The Cognitive, Age, Functioning, and Apolipoprotein E4 (CAFE) Scorecard to Predict the Development of Alzheimer's Disease: A White-Box Approach

Yumiko **Wiranto**[a*+], Devin R **Setiawan**[b+], Amber **Watts**[a,c], Arian **Ashourvan**[a,] and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]Department of Psychology, University of Kansas, Lawrence, Kansas, United States of America

[b]Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, Kansas, United States of America

[c]University of Kansas, Alzheimer's Disease Research Center, Fairway, Kansas, United States of America

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

[+] These authors contributed equally to this work.

* Corresponding author

Yumiko Wiranto

Department of Psychology, University of Kansas, 1415 Jayhawk Boulevard, Lawrence, KS 66044

Email: yumiko.wiranto@ku.edu

Phone: +1 785-864-4131

Abstract

Objective: This study aimed to bridge the gap between the costliness and complexity of diagnosing Alzheimer's disease by developing a scoring system with interpretable machine learning to predict the risk of Alzheimer's using obtainable variables to promote accessibility and early detection.

Participants and Methods: We analyzed 713 participants with normal cognition or mild cognitive impairment from the Alzheimer's Disease Neuroimaging Initiative. We integrated cognitive test scores from various domains, informant-reported daily functioning, *APOE* genotype, and demographics to generate the scorecards using the FasterRisk algorithm.

Results: Various combinations of 5 features were selected to generate ten scorecards with a test area under the curve ranging from 0.867 to 0.893. The best performance scorecard generated the following point assignments: age < 76 (-2 points); no *APOE ε4* alleles (-3 points); Rey Auditory Verbal Learning Test <= 36 items (4 points); Logical Memory delayed recall <= 3 items (5 points); and Functional Assessment Questionnaire <= 2 (-5 points). The probable Alzheimer's development risk was 4.3% for a score of -10, 31.5% for a score of -3, 50% for a score of -1, 76.3% for a score of 1, and greater than 95% for a score of > 6.

Conclusions: Our findings highlight the potential of these interpretable scorecards to predict the likelihood of developing Alzheimer's disease using obtainable information, allowing for applicability across diverse healthcare environments. While our initial scope centers on Alzheimer's disease, the foundation we have established paves the way for similar methodologies to be applied to other types of dementia.

Keywords: Alzheimer's disease; Machine learning; Cognition; Apolipoprotein ε4

Introduction

As the prevalence of Alzheimer's disease (AD) continues to rise, timely and accurate diagnosis becomes increasingly urgent. The diagnostic process for AD typically includes neurological evaluations, cognitive and functional assessments, brain imaging, cerebrospinal fluid analysis, and blood tests. However, this diagnostic approach presents challenges, such as high financial costs, invasiveness of some procedures, and limited accessibility, particularly in resource-limited or rural areas. Another significant barrier to timely diagnosis is the initial point of contact for many patients: their primary care physicians (PCPs). When individuals first notice memory-related issues, the first healthcare professional they typically go to is their PCP. However, many PCPs may not possess the specialized expertise required to identify the nuanced signs and symptoms of early AD or feel confident in delivering a conclusive diagnosis.[1, 2] As a result, patients might experience delays in obtaining appropriate care, or, in some cases, may not be referred for further evaluation at all. Therefore, the solution lies in bridging this diagnostic gap at the primary care level by developing an easily administered and interpretable method to screen for AD risk.

The advancement of machine learning models offers a vast avenue for aiding the diagnostic process due to their speed, consistency, and data-driven decisions that often excel in comparison to humans.[3] Recent efforts to develop machine learning models to assist clinicians in identifying early-stage AD, such as Convolutional Neural Networks (CNN) and Gradient Boosting Machines (GBM), have demonstrated robust accuracy.[4,5] However, the use of these models has raised important issues pertaining to interpretability. To further elucidate this point, a CNN is a type of neural network that uses image data and employs convolution layers (i.e., scanning a group of pixels) and pooling layers (i.e., size reduction) to process the image

efficiently for image classification tasks.[6] Meanwhile, A GBM is a type of machine learning algorithm that combines multiple simple models, typically decision trees, where each new tree aims to correct the errors made by the previous ones to create a powerful predictive model.[7] While CNNs and GBMs allow for accurate image categorization and model prediction, respectively, the complexity of these methods creates a "black box" effect where it becomes difficult to understand how a particular decision is made, potentially leading to a lack of trust in the outputs from clinicians.

Interpretable machine learning models (i.e., white-box approach), on the other hand, do not suffer from the same issues. Interpretable models aim to provide the "why" of outputs, offering insights into how specific features contribute to predictions and allowing for transparent and understandable decision-making processes. This transparency promotes human-computer interaction, in the case of clinical settings, trust between clinicians and the machine learning outputs.[8] Previous research has yielded reasonable accuracy in predicting the risk of a medical condition, such as epileptic seizure, using such an approach.[9]

In this study, we developed risk scores that were presented in a scorecard model to assess the risk of developing AD. Risk scores are predictive models that have been used in various fields, including medicine, to aid decision-making processes through basic mathematical calculation.[10–12] We selected the following variables to develop the scorecards due to their accessibility and comprehensive representation of factors influencing AD: demographic information, cognitive tests from various domains, daily functioning, and the apolipoprotein ε4 allele (*APOE4*). Although these variables are well-known for their contribution to AD development, many PCPs are unsure about the appropriate timing or severity level to seek further interventions. Therefore, we designed the scorecards to inform clinicians of the probable

risk of developing AD based on a patient's presentation. This could help clinicians decide when to refer patients to specialists or initiate interventions.

