#### Scorecard to Predict Alzheimer's Disease

4	4	
-		

2	The Cognitive, Age, Functioning, and Apolipoprotein E4 (CAFE) Scorecard to Predict the
3	Development of Alzheimer's Disease: A White-Box Approach
4	Yumiko Wiranto <sup>a*+</sup> , Devin R Setiawan <sup>b+</sup> , Amber Watts <sup>a,c</sup> , Arian Ashourvan <sup>a,</sup> and for the
5	Alzheimer's Disease Neuroimaging Initiative <sup>1</sup>
6	
7	<sup>a</sup> Department of Psychology, University of Kansas, Lawrence, Kansas, United States of America
8	<sup>b</sup> Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence,
9	Kansas, United States of America
10	<sup>c</sup> University of Kansas, Alzheimer's Disease Research Center, Fairway, Kansas, United States of
11	America
12	<sup>1</sup> Data used in preparation of this article were obtained from the Alzheimer's Disease
13	Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within
14	the ADNI contributed to the design and implementation of ADNI and/or provided data but did
15	not participate in analysis or writing of this report.
16	<sup>+</sup> These authors contributed equally to this work.
17	* Corresponding author
18	Yumiko Wiranto
19	Department of Psychology, University of Kansas, 1415 Jayhawk Boulevard, Lawrence, KS
20	66044
21	Email: yumiko.wiranto@ku.edu
22	Phone: +1 785-864-4131

Scorecard to Predict Alzheimer's Disease

2	2
~	J

### Abstract

24	Objective: This study aimed to bridge the gap between the costliness and complexity of
25	diagnosing Alzheimer's disease by developing a scoring system with interpretable machine
26	learning to predict the risk of Alzheimer's using obtainable variables to promote accessibility
27	and early detection.
28	Participants and Methods: We analyzed 713 participants with normal cognition or mild cognitive
29	impairment from the Alzheimer's Disease Neuroimaging Initiative. We integrated cognitive test
30	scores from various domains, informant-reported daily functioning, APOE genotype, and
31	demographics to generate the scorecards using the FasterRisk algorithm.
32	Results: Various combinations of 5 features were selected to generate ten scorecards with a test
33	area under the curve ranging from 0.867 to 0.893. The best performance scorecard generated the
34	following point assignments: age < 76 (-2 points); no APOE $\varepsilon 4$ alleles (-3 points); Rey Auditory
35	Verbal Learning Test <= 36 items (4 points); Logical Memory delayed recall <= 3 items (5
36	points); and Functional Assessment Questionnaire <= 2 (-5 points). The probable Alzheimer's
37	development risk was 4.3% for a score of -10, 31.5% for a score of -3, 50% for a score of -1,
38	76.3% for a score of 1, and greater than 95% for a score of $> 6$ .
39	Conclusions: Our findings highlight the potential of these interpretable scorecards to predict the
40	likelihood of developing Alzheimer's disease using obtainable information, allowing for
41	applicability across diverse healthcare environments. While our initial scope centers on
42	Alzheimer's disease, the foundation we have established paves the way for similar
43	methodologies to be applied to other types of dementia.
44	Keywords: Alzheimer's disease; Machine learning; Cognition; Apolipoprotein ε4

Scorecard to Predict Alzheimer's Disease

45

## Introduction

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

60 The advancement of machine learning models offers a vast avenue for aiding the 61 diagnostic process due to their speed, consistency, and data-driven decisions that often excel in comparison to humans.<sup>3</sup> Recent efforts to develop machine learning models to assist clinicians in 62 63 identifying early-stage AD, such as Convolutional Neural Networks (CNN) and Gradient Boosting Machines (GBM), have demonstrated robust accuracy.<sup>4,5</sup> However, the use of these 64 65 models has raised important issues pertaining to interpretability. To further elucidate this point, a 66 CNN is a type of neural network that uses image data and employs convolution layers (i.e., 67 scanning a group of pixels) and pooling layers (i.e., size reduction) to process the image

