Machine learning prediction of mild cognitive impairment and its progression to Alzheimer's disease

1 | INTRODUCTION

It is estimated that the number of people with dementia will reach 78 million by 2030 and 139 million by 2050, costing over 2.8 trillion dollars worldwide.[1] Effective screening for mild cognitive impairment (MCI) as a risk factor for developing Alzheimer's disease (AD) is a crucial step in helping aging population with their needs.[2] Early detection and automated screening for MCI and dementia could offer opportunities for deliberate study and recruitment into trials for developing other potentially useful therapeutics or interventions.[3–5] Here, we systematically compare multiple automated machine learning (ML) models in predicting MCI and its progression to AD using real-world structured and unstructured electronic health records (EHRs) data. Our objective is to comprehensively evaluate the predictive accuracy, measured by the area under the curve (AUC) of the receiver operating characteristic (ROC), for future MCI and progression to AD based on routine EHR data, among a diverse population of primary care patients aged 65 years or older.

2 | METHODS

This is a retrospective cohort study using Stanford Healthcare data from 1999 to 2022. The use of this data for this study was approved by Stanford's Institutional Review Board. Our data are formatted in the Observational Medical Outcomes Partnership (OMOP) model.[6] The cohort consists of 157,804 (MCI and non-MCI) patients, who had at least one primary care visit after reaching the age of 65; with an average age of 73 and 57.7% were females. 15.1% of patients were Asian, 6.4% were Black, 0.2% were American Indian, 0.9% were Native Hawaiian, 64.3% were White, and 13.1% had other/unknown races or declined to state their race. Our study includes two main components: (a) MCI prediction and (b) MCI to AD progression prediction. We extracted 531,387 primary care visits (for all 157,804 patients in our cohort; each patient has multiple visits) where the patients were at least 65 years old at the time of their appointment. All historical EHR records, including diagnoses, prescriptions, procedures, and clinical notes before the primary care visits, were extracted. Note clinical note features are pre-processed and

extracted in the form of standardized SNOMED structure concepts from patients' notes as part of OMOP data model.[7] The OMOP Common Data Model standardizes healthcare data for research. By standardizing the representation of patient information and healthcare data elements, OMOP enables researchers to produce reliable evidence, conduct large-scale and multisite studies, and develop predictive models using data from multiple institutions, enhancing our understanding of health outcomes and treatment effectiveness.

MCI prediction component was created using supervised ML models including logistic regression,[8] random forest,[9] and xgboost[10] to predict MCI diagnosis within 1 year of primary care visit and using 480 predictors extracted from structured and unstructured EHR data. Models were trained using data in or before 2019 and tested using data in 2020 and after. The second component, MCI to AD progression prediction model, was trained using 7425 MCI patients' data and 373 predictors extracted from structured and unstructured EHR data before MCI onset. Further, we analyzed and presented possible risk factors for progression from MCI to AD in our data.

3 | RESULTS

Table 1 shows the MCI and MCI to AD progression prediction results. Random forest was the best-performing model in predicting MCI onset as well as predicting its progression to AD. Additionally, we utilized age-stratified test data to evaluate the performance of our models. We divided our test data sets into distinct age groups (65–74, 75–85, and 85+ years old), and tested our models separately on each age group. For MCI prediction, the random forest model outperformed the other models in the age groups of 65–74 (ROC-AUC = 64.3 ± 1.2), 75–84 (ROC-AUC = 60.6 ± 1.4), and 85 years and older (ROC-AUC = 60.8 ± 2.2). Similarly, in MCI to AD progression prediction, the random forest model exhibited the highest ROC-AUC compared to all other models in the age groups of 65–74 (62.4 ± 4.1), 75–84 (58.2 ± 1.9), and 85 years and older (62.0 ± 3.6). This approach allowed us to examine the effectiveness of our models in different age cohorts, providing insights into potential age-related variations in model performance. The utilization of age-stratified data in our analysis enhances the robustness and generalizability of our findings,

TABLE 1 Performance of MCI onset prediction and progression from MCI to AD using machine learning.