The scorecards in this study were constructed using the FasterRisk algorithm, a recent advancement that significantly improves the creation of high-quality risk scores.[13] Traditional methods, such as rounding logistic regression coefficients or non-data-driven approaches, often result in suboptimal risk scores that either fail to accurately capture the data's complexity or require extensive computational resources. The FasterRisk algorithm is not only computationally efficient, completing within minutes, but also provides multiple high-quality risk scores for consideration, enhancing the robustness of the model. [13] This transparency and efficiency make FasterRisk an ideal choice for developing interpretable models that clinicians can trust and easily use in primary care settings to improve the timely diagnosis of AD. We predicted that our framework could generate a scoring system with robust predictive power using accessible variables.

Materials and methods

Participants

We included data from 713 baseline visits from all the Alzheimer's Disease Neuroimaging Initiative cohorts (ADNI 1, 2, GO, and 3) as of August 2023. ADNI is a multi-site study that has collected clinical, biomarker, genetic, and neuroimaging data in the U.S. and Canada since 2004. ADNI's broader criteria include age 55-90, a minimum of 6 years of education, consistent medication for the past 4 weeks, Hachinski scale < 4 (to rule out vascular dementia), and Geriatric Depression Scale < 6; more information can be found www.adni-info.org. We included participants in our analyses who were classified by ADNI as having normal cognition (NC) or amnestic Mild Cognitive Impairment (aMCI). Participants classified as

NC were those with no subjective memory complaints, Mini-Mental State Exam (MMSE) scores of 24-30, Clinical Dementia Rating (CDR) of 0, and a within-normal score on the Wechsler Memory Scale Logical Memory II during screening. aMCI participants were those with subjective memory complaints, objective memory deficits indicated by neuropsychological tests, and a CDR score of 0.5. A request to access the ADNI dataset was approved for this study. Informed consent was obtained from all participants at the time of study enrollment.

For the present analyses, participants were divided into two groups: stable and progressive. The stable group consisted of individuals who remained at the same diagnosis level over time. The progressive group included those who developed AD. Specifically, participants who progressed from aMCI to AD were placed in the aMCI-AD group. Those who went from NC to aMCI and then to AD were placed in the NC-AD group. Individuals who progressed to aMCI from NC were not included in the analysis.

APOE Genotyping

APOE genotyping was performed on DNA samples obtained from subjects' blood, using an APOE genotyping kit, as described in http://www.adniinfo.org/Scientists/Pdfs/adniproceduresmanual12.pdf (also see http://www.adni-info.org for detailed information blood sample collection, DNA preparation, and genotyping methods). *APOE ε4* carriers were defined as participants with one or two copies of the *APOE ε4* allele.

Neuropsychological Tests and Functioning

We selected a range of neuropsychological tests that tapped into a variety of cognitive domains, such as attention, executive function, memory (short-term and long-term), verbal fluency, and global cognition. The selected tests were the Mini-Mental State Examination

(MMSE), Rey Auditory Verbal Learning Task (RAVLT) learning and immediate, Logical Memory delayed (LDEL), Category Animal (CATANIMSC), Trail Making Test A (TMT A), and Trail Making Test B (TMT B). These tests were selected because they were administered across all ADNI cohorts. Additionally, we included the informant-reported instrumental activities of daily living measured with the Functional Activities Questionnaire (FAQ).

Data Preprocessing

The final dataset encompasses a comprehensive set of features that play a crucial role in understanding the factors associated with the progression of the condition under investigation. The final set of features that we selected for training the FasterRisk machine learning model are age, sex, education, *APOE ε4* carrier status, MMSE, RAVLT immediate, RAVLT learning, LDEL, CATANIMSC, TMT A, TMT B, and FAQ. These features represent a combination of demographic information, cognitive assessments, informant-reported daily functioning, and a genetic marker of AD.

To prepare the data for analysis, we converted categorical variables into numerical representations through Scikit-learn Labelencoder. For 'diagnosis,' -1 represents a sample belonging to the stable group, and 1 represents an unstable group sample. For 'PTGENDER,' 0 represents female, and 1 represents male. Participants (n = 15; 2.03%) with invalid or missing values were identified and removed from the dataset. The dataset was further filtered based on the following conversion rate statistics. To be included in the stable group, the sample had to contain data indicating this diagnosis for at least 3 years to be classified as aMCI and 5 years for NC to account for the conversion rate.[14,15] This decision was based on previous studies and to exclude those who converted from normal to aMCI shortly after the initial visit. The next preprocessing step was applying binarization using the FasterRisk build-in binarization module

to convert the features from continuous into binary features (Figure 1). This ensures the proper input data format for the algorithm. All computations were performed on Python version 3.11.9 and data preprocessing was done using Numpy 1.23.5.

FasterRisk Algorithm

The FasterRisk algorithm aims to find high-quality risk scores, which have been the most popular form of the predictive model used in high-stakes decision-making.[13] It provides an interpretable set of scores that are easily understood, making each decision easier to explain. This is achieved through a three-step framework: a beam-search-based algorithm for logistic regression with bounded coefficients (for Step 1), the search algorithm to find pools of diverse, high-quality continuous solutions (for Step 2), the star ray search technique using multipliers (Step 3), and a theorem guaranteeing the quality of the star ray search.