## Scorecard to Predict Alzheimer's Disease

68	efficiently for image classification tasks. <sup>6</sup> Meanwhile, A GBM is a type of machine learning
69	algorithm that combines multiple simple models, typically decision trees, where each new tree
70	aims to correct the errors made by the previous ones to create a powerful predictive model. <sup>7</sup>
71	While CNNs and GBMs allow for accurate image categorization and model prediction,
72	respectively, the complexity of these methods creates a "black box" effect where it becomes
73	difficult to understand how a particular decision is made, potentially leading to a lack of trust in
74	the outputs from clinicians.
75	Interpretable machine learning models (i.e., white-box approach), on the other hand, do
76	not suffer from the same issues. Interpretable models aim to provide the "why" of outputs,
77	offering insights into how specific features contribute to predictions and allowing for transparent
78	and understandable decision-making processes. This transparency promotes human-computer
79	interaction, in the case of clinical settings, trust between clinicians and the machine learning
80	outputs. <sup>8</sup> Previous research has yielded reasonable accuracy in predicting the risk of a medical
81	condition, such as epileptic seizure, using such an approach.9
82	In this study, we developed risk scores that were presented in a scorecard model to assess
83	the risk of developing AD. Risk scores are predictive models that have been used in various
84	fields, including medicine, to aid decision-making processes through basic mathematical
85	calculation. <sup>10–12</sup> We selected the following variables to develop the scorecards due to their
86	accessibility and comprehensive representation of factors influencing AD: demographic
87	information, cognitive tests from various domains, daily functioning, and the apolipoprotein $\epsilon 4$
88	allele (APOE4). Although these variables are well-known for their contribution to AD
89	development, many PCPs are unsure about the appropriate timing or severity level to seek
90	further interventions. Therefore, we designed the scorecards to inform clinicians of the probable

#### Scorecard to Predict Alzheimer's Disease

- 91 risk of developing AD based on a patient's presentation. This could help clinicians decide when92 to refer patients to specialists or initiate interventions.
- 93 The scorecards in this study were constructed using the FasterRisk algorithm, a recent advancement that significantly improves the creation of high-quality risk scores.<sup>13</sup> Traditional 94 95 methods, such as rounding logistic regression coefficients or non-data-driven approaches, often 96 result in suboptimal risk scores that either fail to accurately capture the data's complexity or 97 require extensive computational resources. The FasterRisk algorithm is not only 98 computationally efficient, completing within minutes, but also provides multiple high-quality risk scores for consideration, enhancing the robustness of the model.<sup>13</sup> This transparency and 99 100 efficiency make FasterRisk an ideal choice for developing interpretable models that clinicians 101 can trust and easily use in primary care settings to improve the timely diagnosis of AD. We 102 predicted that our framework could generate a scoring system with robust predictive power using 103 accessible variables.
- 104

#### Materials and methods

## 105 Participants

106 We included data from 713 baseline visits from all the Alzheimer's Disease 107 Neuroimaging Initiative cohorts (ADNI 1, 2, GO, and 3) as of August 2023. ADNI is a multi-site 108 study that has collected clinical, biomarker, genetic, and neuroimaging data in the U.S. and 109 Canada since 2004. ADNI's broader criteria include age 55-90, a minimum of 6 years of 110 education, consistent medication for the past 4 weeks, Hachinski scale < 4 (to rule out vascular 111 dementia), and Geriatric Depression Scale < 6; more information can be found www.adni-112 info.org. We included participants in our analyses who were classified by ADNI as having 113 normal cognition (NC) or amnestic Mild Cognitive Impairment (aMCI). Participants classified as