Model	MCI	MCI to AD <1 year	MCI to AD <2 years	MCI to AD <3 years	MCI to AD <4 years	MCI to AD <5 years
Logistic regression	57.0 ± 1.1	55.8 ± 1.8	55.5 ± 1.2	55.2 ± 1.5	55.6 ± 1.3	55.5 ± 2.1
XGBoost	66.8 ± 0.8	62.1 ± 1.6	63.4 ± 2.0	63.3 ± 1.5	63.2 ± 1.3	63.6 ± 0.1
Random forest	68.2 ± 0.7	65.0 ± 1.7	65.8 ± 1.5	65.0 ± 1.2	64.5 ± 1.3	64.6 ± 1.4

Note: Models were tested for predicting MCI onset, progression from MCI to AD within 1, 2, 3, 4, and 5 years. Models are assessed using ROC-AUC (*c*-statistic).

TABLE 2 Prevalence of clinical factors.

Variable	Frequency ratio	*p* Value
Organic mental disorder diagnosis code	2.11	*p* < 0.001
Donepezil-related concepts in clinical notes	1.92	*p* < 0.001
Degenerative brain disorder-related concepts in clinical notes	1.86	*p* < 0.001
AD-related concepts in clinical notes	1.86	*p* < 0.001
Neuropsychological testing-related concepts in clinical notes	1.81	*p* < 0.001
Frontotemporal degeneration-related concepts in clinical notes	1.80	*p* < 0.001
Rofecoxib-related concepts in clinical notes	1.69	*p* < 0.001
Atenolol-related concepts in clinical notes	1.66	0.001
Cystocele-related concepts in clinical notes	1.64	*p* < 0.001
Aspiration of cataract by phacoemulsification-related concepts in clinical notes	1.63	*p* < 0.001

Note: Frequency ratio indicates the ratio of frequency of each variable in the MCI to AD progression group to the MCI patients with no further AD diagnosis. Rows are listed based on frequency ratio. Higher frequency ratio indicates more prevalence in MCI to AD progression group. Variables with at least 5% of frequency within both groups are presented. *p* Value is computed using Mann–Whitney test.

Abbreviations: AD, Alzheimer's disease; MCI, mild cognitive impairment.

as it accounts for potential age-related differences and enables a more nuanced understanding of our ML model's performance.

Table 2 shows the top 10 variables significantly associated with the progression from MCI to AD. The majority of these variables are the predictors extracted from patients' clinical notes. Variables related to mental health disorder diagnosis and more memory loss-related concepts in patients' clinical notes are among the top variables that are predictive of progression to AD.

4 | DISCUSSION

Given the complex nature of MCI and AD and sparsity of these events, especially at a visit-based level, random forest can detect MCI and progression to AD reasonably well. Our results also showed that clinical notes include signals that provide increased power in discriminating MCI patients who progressed to AD from MCI patients with no further AD diagnosis. Results illustrate that it is possible to predict MCI onset and AD progression with moderate levels of discrimination accuracy. This suggests an opportunity for population-wide screening mechanisms to identify patients at potential risk, who could then undergo more specific confirmatory evaluation to consider early treatment or recruitment into clinical trials. Novel elements here include the use of extracted clinical note elements that are typically underutilized in clinical risk models, which further illustrate some of the key documented features that are predictive of such important conditions.

Expected effects and utilization of this study include an automated tool for primary care providers and specialists for early detection of ADs. Automated multifactor models demonstrated superior predictive ability in assessing the risk of dementia.[11] Despite the current scarcity of clinical interventions with proven efficacy in altering the progression of MCI and dementia, the identification of individuals at risk can facilitate targeted recruitment into clinical trials, enabling the study of emerging interventions that may demonstrate effectiveness in the early stages of the disease. Furthermore, the acquisition and dissemination of personalized diagnostic evaluation strategies provide an immediately applicable approach to enhance the timely diagnostic assessment of MCI cases, enhance therapeutic approaches to postpone the AD onset,[12–14] improve care or socioeconomic factors[15,16] for the patients at risk, and facilitate the prompt identification of potentially reversible factors such as endocrine, nutritional, and infectious causes.