The FasterRisk algorithm has a parameter 'k' called sparsity, which refers to the number of features with non-zero coefficients. In other words, 'k' controls the number of features in the final scorecard. The beam-search algorithm in FasterRisk operates under the assumption that one of the best models of size k implicitly contains variables from one of the best models of size k-1. It begins by selecting the best feature, constrained to a small coefficient box (e.g., [-5, 5]). Then, it iteratively adds another feature to this set, gradually building up the model. This approach allows the algorithm to focus on the most promising features without searching the entire space of possible combinations. The search algorithm in step 2 defines a tolerance gap level and generates many solutions by replacing one feature with another without affecting its performance more than the defined tolerance gap. The star ray search extends the coefficients by multiplying them to find a solution closer to an integer. This model was chosen due to its quality of solutions

and speed, which is significantly better than RiskSlim, a previous state-of-the-art model for finding risk scores.[16]

Selecting Optimal Sparsity

To select the optimal sparsity, a stratified 5-fold cross-validation is employed to find the best k-value that satisfies a given criterion (Figure 1). A range of k-values is selected, and the criteria is given to the cross-validation algorithm. The selected k-value range is 1-10, with AUC as the selection criteria. The cross-validation algorithm works by calculating the mean performance of the top 10 models for each fold and then averaging those means over the folds. This is done with all the k-values in the range, giving an estimated performance for each sparsity selection. The k-value that has the highest performance is selected as the optimal sparsity.

Evaluation Metrics

After finding the optimal sparsity value, the model is trained with the whole training set, which encompasses 80% of the data, and performance is evaluated on a test set encompassing the 20% that was left out during the training process (Figure 1). Ten optimal models were generated, along with their accuracy and area under the curve (AUC) performance on the test set. The decision to generate ten models stemmed from the need to explore a diverse range of "good" models, enabling researchers to delve into the interpretable features extracted from the ten scorecards created. While it is feasible to generate more models, ten was chosen as it allows for capturing all features present in the scorecards. Higher model counts do not significantly differ in features but can consume additional resources without commensurate benefits, thus our approach values parsimony. Importantly, the algorithm often generates different numbers of models to choose from, but we can always guarantee that 10 models will be generated and available for us at any given iteration of the experiment. A set of features and their bounds were generated and

the corresponding points to the right of it. The point is assigned when the criteria for the feature and its bounds are met. The points would then be added to obtain the final score. The score can be mapped to a percentage risk using the score-to-risk table generated by the algorithm. The accuracy metrics were calculated by assigning negative predictions whenever the risk is below 50% and assigning positive predictions whenever the risk is above 50%. The AUCs were calculated from the area of the Receiver Operating Characteristic curve (ROC curve), which represents the ability of the model to distinguish between different classes in a binary classification problem.

Comparison Against Baseline Models

We constructed multiple baseline models using common machine learning algorithms to compare the performance of our scorecard model. The baseline models are built utilizing Logistic Regression, Support Vector Classifier (SVC), and Random Forest Classifier, incorporating all available features from the dataset. These models were chosen to capture different modeling approaches to account for variation in performances across algorithms, giving us a broad range of performance values. Evaluation of these models is conducted through a 5-fold cross-validation approach, similar to how we evaluate our interpretable model to ensure fair comparison. The performance of the baseline models is assessed using the same AUC evaluation metrics employed for the interpretable model, thereby maintaining consistency across the evaluation process.

Results

Participant Characteristics

We included data from 713 participants, 200 with NC and 513 with aMCI at ADNI baseline visit. Over time, 11.5% of the former group and 54.8% of the latter group were

diagnosed with AD. The overall participant characteristics at baseline were as follows: 44.6% were female, the average age of 73.4 years, the average educational level was 16.1 years, and 53.9% did not carry the *APOE4* gene (Table 1). The average transition for the aMCI-AD and the NC-AD groups are 2.5 and 7.2 years, respectively.

Differences in Functioning and Cognitive Performance at Baseline Based on Diagnostic Group

Using t-tests, we observed that the NC-AD group exhibited poorer performance in the TMT A than those whose condition remained stable ("stable normal"), as indicated in Table 1 ($p < 0.05$). A higher score on the TMT A indicates a longer time to complete the test, which is indicative of worse performance. When comparing stable aMCI and aMCI-AD, we found that the aMCI-AD group had significantly lower performance across all cognitive tests included in the model ($p < 0.001$). In terms of functioning level measured by the FAQ, those who eventually progressed to AD showed a higher level of impairment at baseline in relation to their stable counterparts ($p < 0.01$).

Alzheimer Prediction Risk Score

Based on the FasterRisk algorithm, a sparsity level of 5 was selected for the most optimal combination for the generation of the final scorecards to predict AD development. Ten scorecards were generated with a test AUC range of 0.867 to 0.893. The scorecard with the highest test AUC (0.893) shown in Table 2 represents age equal to or less than 76.3 (-2 points); absence of an *APOE ε4* allele (-3 points); RAVLT immediate of 36 or less (4 points); LDEL of 3 or less (5 points); and FAQ of 2 or less (-5 points). Positive points indicate an elevated risk of AD, while negative points suggest a reduced risk. The probable AD development risk was 4.3% for a total score of -10, 12.5% for a score of -7, 31.5% for a score of -3, 50% for a score of -1,

76.3% for a score of 1, 87.5% for a score of 3, and greater than 95% for a score of 6, 7, or 9 (Table 2). In sum, younger age, absence of *APOE ε4* alleles, higher cognitive performance, and better daily functioning contributed to reduced AD risk. Other variations of the scorecard can be found in the Supplementary Figure 1.