## Scorecard to Predict Alzheimer's Disease

114	NC were those with no subjective memory complaints, Mini-Mental State Exam (MMSE) scores
115	of 24-30, Clinical Dementia Rating (CDR) of 0, and a within-normal score on the Wechsler
116	Memory Scale Logical Memory II during screening. aMCI participants were those with
117	subjective memory complaints, objective memory deficits indicated by neuropsychological tests,
118	and a CDR score of 0.5. A request to access the ADNI dataset was approved for this study.
119	Informed consent was obtained from all participants at the time of study enrollment.
120	For the present analyses, participants were divided into two groups: stable and
121	progressive. The stable group consisted of individuals who remained at the same diagnosis level
122	over time. The progressive group included those who developed AD. Specifically, participants
123	who progressed from aMCI to AD were placed in the aMCI-AD group. Those who went from
124	NC to aMCI and then to AD were placed in the NC-AD group. Individuals who progressed to
125	aMCI from NC were not included in the analysis.
126	APOE Genotyping
127	APOE genotyping was performed on DNA samples obtained from subjects' blood, using
128	an APOE genotyping kit, as described in
129	http://www.adniinfo.org/Scientists/Pdfs/adniproceduresmanual12.pdf (also see http://www.adni-
130	info.org for detailed information blood sample collection, DNA preparation, and genotyping
131	methods). APOE $\varepsilon 4$ carriers were defined as participants with one or two copies of the APOE $\varepsilon 4$
132	allele.
133	Neuropsychological Tests and Functioning
134	We selected a range of neuropsychological tests that tapped into a variety of cognitive
135	domains, such as attention, executive function, memory (short-term and long-term), verbal

136 fluency, and global cognition. The selected tests were the Mini-Mental State Examination

### Scorecard to Predict Alzheimer's Disease

137	(MMSE), Rey Auditory Verbal Learning Task (RAVLT) learning and immediate, Logical
138	Memory delayed (LDEL), Category Animal (CATANIMSC), Trail Making Test A (TMT A),
139	and Trail Making Test B (TMT B). These tests were selected because they were administered
140	across all ADNI cohorts. Additionally, we included the informant-reported instrumental
141	activities of daily living measured with the Functional Activities Questionnaire (FAQ).
142	Data Preprocessing
143	The final dataset encompasses a comprehensive set of features that play a crucial role in
144	understanding the factors associated with the progression of the condition under investigation.
145	The final set of features that we selected for training the FasterRisk machine learning model are
146	age, sex, education, APOE ɛ4 carrier status, MMSE, RAVLT immediate, RAVLT learning,
147	LDEL, CATANIMSC, TMT A, TMT B, and FAQ. These features represent a combination of
148	demographic information, cognitive assessments, informant-reported daily functioning, and a
149	genetic marker of AD.
150	To prepare the data for analysis, we converted categorical variables into numerical

151 representations through Scikit-learn Labelencoder. For 'diagnosis,' -1 represents a sample 152 belonging to the stable group, and 1 represents an unstable group sample. For 'PTGENDER,' 0 153 represents female, and 1 represents male. Participants (n = 15; 2.03%) with invalid or missing 154 values were identified and removed from the dataset. The dataset was further filtered based on 155 the following conversion rate statistics. To be included in the stable group, the sample had to 156 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.<sup>14,15</sup> This decision was based on previous studies and to 157 158 exclude those who converted from normal to aMCI shortly after the initial visit. The next 159 preprocessing step was applying binarization using the FasterRisk build-in binarization module

#### Scorecard to Predict Alzheimer's Disease

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.

### 163 FasterRisk Algorithm

164 The FasterRisk algorithm aims to find high-quality risk scores, which have been the most 165 popular form of the predictive model used in high-stakes decision-making.<sup>13</sup> It provides an 166 interpretable set of scores that are easily understood, making each decision easier to explain. This

167 is achieved through a three-step framework: a beam-search-based algorithm for logistic

168 regression with bounded coefficients (for Step 1), the search algorithm to find pools of diverse,

169 high-quality continuous solutions (for Step 2), the star ray search technique using multipliers

170 (Step 3), and a theorem guaranteeing the quality of the star ray search.

171 The FasterRisk algorithm has a parameter 'k' called sparsity, which refers to the number 172 of features with non-zero coefficients. In other words, 'k' controls the number of features in the 173 final scorecard. The beam-search algorithm in FasterRisk operates under the assumption that one 174 of the best models of size k implicitly contains variables from one of the best models of size k-1. 175 It begins by selecting the best feature, constrained to a small coefficient box (e.g., [-5, 5]). Then, 176 it iteratively adds another feature to this set, gradually building up the model. This approach 177 allows the algorithm to focus on the most promising features without searching the entire space 178 of possible combinations. The search algorithm in step 2 defines a tolerance gap level and 179 generates many solutions by replacing one feature with another without affecting its performance 180 more than the defined tolerance gap. The star ray search extends the coefficients by multiplying 181 them to find a solution closer to an integer. This model was chosen due to its quality of solutions

