomarkers like RNA as well as APOE [3]. Jack et al. also proposed a model for biomarkers. describing the interrelationships between AD biomarkers and the temporal evolution between them and the onset and progression of the elicited clinical symptoms [5]. On the other side, for clinicians, due to the limitations in patients' follow-up time and the number of biomarkers available, they are often unable to directly assess a patient's cognitive status solely through biomarkers within a short timeframe. As a result, they typically employ quicker but more subjective rating scales to diagnose the patient's current cognitive condition and ascertain appropriate further treatment strategies. Martin et al. described the relationship between biomarkers and scale score considering without negative influence of time change [9]. The conclusions they showed were great, but there are still some limitations. Many studies tend to focus on a single type factor when exploring the effects of factors on the manifestation of Alzheimer's disease in patients. For instance, they solely investigate the role of biomarkers, while others solely examine the impact of cognition questionnaire assessments. However, it is important to take a more comprehensive and integrated approach by considering multiple factors simultaneously, so that doctors can gain a perspective of the complex nature of AD and its multifactorial etiologic. Therefore, comprehending research on connecting biomarkers, scale score and personal demographic features measuring fluctuations in human cognitive states is of utmost significance.

As for the methods, multiple logistic regression analysis has been widely used by authors to analyse relevant medical issues. It offers a paradigm for multivariate analysis that can periodically assess the relationship between various factors and disease (or all events) after accounting for the impact of other factors [10]. Taku et al. used it to recognize the relation between postoperative severity and factors such as sex, age, type of surgery and so on [11]. Hwang et al. applied it to study the risk factors affecting survival to hospital discharge [12]. Yokoyama et al. demonstrated that the signals of hand-muscle activities associated with finger-oriented tasks that are densely placed have a specific potential for investigation in the logistic regression models [13].

To conclude, after consideration and optimization, this paper aims to further combine the biomarkers' value and scale scores as well as volumetric measurements to describe the patient's cognitive state.

## 2. Methods

### *2.1. Data sources*

Data analysed in this research is collected form the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The data has open access to the public and was launched in April 2013 which was led by Michael W. Weiner, MD. In summarize, the dataset includes 643 records from 643 patients with unique RID (Participant roster ID). All these data below can be found in the table named ADNIMERGE-Key ADNI tables merged into one table [ADNI1, GO,2,3] from the ADNI.

### *2.2. Variable description*

As for patients' demographic features, this research selected baseline age, gender, education time, ethnicity race and their APOE gene types. Besides, there are two biomarkers used in the later studies, including β-amyloid plaques (AV45) from positron emission tomography (PET) scans and phosphorylated-tau 181(ptau) from cerebrospinal fluid (CSF). "AV45" in the chart represents the average SUVR (Standard Uptake Value Ration) of frontal, anterior cingulate, praecuneus, and parietal cortex relative to the cerebellum, which shows the total accumulation of β-amyloid plaques. The other biomarker is ptau, which many researches define it as an important factor indicating the patients' cognition state. Moreover, the study also includes multiple volumetric measurements derived from magnetic resonance imaging (MRI) scans, with ventricles in which flow cerebrospinal fluid, hippocampus, entorhinal, whole brain volumes in order to indicate their overall atrophy. Other

DOI: 10.54254/2753-8818/16/20240577

measurements commonly used in actual life like cognitive function questionnaire scores were also added in the research. The research took the Alzheimer's Disease Assessment Scale-cognitive subscale with 13 assessments (ADAS13), the Functional Assessment Questionnaire (FAQ), the Rey Auditory Verbal Learning Test's learning evaluation (RAVLT learning, measuring improvement in recalling words over 5 trials), the Mini-Mental State Examination (MMSE), the Clinical Dementia Rating Scale Sum of Boxes (CDRSB) to evaluate the potential syndrome manifestation of AD [9]. Specific variable details are in Table 1.