Base Model Comparison

We compared our custom scorecard model with three common machine learning methods: Logistic Regression, Support Vector Classifier (SVC), and Random Forest Classifier. The Logistic Regression and SVC had an AUC score of 0.88, while the Random Forest Classifier had an AUC score of 0.89. These scores demonstrated how well these methods perform using all available features. Our scorecard model, however, only used five key features and still did well, with an AUC score of 0.872 and a range of 0.867 to 0.893. Despite the slight reduction in average AUC to 0.87 when compared to the base ML models, it is important to highlight the tradeoff made for interpretability and parsimony by utilizing only five features in our scorecard model. This compromise highlights the significance of our approach, where maintaining high predictive performance while having a sparse feature set demonstrates the model's effectiveness and practical applicability in real-world scenarios.

Discussion

Our study presents a novel approach to predicting the risk of developing AD that offers promising potential to be applied in clinical settings or in primary care by employing a set of obtainable variables, including demographics, *APOE ε4* status, informant-reported daily functioning, and cognitive performance scores. By utilizing the FasterRisk algorithm, we generated ten scorecards, each demonstrating high predictive accuracy with AUC scores ranging from 0.867 to 0.893. This range indicates a strong balance between sensitivity and specificity in

identifying individuals at risk of developing AD. All scorecards consistently included variables such as *APOE ε4*, daily functioning, and memory-related tests, suggesting the significance of these variables in determining progression to AD. These findings are consistent with existing knowledge in the literature on AD. [17–19] Age appeared as a significant predictor in six of the scorecards, while executive function (TMT B) and verbal fluency (CATANIMSC) were highlighted in fewer scorecards, reflecting the cognitive diversity observed in AD. These findings suggest that including executive function and verbal fluency in the scorecards could potentially capture cognitive decline in those who may have a slightly different presentation, highlighting the heterogeneity of AD.[20]

Furthermore, a notable observation from our study is that none of the scorecards identified TMT A, RAVLT learning, and Mini-Mental State Examination (MMSE) as reliable predictors for the development of AD. These findings could imply that cognitive domains related to attention and learning ability may not be significantly affected in the early stages of cognitive decline and that memory is the first domain to decline in individuals who later develop AD.[21] While MMSE is widely used in clinical practice for diagnosing dementia, its utility in predicting progression to AD may be limited. The MMSE primarily assesses global cognitive function and may lack the sensitivity to detect subtle cognitive changes that precede the onset of AD.[22,23]

In comparison to established models for AD diagnosis, our developed scorecard has exhibited promising performance. Fraser et al. demonstrated an accuracy of 82% utilizing only neuropsychological (NPS) variables with a larger dataset comprising 167 AD samples and 97 healthy controls.[24] When examining MCI discrimination, our scorecard, with an accuracy of 80.4% and an AUC of 0.893, remains competitive. Notably, it compares favorably with Ye et al.'s logistic regression model, which achieved AUCs of 0.77 using only NPS, 0.81 using NPS

and biological data, and 0.86 using NPS, biological, and imaging data, in the context of 142 MCI converters and 177 MCI non-converters.[25] Lastly, our scorecard has a better AUC compared to an interpretable model from an existing study that achieved an AUC of 0.86.[26] While acknowledging the nuanced differences in sample sizes, features, and methodologies across studies, our findings suggest the potential utility and efficacy of our scorecard in contributing to the field of AD diagnosis.

Despite not attaining the highest AUC in comparison to baseline models, a notable advantage of our scorecard lies in its interpretability. While some high-performing models may exhibit superior discrimination metrics, their complexity often renders them opaque in terms of feature contributions. In contrast, our scorecard's interpretability provides clinicians with a clear explanation of the specific neuropsychological and biological features influencing its predictions. This transparency promotes human-computer interaction, in this case, trust between clinicians and the machine learning outputs.[8] Furthermore, these scorecards offer flexibility in their implementation, which allows clinicians to incorporate their knowledge of expertise into the scorecards when predicting the risk of AD development. For example, in our scorecard (Figure 1), being younger than 76 years old would decrease the total points by 2. However, a 75-year-old patient is not significantly younger than 76 years old. In this case, the clinician can incorporate their judgement and assign a 0 to the age feature, indicating that being 75 years old does not decrease the risk of AD development. The balance between performance, interpretability, and flexibility positions our scorecard as a promising tool for practical clinical application, where understanding the rationale behind predictions is paramount for effective and informed decision support.

There are some limitations in our study. The scorecards generated in this study are only applicable to one type of dementia – Alzheimer's. Future work incorporating individuals who develop other types of dementia may result in different results or patterns of the scorecards. For example, a scorecard consisting of individuals with Frontotemporal Dementia (FTD) may highlight neuropsychiatric symptoms and a language feature in the card instead of memory, as seen in our study.[27,28] This future development would also better inform primary care physicians which further tests to refer their patients to confirm their diagnosis, which will cut down some costs compared to sending the patients to all tests/procedures. Additionally, the demographic composition of the ADNI sample, predominantly White and highly educated individuals, highlights the need for further validation in more diverse populations to ensure the generalizability of our findings. Regarding the accessibility of the tests that were included in our scorecard, *APOE* genotyping is primarily used in research settings and is currently not included as a routine test in healthcare settings. Changes in healthcare policy are necessary to disseminate and implement the scorecard in clinical settings.