5 | LIMITATIONS

Note this study is limited as a single-site study; however, the models can be applied to any other site with OMOP data model. The proposed models serve as decision support systems that should be utilized under the supervision of trained healthcare providers, including primary care providers and specialists. Although the proposed ML models may not be as accurate as deliberate diagnostics such as MOCA, they are able to evaluate population-wide automatically through data systems without requiring deliberate in-person evaluation of everyone.

KEYWORDS
Alzheimer's disease, machine learning, mild cognitive impairment

DATA AVAILABILITY STATEMENT
Due to confidentiality restrictions, the data used in this study are not available for public access or sharing.

TRANSPARENCY STATEMENT
The lead author Sajjad Fouladvand, Sajjad Fouladvand affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Sajjad Fouladvand[1]
Morteza Noshad[1]
V. J. Periyakoil[2]
Jonathan H. Chen[1,3,4]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA
[2]Department of Medicine, Stanford University, Stanford, California, USA
[3]Division of Hospital Medicine, Stanford University, Stanford, California, USA
[4]Clinical Excellence Research Center, Stanford University, Stanford, California, USA

Correspondence
Sajjad Fouladvand, Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA.
Email: sajjadf@stanford.edu

ORCID
Sajjad Fouladvand http://orcid.org/0000-0002-9869-1836

TWITTER
Sajjad Fouladvand @SajjadFV

REFERENCES
1. ADI—Dementia Statistics. Accessed October 31, 2022. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/
2. Reitz C, Mayeux R. Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. *Biochem Pharmacol*. 2014;88(4):640-651. doi:10.1016/j.bcp.2013.12.024
3. Fouladvand S, Mielke MM, Vassilaki M, Sauver JSt, Petersen RC, Sohn S. Deep learning prediction of mild cognitive impairment using electronic health records. *Proc IEEE Int Conf Bioinform Biomed*. 2019;2019:799-806. doi:10.1109/bibm47256.2019.8982955
4. Zhang R, Simon G, Yu F. Advancing Alzheimer's research: a review of big data promises. *Int J Med Inform*. 2017;106:48-56. doi:10.1016/j.ijmedinf.2017.07.002
5. Kalb R, Beier M, Benedict RH, et al. Recommendations for cognitive screening and management in multiple sclerosis care. *Mult Scler J*. 2018;24(13):1665-1680. doi:10.1177/1352458518803785
6. OMOP Common Data Model—OHDSI. Accessed September 7, 2022. https://www.ohdsi.org/data-standardization/the-common-data-model/
7. OMOP NOTE_NLP Table. Accessed September 14, 2022. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:note_nlp
8. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. *JAMA*. 2016;316(5):533. doi:10.1001/jama.2016.7653
9. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
10. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16. Association for

Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

11. Stephan BCM, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the population: are screening models accurate? *Nat Rev Neurol*. 2010;6(6):318-326. doi:10.1038/nrneurol.2010.54

12. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am J Public Health*. 1998;88(9):1337-1342.

13. Zissimopoulos J, Crimmins E, St. Clair P. The value of delaying Alzheimer's disease onset. *Forum Health Econ Policy*. 2015;18(1):25-39. doi:10.1515/fhep-2014-0013

14. Park JH, Cho HE, Kim JH, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *npj Digital Med*. 2020;3(1):46. doi:10.1038/s41746-020-0256-0

15. Kim JI, Kim G. Factors affecting the survival probability of becoming a centenarian for those aged 70, based on the human mortality database: income, health expenditure, telephone, and sanitation. *BMC Geriatr*. 2014;14:113. doi:10.1186/1471-2318-14-113

16. Kin JI. *The Sociology of Longevity: Socioecological Factors of Survival Probability*. Cambridge Scholar Publishing; 2022. https://www.cambridgescholars.com/product/978-1-5275-8062-6