**Table 1.** Explanations of Factors.

| Element | Description | Logogram | Collection method |
|---|---|---|---|
| RID | Unique Participant Roter ID | | |
| DX_bl | Baseline Statement | Y | |
| AGE | Baseline Age | | |
| PTGENDER | Sex | | |
| PTEDUCAT | Education | | Demographic |
| PTETHCAT | Ethnicity | | |
| PTRACCAT | Race | | |
| APOE4 | APOE4 Gene | | |
| AV45_bl | Average AV45 SUVR Relative to the cerebellum | X1 | biomarkers |
| PTAU_bl | Ptau at Baseline | X2 | |
| Hippocampus_bl | Hippocampus at Baseline | X3 | |
| Entorhinal_bl | Entorhinal at Baseline | X4 | |
| WholeBrain_bl | Whole Brain at Baseline | X5 | |
| ADAS13_bl | ADAS13 at Baseline | X6 | |
| FAQ_bl | FAQ at Baseline | X7 | |
| RAVLT_learning _bl | RAVLT_learning at Baseline | X8 | cognition tests |
| MMSE_bl | MMSE at Baseline | X9 | |
| CDRSB_bl | CDRSB at Baseline | X10 | |

### *2.3. Research methods*

The primary approach employed for data processing and modelling is multiple logistic regression analysis, which can help to manifest the association between variable Y and many variable X.

*2.3.1. Correlation analysis.* Pearson correlation analysis was chosen to calculate the correlations among ten independent variables X1, X2…X10, helping avoid influences among independent variables in the later study.

*2.3.2. Multiple logistic regression.* Multiple logistic regression offers a method of using models to fit and predict a categorical result or dependent variable from a combination of explanatory variables for it is crucial to calculate whether other factors relevant to both variables may have some effects [14]. The odds ratio (OR) of prognostic factors is assessed by logit transformation of the development of the end point to investigate the accusation between the outcome variable and independent variables [15]. If the 95% confidence interval (CI) does not include 1, then the value if significant [16]. The model is assumed

DOI: 10.54254/2753-8818/16/20240577

as below: b is the selected benchmark group, J is the category amount, j=1,2 ⋯ J. Through calculating J sets of equations in (1), the following probability (2) in every category can be deduced.

$$\ln\left(\frac{\pi_{ij}}{\pi_{ib}}\right) = \ln\left(\frac{P(y_i = j \mid x)}{P(y_i = b \mid x)}\right) = x_i'{}'\beta_j \tag{1}$$

$$\pi_{ij} = P(y_i = j \mid x) = \frac{\exp\left(x_i'\beta_j\right)}{\sum_{m=1}^{J} \exp\left(x_i'\beta_m\right)} \tag{2}$$

As for this passage, the dependent variable is patients' baseline cognition statement, including five stages: CN, subjective memory complaints (SMC), EMCI, mild cognitive impairment (LMCI) and AD. In the study, based on the clinical separation, these five different were correspondingly regarded as number from one to five. Campbell et al. listed two main advantages using this model, it is not only an example of a multivariate approach, but also have some underlying assumptions [17], so that it doesn't require the data to fit the Normal distribution [14].

## 3. Results and Discussion

### *3.1. Descriptive results*

Table 2 shows the selected patients' baseline results. Based on all the available data, the means of all independent variables do not differ significantly and are within the same order of magnitude. Comparing the numbers in each category, the majority of patients are concentrated in the EMCI state, while the number of patients in the SMC state is minimal.