Our study lays the groundwork for a more accessible and population-wide approach to screening for Alzheimer's disease. As the field advances, the integration of emerging and readily available biomarkers, such as blood plasma tests, holds promise for enhancing the predictive accuracy of our scorecards. Recent advancements in blood plasma biomarkers for AD, such as the measurement of amyloid-beta and tau proteins, offer a non-invasive and cost-effective method for early detection, showing promising results in correlating with traditional neuroimaging and cerebrospinal fluid markers.[29,30] Moving forward, our next objective is to validate these scorecards using an independent dataset to assess their stability and generalizability across diverse populations. Additionally, we aim to collaborate with primary

care physicians to collect both qualitative and quantitative data on the feasibility and potential impact of implementing these scorecards in routine clinical practice. This collaboration will provide valuable insights into the practical challenges and opportunities for integrating our tool into the healthcare system.

Conclusion

Our study generated a robust scoring system for predicting the likelihood of developing Alzheimer's disease using accessible and cost-efficient variables through interpretable machine learning. This framework's interpretability may aid primary care physicians in providing early detection to their patients, including those residing in resource-constrained areas.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

The authors have no funding to report.

Conflict of Interest

Scorecard to Predict Alzheimer's Disease

The authors have no conflict of interest to report.

Datasets/Data Availability Statement

Data used in the analysis were obtained from 722 the Alzheimer's Disease Neuroimaging Initiative 723 (ADNI) database (https://adni.loni.usc.edu/)

References

1. Bradford A, Kunik ME, Schulz P, et al. Missed and Delayed Diagnosis of Dementia in Primary Care: Prevalence and Contributing Factors. *Alzheimer Disease & Associated Disorders* 2009; 23: 306.

2. Koch T, Iliffe S, the EVIDEM-ED project. Rapid appraisal of barriers to the diagnosis and management of patients with dementia in primary care: a systematic review. *BMC Family Practice* 2010; 11: 52.

3. Kumar Y, Koul A, Singla R, et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Human Comput* 2023; 14: 8459–8486.

4. Helaly HA, Badawy M, Haikal AY. Deep Learning Approach for Early Detection of Alzheimer's Disease. *Cogn Comput* 2022; 14: 1711–1727.

5. Kavitha C, Mani V, Srividhya SR, et al. Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. *Front Public Health*; 10. Epub ahead of print 3 March 2022. DOI: 10.3389/fpubh.2022.853294.

6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436–444.

7. Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 2001; 29: 1189–1232.

8. Nasarian E, Alizadehsani R, Acharya UR, et al. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Information Fusion* 2024; 108: 102412.

9. Struck AF, Ustun B, Ruiz AR, et al. Association of an Electroencephalography-Based Risk Score With Seizure Probability in Hospitalized Patients. *JAMA Neurol* 2017; 74: 1419–1424.

10. Moreno RP, Metnitz PGH, Almeida E, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; 31: 1345–1355.

11. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. *Neth Heart J* 2008; 16: 191–196.

12. Than M, Flaws D, Sanders S, et al. Development and validation of the Emergency Department Assessment of Chest pain Score and 2 h accelerated diagnostic protocol. *Emerg Med Australas* 2014; 26: 34–44.

13. Liu J, Zhong C, Li B, et al. FasterRisk: Fast and Accurate Interpretable Risk Scores, http://arxiv.org/abs/2210.05846 (2022, accessed 14 June 2024).

14. Cabral C, Morgado PM, Campos Costa D, et al. Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Computers in Biology and Medicine* 2015; 58: 101–109.

15. García-Herranz S, Díaz-Mardomingo MC, Peraita H. Neuropsychological predictors of conversion to probable Alzheimer disease in elderly with mild cognitive impairment. *Journal of Neuropsychology* 2016; 10: 239–255.

16. Ustun B, Rudin C. Learning Optimized Risk Scores. *Journal of Machine Learning Research* 2019; 20: 1–75.

17. Gainotti G, Quaranta D, Vita MG, et al. Neuropsychological Predictors of Conversion from Mild Cognitive Impairment to Alzheimer's Disease. *Journal of Alzheimer's Disease* 2014; 38: 481–495.

18. Li J-Q, Tan L, Wang H-F, et al. Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies. *J Neurol Neurosurg Psychiatry* 2016; 87: 476–484.

19. Chen Y, Qian X, Zhang Y, et al. Prediction Models for Conversion From Mild Cognitive Impairment to Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Front Aging Neurosci*; 14. Epub ahead of print 7 April 2022. DOI: 10.3389/fnagi.2022.840386.

20. Martorelli M, Sudo FK, Charchat-Fichman H. This is not only about memory: A systematic review on neuropsychological heterogeneity in Alzheimer's disease. *Psychology & Neuroscience* 2019; 12: 271–281.

21. Wilson RS, Leurgans SE, Boyle PA, et al. Cognitive Decline in Prodromal Alzheimer Disease and Mild Cognitive Impairment. *Archives of Neurology* 2011; 68: 351–356.

22. Ciesielska N, Sokołowski R, Mazur E, et al. Is the Montreal Cognitive Assessment (MoCA) test better suited than the Mini-Mental State Examination (MMSE) in mild cognitive impairment (MCI) detection among people aged over 60? Meta-analysis. *Psychiatr Pol* 2016; 50: 1039–1052.