#### Scorecard to Predict Alzheimer's Disease

and speed, which is significantly better than RiskSlim, a previous state-of-the-art model for
 finding risk scores.<sup>16</sup>

#### 184 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.

### **192 Evaluation Metrics**

193 After finding the optimal sparsity value, the model is trained with the whole training set, 194 which encompasses 80% of the data, and performance is evaluated on a test set encompassing 195 the 20% that was left out during the training process (Figure 1). Ten optimal models were 196 generated, along with their accuracy and area under the curve (AUC) performance on the test set. 197 The decision to generate ten models stemmed from the need to explore a diverse range of "good" 198 models, enabling researchers to delve into the interpretable features extracted from the ten 199 scorecards created. While it is feasible to generate more models, ten was chosen as it allows for 200 capturing all features present in the scorecards. Higher model counts do not significantly differ in 201 features but can consume additional resources without commensurate benefits, thus our approach 202 values parsimony. Importantly, the algorithm often generates different numbers of models to 203 choose from, but we can always guarantee that 10 models will be generated and available for us 204 at any given iteration of the experiment. A set of features and their bounds were generated and

## Scorecard to Predict Alzheimer's Disease

205	the corresponding points to the right of it. The point is assigned when the criteria for the feature
206	and its bounds are met. The points would then be added to obtain the final score. The score can
207	be mapped to a percentage risk using the score-to-risk table generated by the algorithm. The
208	accuracy metrics were calculated by assigning negative predictions whenever the risk is below
209	50% and assigning positive predictions whenever the risk is above 50%. The AUCs were
210	calculated from the area of the Receiver Operating Characteristic curve (ROC curve), which
211	represents the ability of the model to distinguish between different classes in a binary
212	classification problem.
213	Comparison Against Baseline Models
214	We constructed multiple baseline models using common machine learning algorithms to
215	compare the performance of our scorecard model. The baseline models are built utilizing
216	Logistic Regression, Support Vector Classifier (SVC), and Random Forest Classifier,
217	incorporating all available features from the dataset. These models were chosen to capture
218	different modeling approaches to account for variation in performances across algorithms, giving
219	us a broad range of performance values. Evaluation of these models is conducted through a 5-
220	fold cross-validation approach, similar to how we evaluate our interpretable model to ensure fair
221	comparison. The performance of the baseline models is assessed using the same AUC evaluation
222	metrics employed for the interpretable model, thereby maintaining consistency across the
223	evaluation process.
224	Results
225	Participant Characteristics
226	We included data from 713 participants, 200 with NC and 513 with aMCI at ADNI
227	baseline visit. Over time, 11.5% of the former group and 54.8% of the latter group were

#### Scorecard to Predict Alzheimer's Disease

228	diagnosed with AD. The overall participant characteristics at baseline were as follows: 44.6%
229	were female, the average age of 73.4 years, the average educational level was 16.1 years, and
230	53.9% did not carry the APOE4 gene (Table 1). The average transition for the aMCI-AD and the
231	NC-AD groups are 2.5 and 7.2 years, respectively.
232	Differences in Functioning and Cognitive Performance at Baseline Based on Diagnostic
233	Group
234	Using t-tests, we observed that the NC-AD group exhibited poorer performance in the
235	TMT A than those whose condition remained stable ("stable normal"), as indicated in Table 1 ( $p$
236	< 0.05). A higher score on the TMT A indicates a longer time to complete the test, which is
237	indicative of worse performance. When comparing stable aMCI and aMCI-AD, we found that
238	the aMCI-AD group had significantly lower performance across all cognitive tests included in
239	the model ( $p < 0.001$ ). In terms of functioning level measured by the FAQ, those who eventually
240	progressed to AD showed a higher level of impairment at baseline in relation to their stable
241	counterparts ( $p < 0.01$ ).