**Table 2.** Descriptive Results.

| | CN | SMC | EMCI | LMCI | AD |
|---|---|---|---|---|---|
| N | 133 | 78 | 225 | 119 | 88 |
| AGE | 73.32±5.97 | 71.97±5.35 | 71.10±7.22 | 71.40±7.23 | 73.35±8.44 |
| PTEDUCAT | 16.55±2.46 | 16.69±2.49 | 15.99±2.65 | 16.69±2.59 | 15.75±2.45 |
| AV45_bl | 1.11±0.18 | 1.14±0.20 | 1.18±0.21 | 1.31±0.23 | 1.43±0.20 |
| PTAU_bl | 21.43±8.96 | 22.42±9.74 | 24.58±13.84 | 30.03±14.36 | 39.07±15.23 |
| Hippocampus_bl | 7.52±0.81 | 7.64±0.87 | 7.26±1.04 | 6.71±1.12 | 5.99±0.94 |
| Entorhinal_bl | 3.86±0.59 | 3.87±0.57 | 3.78±0.68 | 3.40±0.70 | 2.93±0.70 |
| WholeBrain_bl | 10.51±1.03 | 10.67±1.00 | 10.70±1.05 | 10.43±1.00 | 10.15±1.16 |
| ADAS13_bl | 9.04±4.53 | 8.36±3.71 | 12.61±5.21 | 18.36±6.92 | 32.05±8.60 |
| FAQ_bl | 0.18±0.68 | 0.41±0.97 | 2.16±3.34 | 3.76±4.26 | 13.03±7.10 |
| RAVLT_learning_bl | 5.82±2.39 | 6.10±2.23 | 5.31±2.44 | 3.72±2.50 | 1.88±1.66 |
| MMSE_bl | 29.08±1.07 | 28.97±1.25 | 28.37±1.58 | 27.65±1.79 | 23.14±1.97 |
| CDRSB_bl | 0.03±0.13 | 0.08±0.18 | 1.30±0.79 | 1.77±1.01 | 4.53±1.67 |

Then, those independent viables are illustrated in the below two figures, separately shows the biomarkers and volumetric measurements in five categories. From Figure 1, The value of PTAU consistently remains at the peak among all indicators, the other four indicators' trend are almost parallel. While majority of variables in the right figure crossed together by each other, ending in ADAS13_bl at the top.

DOI: 10.54254/2753-8818/16/20240577

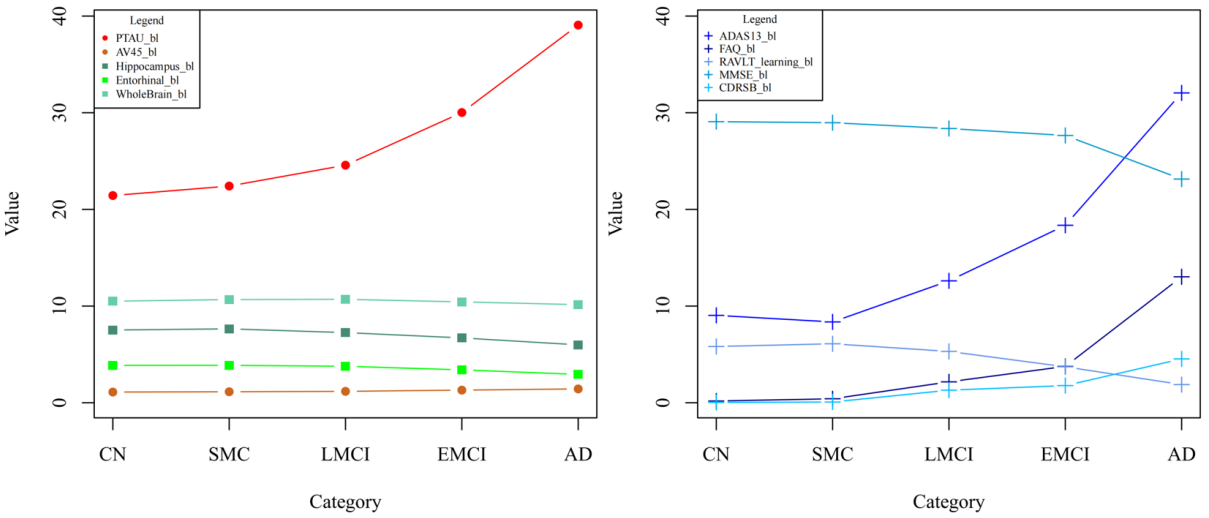


**Figure 1.** Describe by group.

To have a better look of the distribution of the most influential indicators, Figure 2 selects PTAU_bl which is showed at the left and ADA13_bl at the right. The boxplots clearly illustrate their percentages or medians and interquartile ranges.