23. de Jager CA, Schrijnemaekers A-CMC, Honey TEM, et al. Detection of MCI in the clinic: evaluation of the sensitivity and specificity of a computerised test battery, the Hopkins Verbal Learning Test and the MMSE. *Age and Ageing* 2009; 38: 455–460.

24. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease* 2016; 49: 407–422.

25. Ye J, Farnum M, Yang E, et al. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurology* 2012; 12: 46.

26. Das D, Ito J, Kadowaki T, et al. An interpretable machine learning model for diagnosis of Alzheimer's disease. *PeerJ* 2019; 7: e6543.

27. Johnson DK, Watts AS, Chapin BA, et al. Neuropsychiatric profiles in dementia. *Alzheimer Dis Assoc Disord* 2011; 25: 326–332.

28. Hutchinson AD, Mathias JL. Neuropsychological deficits in frontotemporal dementia and Alzheimer's disease: a meta-analytic review. *Journal of Neurology, Neurosurgery & Psychiatry* 2007; 78: 917–928.

29. Pereira JB, Janelidze S, Stomrud E, et al. Plasma markers predict changes in amyloid, tau, atrophy and cognition in non-demented subjects. *Brain* 2021; 144: 2826–2836.

30. Risacher SL, Fandos N, Romero J, et al. Plasma amyloid beta levels are associated with cerebral amyloid and tau deposition. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 2019; 11: 510–519.

Table 1. Table of demographic and cognition data by diagnostic group

Participants characteristics	Stable normal (*n* = 177)	NC-AD (*n* = 23)	Stable aMCI (*n* = 232)	aMCI-AD (*n*= 281)
Age (years)	73.1 (6)*	75.9 (4)*	72.4 (7.5)**	74.3 (6.9)**
Education (years)	16.6 (2.6)	16 (2.8)	16 (2.8)	15.8 (2.8)
	n (%)	*n* (%)	*n* (%)	*n* (%)
Female	96 (54.2%)	13 (56.5%)	94 (40.5%)	115 (40.9%)
APOE e4 non-carriers	135 (76.3%)	12 (52.2 %)	142 (61.2 %)	95 (33.8%)
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
FAQ	0.1 (0.4)**	0.7 (2.8)**	1.5 (2.8)***	5.2 (4.9)***
Cognition				
MMSE	29.1 (1.1)	29.5 (0.6)	28 (1.7)***	27 (1.8)***
LDEL	13.7 (3)	13 (4.2)	7 (3)***	3.6 (3.1)***
TMT A	32.2 (9.7)*	37 (14.9)*	37.9 (16.1)***	46.9 (24.4)***
TMT B	76.4 (35.9)	89.5 (36.2)	95.5 (48.8)***	139 (75.1)***
RAVLT immediate	47.1 (9.6)	44.3 (9.5)	37.8 (10.5)***	28.8 (7.4)***
RAVLT learning	6.4 (2.2)	6 (2.6)	4.8 (2.5)***	3 (2.2)***
CATANIMSC	21.2 (5.1)	19.9 (5.3)	18.3 (5.1)***	15.6 (4.8)***

NC-AD = Normal cognition to Alzheimer's; aMCI = amnestic mild cognitive impairment; *APOE e4* = Apolipoprotein e4 allele; FAQ=Functional Activities Questionnaire; MMSE = Mini-Mental State Examination; LDEL = Logical Memory delayed recall; TMT = Trail Making Test; RAVLT = Rey Auditory Verbal Learning Test; CATANIMSC = Category Fluency (Animals); *M* = mean. *SD* = standard deviation.

Differences between stable normal vs. NC-AD or stable aMCI vs. aMCI-AD, * $p < 0.05$, ** $p<0.01$, *** $p<0.001$

FAQ and Cognition adjusted for age, education, and *APOE4*.

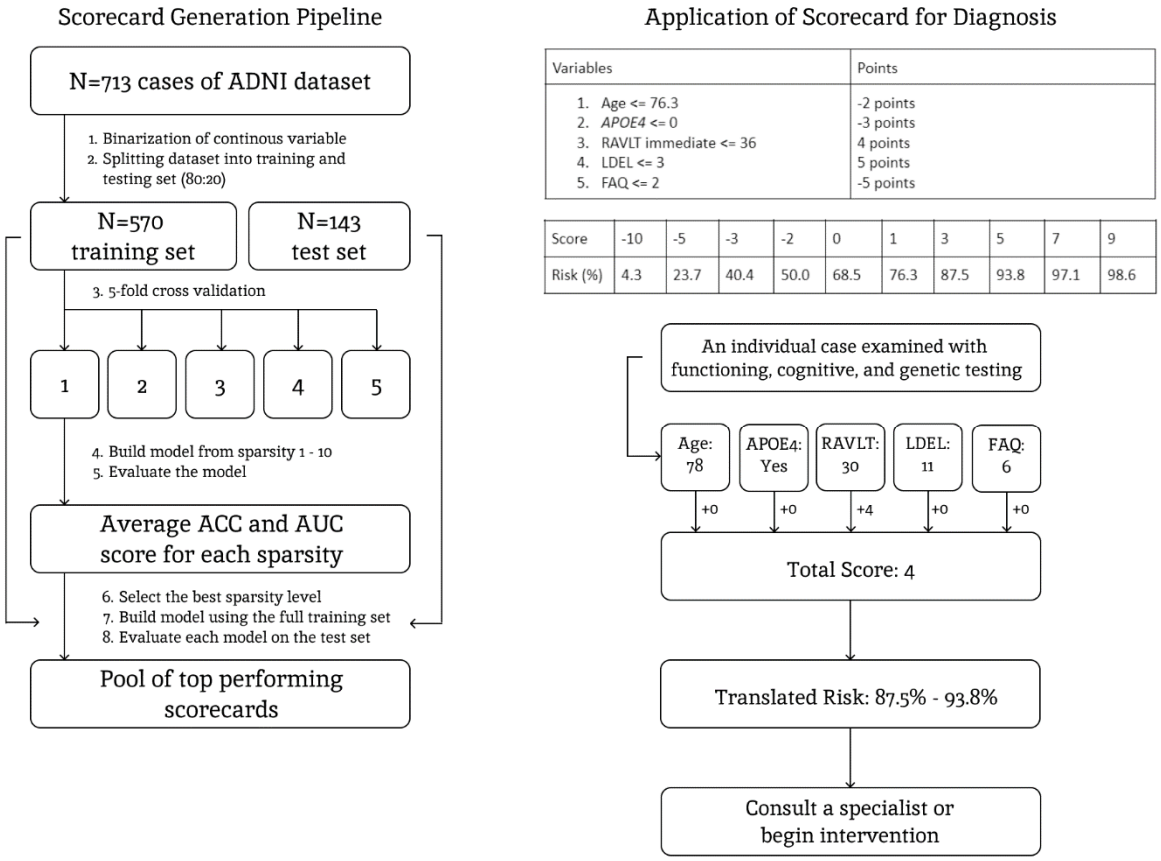
Table 2. Scorecard with the highest AUC and Risk Score to assess AD development probability