#### 242 Alzheimer Prediction Risk Score

243 Based on the FasterRisk algorithm, a sparsity level of 5 was selected for the most optimal 244 combination for the generation of the final scorecards to predict AD development. Ten 245 scorecards were generated with a test AUC range of 0.867 to 0.893. The scorecard with the 246 highest test AUC (0.893) shown in Table 2 represents age equal to or less than 76.3 (-2 points); 247 absence of an APOE ɛ4 allele (-3 points); RAVLT immediate of 36 or less (4 points); LDEL of 3 248 or less (5 points); and FAQ of 2 or less (-5 points). Positive points indicate an elevated risk of 249 AD, while negative points suggest a reduced risk. The probable AD development risk was 4.3% 250 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,

### Scorecard to Predict Alzheimer's Disease

251	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
252	(Table 2). In sum, younger age, absence of APOE $\varepsilon 4$ alleles, higher cognitive performance, and
253	better daily functioning contributed to reduced AD risk. Other variations of the scorecard can be
254	found in the Supplementary Figure 1.
255	Base Model Comparison
256	We compared our custom scorecard model with three common machine learning
257	methods: Logistic Regression, Support Vector Classifier (SVC), and Random Forest Classifier.
258	The Logistic Regression and SVC had an AUC score of 0.88, while the Random Forest
259	Classifier had an AUC score of 0.89. These scores demonstrated how well these methods
260	perform using all available features. Our scorecard model, however, only used five key features
261	and still did well, with an AUC score of 0.872 and a range of 0.867 to 0.893. Despite the slight
262	reduction in average AUC to 0.87 when compared to the base ML models, it is important to
263	highlight the tradeoff made for interpretability and parsimony by utilizing only five features in
264	our scorecard model. This compromise highlights the significance of our approach, where
265	maintaining high predictive performance while having a sparse feature set demonstrates the
266	model's effectiveness and practical applicability in real-world scenarios.
267	Discussion
268	Our study presents a novel approach to predicting the risk of developing AD that offers
269	promising potential to be applied in clinical settings or in primary care by employing a set of
270	obtainable variables, including demographics, APOE $\epsilon$ 4 status, informant-reported daily
271	functioning, and cognitive performance scores. By utilizing the FasterRisk algorithm, we
272	generated ten scorecards, each demonstrating high predictive accuracy with AUC scores ranging

273 from 0.867 to 0.893. This range indicates a strong balance between sensitivity and specificity in

## Scorecard to Predict Alzheimer's Disease

274	identifying individuals at risk of developing AD. All scorecards consistently included variables
275	such as APOE $\epsilon 4$ , daily functioning, and memory-related tests, suggesting the significance of
276	these variables in determining progression to AD. These findings are consistent with existing
277	knowledge in the literature on AD. <sup>17–19</sup> Age appeared as a significant predictor in six of the
278	scorecards, while executive function (TMT B) and verbal fluency (CATANIMSC) were
279	highlighted in fewer scorecards, reflecting the cognitive diversity observed in AD. These
280	findings suggest that including executive function and verbal fluency in the scorecards could
281	potentially capture cognitive decline in those who may have a slightly different presentation,
282	highlighting the heterogeneity of AD. <sup>20</sup>
283	Furthermore, a notable observation from our study is that none of the scorecards
284	identified TMT A, RAVLT learning, and Mini-Mental State Examination (MMSE) as reliable
285	predictors for the development of AD. These findings could imply that cognitive domains related
286	to attention and learning ability may not be significantly affected in the early stages of cognitive
287	decline and that memory is the first domain to decline in individuals who later develop AD. <sup>21</sup>
288	While MMSE is widely used in clinical practice for diagnosing dementia, its utility in predicting
289	progression to AD may be limited. The MMSE primarily assesses global cognitive function and
290	may lack the sensitivity to detect subtle cognitive changes that precede the onset of AD. <sup>22,23</sup>
291	In comparison to established models for AD diagnosis, our developed scorecard has
292	exhibited promising performance. Fraser et al. demonstrated an accuracy of 82% utilizing only
293	neuropsychological (NPS) variables with a larger dataset comprising 167 AD samples and 97
294	healthy controls. <sup>24</sup> When examining MCI discrimination, our scorecard, with an accuracy of
295	80.4% and an AUC of 0.893, remains competitive. Notably, it compares favorably with Ye et
296	al.'s logistic regression model, which achieved AUCs of 0.77 using only NPS, 0.81 using NPS

#### Scorecard to Predict Alzheimer's Disease

and biological data, and 0.86 using NPS, biological, and imaging data, in the context of 142 MCI
converters and 177 MCI non-converters.<sup>25</sup> Lastly, our scorecard has a better AUC compared to
an interpretable model from an existing study that achieved an AUC of 0.86.<sup>26</sup> 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.