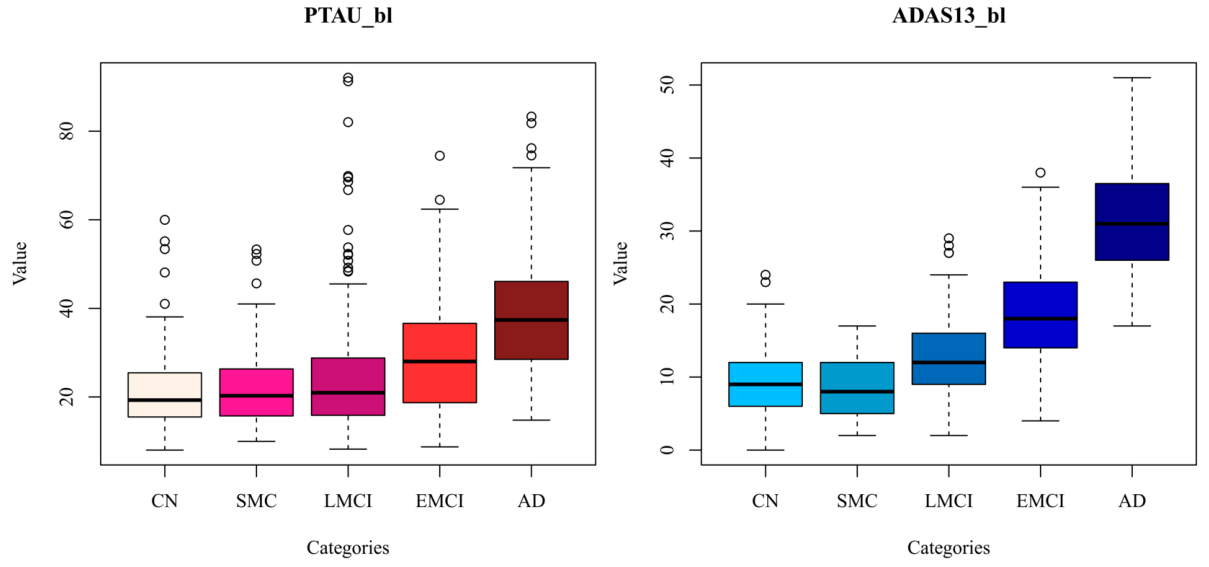


**Figure 2.** PTAU_bl and ADA13_bl Value.

### *3.2. Correlation results*

Figure 3 provides the Pearson correlation results among ten independent variables. With blue represents the lower relationship, and red stands for the stronger ones. FAQ_bl has the strongest correlation with CDRSB_bl beyond 0.8, ADAS13_bl correlated MMSE_bl and CDRSB_bl with negative affection at 0.758 and 0.747 at the positive side.