Variables	Points	
1. Age <= 76.3	-2 points	…
2. *APOE4* <= 0	-3 points	+ …
3. RAVLT immediate <= 36	4 points	+ …
4. LDEL <= 3	5 points	+ …
5. FAQ <= 2	-5 points	+
	SCORE	=

APOE4 = Apolipoprotein e4 allele; RAVLT = Rey Auditory Verbal Learning; LDEL = Logical Memory delayed recall; Test; FAQ=Functional Activities Questionnaire

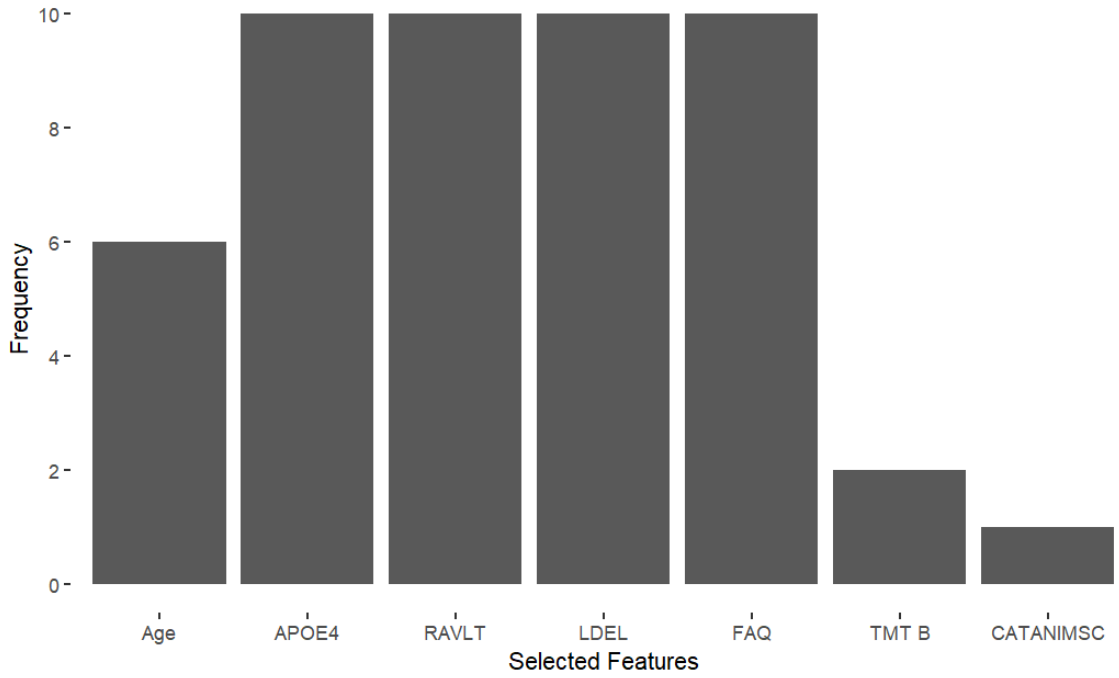
Score	-10	-5	-3	-2	0	1	3	5	7	9
Risk (%)	4.3	23.7	40.4	50.0	68.5	76.3	87.5	93.8	97.1	98.6

Figure 1. Pipeline of conducting FasterRisk algorithm to generate the CAFE scorecard and its clinical application.