303 Despite not attaining the highest AUC in comparison to baseline models, a notable 304 advantage of our scorecard lies in its interpretability. While some high-performing models may 305 exhibit superior discrimination metrics, their complexity often renders them opaque in terms of 306 feature contributions. In contrast, our scorecard's interpretability provides clinicians with a clear 307 explanation of the specific neuropsychological and biological features influencing its predictions. 308 This transparency promotes human-computer interaction, in this case, trust between clinicians and the machine learning outputs.<sup>8</sup> Furthermore, these scorecards offer flexibility in their 309 310 implementation, which allows clinicians to incorporate their knowledge of expertise into the 311 scorecards when predicting the risk of AD development. For example, in our scorecard (Figure 312 1), being younger than 76 years old would decrease the total points by 2. However, a 75-year-old 313 patient is not significantly younger than 76 years old. In this case, the clinician can incorporate 314 their judgement and assign a 0 to the age feature, indicating that being 75 years old does not 315 decrease the risk of AD development. The balance between performance, interpretability, and 316 flexibility positions our scorecard as a promising tool for practical clinical application, where 317 understanding the rationale behind predictions is paramount for effective and informed decision 318 support.

## Scorecard to Predict Alzheimer's Disease

319	There are some limitations in our study. The scorecards generated in this study are only
320	applicable to one type of dementia – Alzheimer's. Future work incorporating individuals who
321	develop other types of dementia may result in different results or patterns of the scorecards. For
322	example, a scorecard consisting of individuals with Frontotemporal Dementia (FTD) may
323	highlight neuropsychiatric symptoms and a language feature in the card instead of memory, as
324	seen in our study. <sup>27,28</sup> This future development would also better inform primary care physicians
325	which further tests to refer their patients to confirm their diagnosis, which will cut down some
326	costs compared to sending the patients to all tests/procedures. Additionally, the demographic
327	composition of the ADNI sample, predominantly White and highly educated individuals,
328	highlights the need for further validation in more diverse populations to ensure the
329	generalizability of our findings. Regarding the accessibility of the tests that were included in our
330	scorecard, APOE genotyping is primarily used in research settings and is currently not included
331	as a routine test in healthcare settings. Changes in healthcare policy are necessary to disseminate
332	and implement the scorecard in clinical settings.
333	Our study lays the groundwork for a more accessible and population-wide approach to
334	screening for Alzheimer's disease. As the field advances, the integration of emerging and readily
335	available biomarkers, such as blood plasma tests, holds promise for enhancing the predictive
336	accuracy of our scorecards. Recent advancements in blood plasma biomarkers for AD, such as
337	the measurement of amyloid-beta and tau proteins, offer a non-invasive and cost-effective
338	method for early detection, showing promising results in correlating with traditional
339	neuroimaging and cerebrospinal fluid markers. <sup>29,30</sup> Moving forward, our next objective is to
340	validate these scorecards using an independent dataset to assess their stability and

341 generalizability across diverse populations. Additionally, we aim to collaborate with primary

# Scorecard to Predict Alzheimer's Disease

342	care physicians to collect both qualitative and quantitative data on the feasibility and potential
343	impact of implementing these scorecards in routine clinical practice. This collaboration will
344	provide valuable insights into the practical challenges and opportunities for integrating our tool
345	into the healthcare system.
346	Conclusion
347	Our study generated a robust scoring system for predicting the likelihood of developing
348	Alzheimer's disease using accessible and cost-efficient variables through interpretable machine
349	learning. This framework's interpretability may aid primary care physicians in providing early
350	detection to their patients, including those residing in resource-constrained areas.
351	
352	
353	
354	
355	
356	
357	
358	