DOI: 10.54254/2753-8818/16/20240577

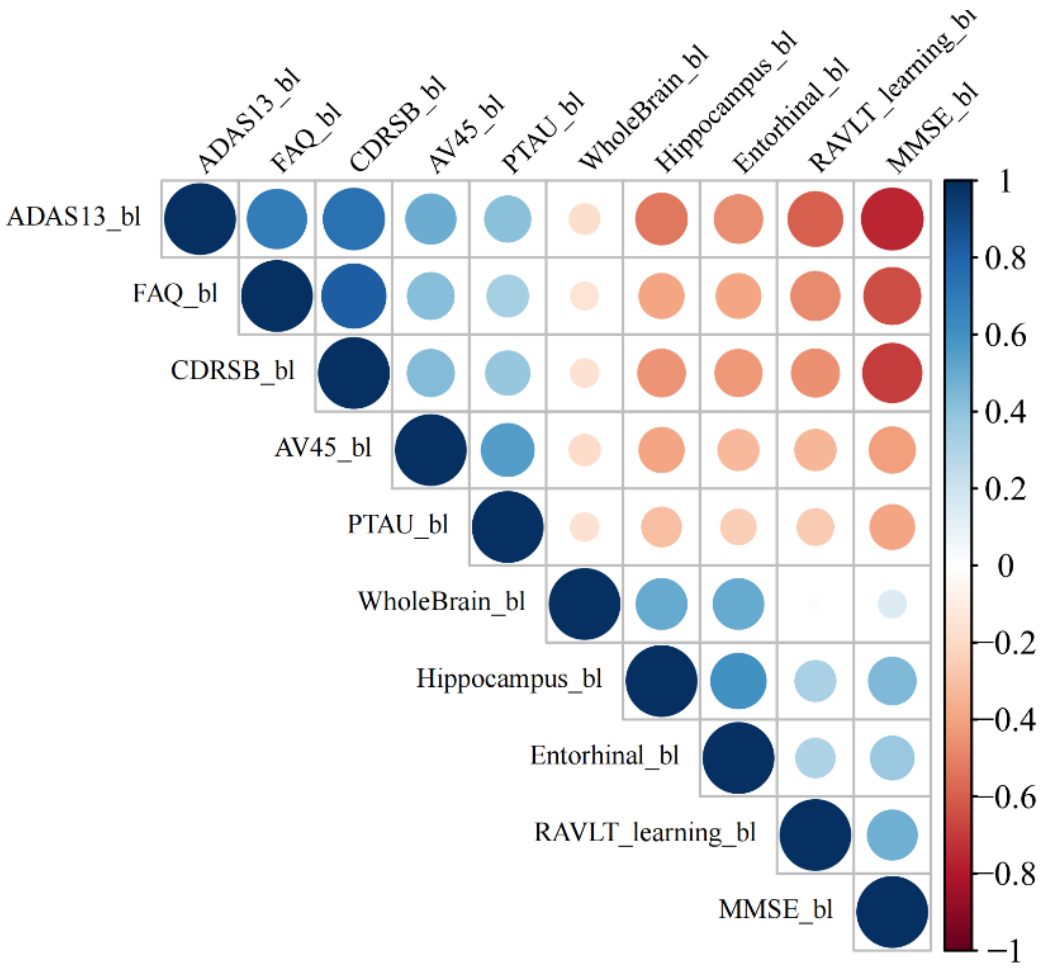


**Figure 3.** Correlation Results.

*3.3. Multiple logistic regression results*

In order to mitigate errors caused by varying magnitudes, X3 and X4 have been reduced by a factor of one thousand, and X5 is reduced by 10000. According to the correlation results listed below, Entorhinal_bl, Hippocampus_bl, ADAS13_bl, FAQ_bl, RAVLT_learning_bl, MMSE_bl, CDRSB_bl have strong correlation between each other. After numerous iterations of the model, this research has selected the optimal outcome for analysis with the lowest AIC (Akaike information criterion) value at 881.979.

The initial hypothesis for model validation in this case was whether the inclusion of independent variables (PTAU_bl, AV45_bl, WholeBrain_bl, RAVLT_learning_bl, MMSE_bl, CDRSB_bl, ADAS13_bl) would have the same impact on the model's quality. The p-value in Table 3 is less than 0.05, indicating rejection of the initial hypothesis. This suggests that the inclusion of these independent variables in the model is meaningful and valid, making this model construction significant.

**Table 3.** Descriptive Results.

| Model | $\chi^2$ | df | p | AIC | BIC |
|---|---|---|---|---|---|
| Only intercept | | | | | |
| Final model | 1154.316 | 28 | 0.000 | 881.979 | 1024.895 |

DOI: 10.54254/2753-8818/16/20240577

Table 4 displayed statistical significance in every independent variable in four groups. MMSE_bl only shows significance in AD at 0.01 level, while CDRSB_bl presents significance in all categories. Besides, ADAS13_bl in LMCI and AD is both significant. This study uses three $R^2$ to measure variables that Cox & Snell $R^2$ and Nagelkerke $R^2$ surpass 0.83.