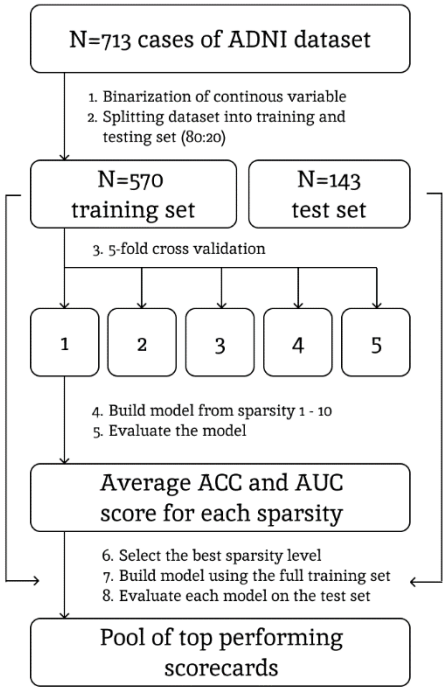
ACC = accuracy; AUC = area under the curve; *APOE e4* = Apolipoprotein e4 allele; RAVLT = Rey Auditory Verbal Learning Test; LDEL = Logical Memory delayed recall; FAQ=Functional Activities Questionnaire

Figure 2. Frequency of selected features in the 10 scorecards



APOE e4 = Apolipoprotein e4 allele; RAVLT = Rey Auditory Verbal Learning Test; LDEL = Logical Memory delayed recall; FAQ=Functional Activities Questionnaire; TMT B= Trail Making Test B; CATANIMSC = Category Fluency (Animals)

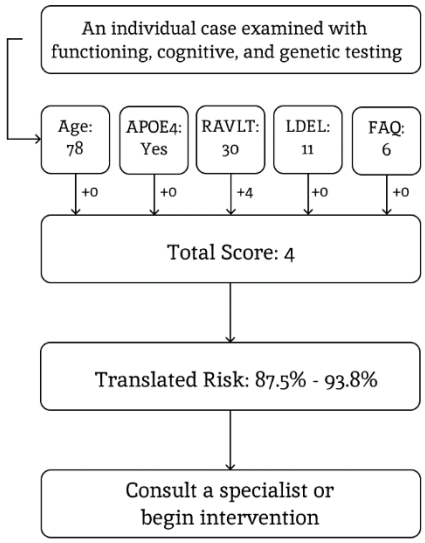
Scorecard Generation Pipeline

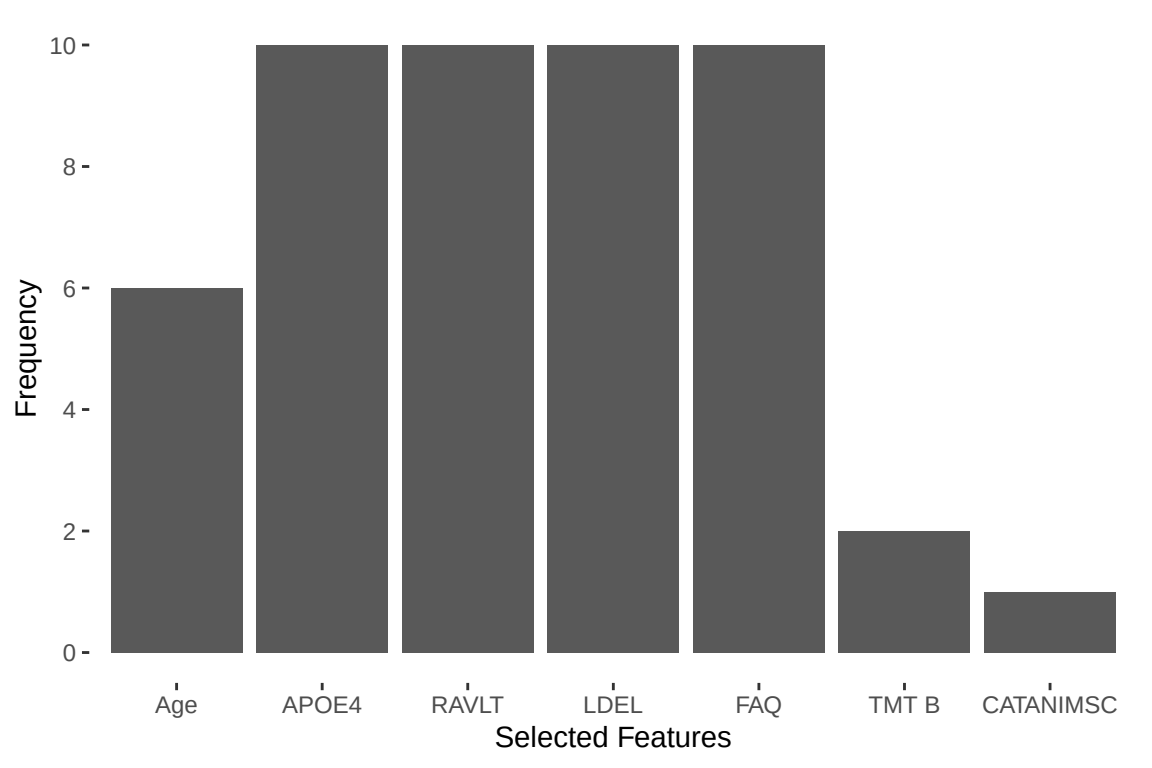


Application of Scorecard for Diagnosis

Variables	Points
1. Age <= 76.3	-2 points
2. *APOE4* <= 0	-3 points
3. RAVLT immediate <= 36	4 points
4. LDEL <= 3	5 points
5. FAQ <= 2	-5 points

Score	-10	-5	-3	-2	0	1	3	5	7	9
Risk (%)	4.3	23.7	40.4	50.0	68.5	76.3	87.5	93.8	97.1	98.6




10
8
6
4
2
0
Frequency
Age
APOE4
RAVLT
LDEL
FAQ
TMT B
CATANIMSC
Selected Features