Scorecard to Predict Alzheimer's Disease

## 359 Acknowledgments

360 Data collection and sharing for this project was funded by the Alzheimer's Disease

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

379 Funding

380 The authors have no funding to report.

### **381 Conflict of Interest**

Scorecard to Predict Alzheimer's Disease

382 The authors have no conflict of interest to report.

## 383 Datasets/Data Availability Statement

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

Scorecard to Predict Alzheimer's Disease

386		References
387 388 389	1.	Bradford A, Kunik ME, Schulz P, et al. Missed and Delayed Diagnosis of Dementia in Primary Care: Prevalence and Contributing Factors. <i>Alzheimer Disease &amp; Associated Disorders</i> 2009; 23: 306.
390 391 392	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. <i>BMC Family Practice</i> 2010; 11: 52.
393 394 395	3.	Kumar Y, Koul A, Singla R, et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. <i>J Ambient Intell Human Comput</i> 2023; 14: 8459–8486.
396 397	4.	Helaly HA, Badawy M, Haikal AY. Deep Learning Approach for Early Detection of Alzheimer's Disease. <i>Cogn Comput</i> 2022; 14: 1711–1727.
398 399 400	5.	Kavitha C, Mani V, Srividhya SR, et al. Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. <i>Front Public Health</i> ; 10. Epub ahead of print 3 March 2022. DOI: 10.3389/fpubh.2022.853294.
401	6.	LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521: 436-444.
402 403	7.	Friedman JH. Greedy function approximation: A gradient boosting machine. <i>The Annals of Statistics</i> 2001; 29: 1189–1232.
404 405 406	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. <i>Information Fusion</i> 2024; 108: 102412.
407 408 409	9.	Struck AF, Ustun B, Ruiz AR, et al. Association of an Electroencephalography-Based Risk Score With Seizure Probability in Hospitalized Patients. <i>JAMA Neurol</i> 2017; 74: 1419–1424.
410 411 412	10.	Moreno RP, Metnitz PGH, Almeida E, et al. SAPS 3From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. <i>Intensive Care Med</i> 2005; 31: 1345–1355.
413 414	11.	Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. <i>Neth Heart J</i> 2008; 16: 191–196.
415 416 417	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. <i>Emerg Med Australas</i> 2014; 26: 34–44.
418 419	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).

Scorecard to Predict Alzheimer's Disease

- 420 14. Cabral C, Morgado PM, Campos Costa D, et al. Predicting conversion from MCI to AD
  421 with FDG-PET brain images at different prodromal stages. *Computers in Biology and*422 *Medicine* 2015; 58: 101–109.
- 423 15. García-Herranz S, Díaz-Mardomingo MC, Peraita H. Neuropsychological predictors of
  424 conversion to probable Alzheimer disease in elderly with mild cognitive impairment.
  425 *Journal of Neuropsychology* 2016; 10: 239–255.
- 426 16. Ustun B, Rudin C. Learning Optimized Risk Scores. *Journal of Machine Learning Research* 2019; 20: 1–75.
- 428 17. Gainotti G, Quaranta D, Vita MG, et al. Neuropsychological Predictors of Conversion from
  429 Mild Cognitive Impairment to Alzheimer's Disease. *Journal of Alzheimer's Disease* 2014;
  430 38: 481–495.
- 431 18. Li J-Q, Tan L, Wang H-F, et al. Risk factors for predicting progression from mild cognitive
  432 impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies.
  433 *J Neurol Neurosurg Psychiatry* 2016; 87: 476–484.
- 434 19. Chen Y, Qian X, Zhang Y, et al. Prediction Models for Conversion From Mild Cognitive
  435 Impairment to Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Front Aging*436 *Neurosci*; 14. Epub ahead of print 7 April 2022. DOI: 10.3389/fnagi.2022.840386.
- 437 20. Martorelli M, Sudo FK, Charchat-Fichman H. This is not only about memory: A systematic
  438 review on neuropsychological heterogeneity in Alzheimer's disease. *Psychology &*439 *Neuroscience* 2019; 12: 271–281.
- 440 21. Wilson RS, Leurgans SE, Boyle PA, et al. Cognitive Decline in Prodromal Alzheimer
  441 Disease and Mild Cognitive Impairment. *Archives of Neurology* 2011; 68: 351–356.
- 442 22. Ciesielska N, Sokołowski R, Mazur E, et al. Is the Montreal Cognitive Assessment (MoCA)
  443 test better suited than the Mini-Mental State Examination (MMSE) in mild cognitive
  444 impairment (MCI) detection among people aged over 60? Meta-analysis. *Psychiatr Pol*445 2016; 50: 1039–1052.
- 446 23. de Jager CA, Schrijnemaekers A-CMC, Honey TEM, et al. Detection of MCI in the clinic:
  447 evaluation of the sensitivity and specificity of a computerised test battery, the Hopkins
  448 Verbal Learning Test and the MMSE. *Age and Ageing* 2009; 38: 455–460.
- 449 24. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in 450 Narrative Speech. *Journal of Alzheimer's Disease* 2016; 49: 407–422.
- 451 25. Ye J, Farnum M, Yang E, et al. Sparse learning and stability selection for predicting MCI to
  452 AD conversion using baseline ADNI data. *BMC Neurology* 2012; 12: 46.
- 453 26. Das D, Ito J, Kadowaki T, et al. An interpretable machine learning model for diagnosis of
  454 Alzheimer's disease. *PeerJ* 2019; 7: e6543.