**Table 4.** Results of Multiclass Logistic Regression Analysis.

| | SMC | EMCI | LMCI | AD |
|---|---|---|---|---|
| $X_1$ | 0.780 (0.880) | -0.658 (-0.483) | -0.033 (-0.023) | -0.041 (-0.021) |
| $X_2$ | 0.004 (0.259) | -0.001 (-0.045) | -0.000 (-0.003) | 0.013 (0.394) |
| $X_5$ | 0.222 (1.498) | 0.316 (1.114) | 0.134 (0.443) | 0.534 (1.251) |
| $X_6$ | -0.046 (-1.189) | 0.129 (1.750) | 0.250** (3.239) | 0.340** (3.689) |
| $X_8$ | 0.051 (0.751) | 0.178 (1.290) | 0.077 (0.525) | 0.201 (0.870) |
| $X_9$ | -0.139 (-1.042) | -0.262 (-1.195) | -0.244 (-1.069) | -1.281** (-3.968) |
| $X_{10}$ | 2.023* (2.044) | 14.546** (6.563) | 14.867** (6.698) | 16.034** (7.173) |
| Intercept | 0.178 (0.040) | -2.516 (-0.336) | -4.403 (-0.563) | 10.999 (1.071) |
| Likelihood Ratio Test | $\chi^2(28)$ =1154.316, p=0.000 | | | |
| | Dependent value: DX_bl | | | |
| | Cox & Snell $R^2$: 0.834 | | | |
| | Nagelkerke $R^2$: 0.875 | | | |
| | * p<0.05 ** p<0.01 | | | |

To describe independent variables more detailly, PTAU_bl, AV45_bl, WholeBrain_bl, RAVLT_learning_bl, MMSE_bl, CDRSB_bl, ADAS13_bl, were taken as independent variables, while DX_bl was taken as the dependent variable for the multicategory logistic regression analyses. Moreover, the first item of DX_bl: CN was taken as the reference item for the comparative analysed.

As opposed to CN, under the premise of SMC, all selected variables did not show significance, meaning that they all have no influence on DX_bl. As for statement EMCI, only CDRSB_bl presents significance at the 0.01 level. With Z score at 6.563, CDRSB_bl will have a significant positive influence relationship on DX_bl. Besides, the dominance ratio (OR value) is 2075998.465, which supposes that when controlling for other variables holding constant, CDRSB_bl increases by one unit, the magnitude of increase is 2075998.465 times. When considering LMCI, CDRSB_bl acts similar when it in EMCI. Moreover, ADAS13_bl also shows significance at 0.01 level, as well as having positive influence on DX_bl. With the OR value at 1.284, the change in every unit is lower than that in CDRSB_bl. Relative to CN and on the premise that AD, MMSE_bl shows significance with its coefficient at -1.281. The OR value different huge among three significant independent variables: MMSE_bl, CDRSB_bl and ADAS13_bl. OR value in CDRSB_bl is 9189579.290, while in MMSE_bl is just 0.278.

The final model equation is as follows:

$$\ln\left(\frac{SMC}{CN}\right) = 0.178 + 0.780 * X1 + 0.004 * X2 + 0.222 * X5 - 0.046 * X6 + 0.051 * X8 - 0.139 * X9 + 2.023 * X10 \quad (3)$$

$$\ln\left(\frac{\text{EMCI}}{CN}\right) = -2.516 - 0.658 * X1 - 0.001 * X2 + 0.316 * X5 + 0.129 * X6 + 0.178 * X8 - 0.262 * X9 + 14.546 * X10 \quad (4)$$

$$\ln\left(\frac{\text{LMCI}}{CN}\right) = -4.403 - 0.033 * X1 + 0.134 * X5 + 0.250 * X6 + 0.077 * X8 - 0.244 * X9 + 14.867 * X10 \quad (5)$$

DOI: 10.54254/2753-8818/16/20240577

$$\ln\left(\frac{\mathrm{AD}}{CN}\right) = +10.999 - 0.041 * X1 + 0.013 * X2 + 0.534 * X5 + 0.340 * X6 + 0.201 * X8 - 1.281 * X9 + 16.034 * X10 \quad (6)$$

The model fitness performance is relatively good since the predictive accuracy rates in Table 5 are nearly 89.47% in CN, 89.33% in EMCI and 89.77% in AD. proving that this model is good. Among five stages, the lowest predictive accuracy rate appears in SMC. This is possibly due to several reasons, such as the samples with SMC is not enough as well as SMC is not easy to discover and define