Scorecard to Predict Alzheimer's Disease

- 455 27. Johnson DK, Watts AS, Chapin BA, et al. Neuropsychiatric profiles in dementia. *Alzheimer*456 *Dis Assoc Disord* 2011; 25: 326–332.
- 457 28. Hutchinson AD, Mathias JL. Neuropsychological deficits in frontotemporal dementia and
  458 Alzheimer's disease: a meta-analytic review. *Journal of Neurology, Neurosurgery & amp;*459 *Psychiatry* 2007; 78: 917–928.
- 460 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.
- 462 30. Risacher SL, Fandos N, Romero J, et al. Plasma amyloid beta levels are associated with
  463 cerebral amyloid and tau deposition. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 2019; 11: 510–519.

Scorecard to Predict Alzheimer's Disease

Participants	Stable normal	NC-AD	Stable aMCI	aMCI-AD
characteristics	( <i>n</i> = 177)	( <i>n</i> = 23)	( <i>n</i> = 232)	( <i>n</i> =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 (%)	<i>n</i> (%)	n (%)	<i>n</i> (%)
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)***

466	Table 1.	Table of	demograp	hic and c	cognition	data b	y diagnostic	group
			62 1					

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 a MCI vs. a MCI-AD, \* p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001

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

Scorecard to Predict Alzheimer's Disease

able 2. Scorecard with	the highest	AUC and Risk	Score to assess	AD develo	pment	probability	1
	able 2. Scorecard with	able 2. Scorecard with the highest	able 2. Scorecard with the highest AUC and Risk	able 2. Scorecard with the highest AUC and Risk Score to assess	able 2. Scorecard with the highest AUC and Risk Score to assess AD develo	able 2. Scorecard with the highest AUC and Risk Score to assess AD development	able 2. Scorecard with the highest AUC and Risk Score to assess AD development probability

Variables	Points	
<ol> <li>Age &lt;= 76.3</li> <li>APOE4 &lt;= 0</li> <li>RAVLT immediate &lt;= 36</li> <li>LDEL &lt;= 3</li> <li>FAQ &lt;= 2</li> </ol>	-2 points -3 points 4 points 5 points -5 points	···· + ··· + ··· + ··· +
	SCORE	=

 <sup>469</sup> APOE4 = Apolipoprotein e4 allele; RAVLT = Rey Auditory Verbal Learning; LDEL = Logical
 470 Memory delayed recall; Test; FAQ=Functional Activities Questionnaire

471

\_\_\_\_\_

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

## Scorecard to Predict Alzheimer's Disease

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



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

487 Activities Questionnaire

## Scorecard to Predict Alzheimer's Disease





500 APOE e4 = Apolipoprotein e4 allele; RAVLT = Rey Auditory Verbal Learning Test; LDEL =

501 Logical Memory delayed recall; FAQ=Functional Activities Questionnaire; TMT B= Trail

502 Making Test B; CATANIMSC = Category Fluency (Animals)





#### 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