**Table 5.** Predict Accuracy Rate.

| | | predict | | | | | Predictive Accuracy Rate | Prediction error rate |
|---|---|---|---|---|---|---|---|---|
| | | CN | SMC | EMCI | LMCI | AD | | |
| Actual | CN(n=133) | 119 | 7 | 7 | 0 | 0 | 89.47% | 10.526% |
| | SMC(n=78) | 64 | 3 | 11 | 0 | 0 | 3.85% | 96.154% |
| | EMCI(n=225) | 0 | 0 | 201 | 22 | 2 | 89.33%s | 10.667% |
| | LMCI(n=119) | 0 | 0 | 65 | 44 | 10 | 36.97% | 63.025% |
| | AD(n=88) | 0 | 0 | 2 | 7 | 79 | 89.77% | 10.227% |
| | Sum | | | | | | 69.36% | 30.640% |

## 4. Conclusion

In conclusion, Multiple logistic regression showed a good performance especially among CN, EMCI and AD, with the accuracy beyond 89%. After testing the correlation of independent variables, this study removed variables with high correlation indices. Compare to other researches, instead of only concentrated on biomarkers or cognition questionnaire scores, this study integrates them comprehensively and use the multiple logistic regression to evaluate how well it measure patients' cognition states. CDRSB_bl is the most influential indicators that describe patients' cognition states with the highest significant level. This research intends to contribute to the early detection and proactive management of patients suffering from cognitive problems, with the aim of implementing targeted prevention strategies in advance and interventions to enhance health outcomes for individuals at risk.

## References

[1] Yang B H, Chen J C, Chou W H, et al. Classification of Alzheimer's Disease from 18 F-FDG and 11 C-PiB PET Imaging Biomarkers Using Support Vector Machine. Journal of Medical and Biological Engineering, 2020, 40: 545-554.

[2] Brookmeyer R, et al. 2007 Forecasting the global burden of Alzheimer's disease. Alzheimer's & dementia, 3(3), 186-191.

[3] Chen Y and Zhang B R 2014 Research progress on core biomarkers related to Alzheimer's disease. Life Sciences, 26(1), 8.

[4] Catala L F 2019 GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050. *Global Burden of Disease Study*.

[5] Jack C R, et al. 2010 Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1), 119-128.

[6] Mattsson N, et al. 2015 Revolutionizing Alzheimer's disease and clinical trials through biomarkers. *Alzheimer's & Dementia: Diagnosis*, 1(4), 412-419.

[7] Fisher C K, Smith A M and Walsh J R 2019 Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Scientific reports*, 9(1), 13622.

[8] Abuhmed T, El-Sappagh S and Alonso J M 2021 Robust hybrid deep learning models for Alzheimer's progression detection. *Knowledge-Based Systems*, 213, 106688.

DOI: 10.54254/2753-8818/16/20240577

[9] Saint J M, et al. 2023 Disease progression modelling of Alzheimer's disease using probabilistic principal components analysis. *NeuroImage*, 278, 120279.
[10] Harada H, Shiraishi K and Kato T 2001 Prognosis of acute acoustic trauma: a retrospective study using multiple logistic regression analysis. *Auris Nasus Larynx*, 28(2), 117-120.
[11] Suzuki T, et al. 2017 Predictors of postoperative outcomes of cubital tunnel syndrome treatments using multiple logistic regression analysis. *Journal of Orthopaedic Science*, 22(3), 453-456.
[12] Hwang J, et al. 2021 Prognostic factors of Veno-arterial ECMO patient. Chest, 160(4), 1138.
[13] Yokoyama M and Yanagisawa M 2019 Logistic regression analysis of multiple interosseous hand-muscle activities using surface electromyography during finger-oriented tasks. *Journal of Electromyography and Kinesiology*, 44, 117-123.
[14] Campbell M 2004 Multiple logistic regression models--what are they? *Midwifery*, 20(3), 236-239.
[15] Sonis J 2018 Odds ratios vs risk ratios. *JAMA*, 320(19), 2041-2041.
[16] Zhou S, et al. 2008 Multiple logistic regression analysis of risk factors for carcinogenesis of oral submucous fibrosis in mainland China. *International journal of oral and maxillofacial surgery*, 37(12), 1094-1098.
[17] Kirkwood B R and Sterne J A C 2010 Essential medical statistic. *John Wiley & Sons*.