Acta Pharmaceutica Sinica B 2025;15(9):4427−4437

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

TOOLS

AI-powered model for accurate prediction of MCI-to-AD progression

Ahmed Abdelhameed[a], Jingna Feng[a], Xinyue Hu[a], Fang Li[a], Sori Lundin[b], Paul E. Schulz[c], Cui Tao[a,*]

[a]*Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL 32224, USA*
[b]*McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[c]*McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX 77030, USA*

Received 27 November 2024; received in revised form 29 December 2024; accepted 29 January 2025

KEY WORDS

BiGRU;
Predictive modeling;
Machine learning;
Longitudinal claim data;
Risk stratification;
Electronic health records

Abstract Alzheimer's disease (AD) remains a formidable challenge in modern healthcare, necessitating innovative approaches for its early detection and intervention. This study aimed to enhance the identification of individuals with mild cognitive impairment (MCI) at risk of developing AD. Leveraging advances in computational power and the extensive availability of healthcare data, we explored the potential of deep learning models for early prediction using medical claims data. We employed a bidirectional gated recurrent unit (BiGRU) deep learning model for predictive modeling of MCI progression across various prediction intervals, extending up to five years post-initial MCI diagnosis. The performance of the BiGRU model was rigorously compared with several machine-learning model baselines to evaluate its efficacy. Using a robust cross-validation methodology, the BiGRU emerged as the top-performing model, achieving an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.833 (95% CI: 0.822, 0.843), an Area Under the Precision-Recall Curve (AUC-PR) of 0.856 (95% CI: 0.845, 0.867), and an F1-Score of 0.71 (95% CI: 0.694, 0.724) for a five-year prediction interval. The results indicate that BiGRU, utilizing longitudinal claims data, reliably predicts MCI-to-AD progression over a lengthy interval following the initial MCI diagnosis, offering clinicians a valuable tool for targeted risk identification and stratification.

This article is part of a special issue entitled: AI, computational and systems biology-guided translational research in Alzheimer's disease published in Acta Pharmaceutica Sinica B.

*Corresponding author.
E-mail address: Tao.Cui@Mayo.edu (Cui Tao).

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.01.027

1. Introduction

Alzheimer's disease (AD) is the most common degenerative neurological disorder, characterized by a spectrum of symptoms ranging from short-term memory lapses to loss of bodily function and eventual death[1,2]. Presently, AD afflicts approximately 6.7 million Americans aged 65 and older. With an aging population and increasing life spans, the incidence and prevalence of AD are predicted to rise substantially over the next several decades. In the absence of significant medical and technological breakthroughs to prevent, arrest, or cure AD, estimates suggest that the affected population could increase to 13.8 million by the year 2060[2]. The progression of AD unfolds from subtle brain changes to significant alterations that lead to memory impairments and eventual physical incapacitation in affected individuals. The disease trajectory is delineated into three broad stages: preclinical AD, mild cognitive impairment (MCI) attributable to AD, and dementia ascribable to AD, also termed Alzheimer's dementia. The latter category is further subdivided into mild, moderate, and severe stages[3-5].

MCI identifies individuals who do not manifest dementia but exhibit cognitive abnormalities, such as performing below average on memory tasks while sustaining their activities of daily living (ADLs). The progression of MCI to AD dementia may or may not occur, a transition influenced by a variety of risk factors[6]. Identifying patients clinically diagnosed with MCI who are at heightened risk of developing AD is critical for effective risk stratification and early intervention, particularly when newly developed medications prove efficacious.

Existing studies offer diverse approaches to predict the likelihood of MCI progressing to AD and to identify the significant risk factors for this progression. Various methods have utilized distinct data modalities, including neuroimaging data such as magnetic resonance imaging (MRI)[7-9], positron emission tomography (PET)[10,11], magnetoencephalography (MEG)[12,13], as well as combinations of these modalities[14]. Although these imaging techniques have proven invaluable, their limited availability and substantial cost have prompted researchers to investigate more accessible alternatives.

Recent years have seen significant advancements in PET imaging and fluid biomarkers. PET tracers like ^{18}F-FDG assess glucose metabolism, while amyloid tracers such as ^{18}F-Florbetapir and tau-specific tracers like ^{18}F-Flortaucipir provide insights into amyloid plaques and tau tangles, respectively[15]. Similarly, while cerebrospinal fluid (CSF) biomarkers were traditionally used to confirm AD in later stages[16], blood-based biomarkers now present less invasive alternatives for early detection[17]. However, these specialized tests remain inaccessible to many patients in routine clinical practice. Despite the availability of biomarkers in neuroimaging and fluid data, detecting progression from MCI to AD in clinical practice remains challenging. To address this issue, some studies have tried to integrate neuroimaging datasets with other patient data, such as demographics and genetics[18].

More advanced models have recently employed machine learning, particularly deep learning-based techniques, to study the progression of AD by leveraging the extensive information stored in electronic health records (EHRs). These EHRs encompass a wealth of data, including, but not limited to, patient demographics, diagnoses, medications, procedures, laboratory measurements, and clinical notes. For example, Fouladvand et al.[19] utilized patient EHR data such as demographics and clinical notes, along with patient-provided information, to predict the progression from cognitively unimpaired (CU) to MCI using the long short-term memory (LSTM) deep learning model. Similarly, Gao et al.[20] combined a subset of patient EHR data with genetic data and traditional risk factors and applied machine learning techniques for AD prediction.

In clinical settings, the completeness of EHR and biomarker data exhibits significant variability across patients, attributable to factors such as differing care pathways, the frequency of clinical visits, and access to specialized testing. While some patients may possess comprehensive data including neuroimaging and biomarker measurements, many others have only basic clinical records and claims data. This inconsistency in data availability can impede the broad applicability of predictive models in real-world settings. Furthermore, it's imperative to acknowledge the temporal dimension of the data as a crucial factor in ensuring accurate risk prediction.

Despite these challenges, there has been a scarcity of research focused on utilizing EHRs to predict the complex progression from MCI to AD. For instance, Mao et al.[21] utilized unstructured clinical notes from EHRs to develop a deep learning framework based on the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, with the aim of predicting the likelihood of disease progression from MCI to AD. Recent advancements in artificial intelligence, including Large Language Models (LLMs) and foundation models, have further expanded the potential of EHR analysis. These models have demonstrated promising capabilities in processing EHR and claims data, particularly in tasks involving unstructured clinical text and complex medical knowledge representation[22-24]. While such models open new avenues for healthcare data analysis, sequence-based deep learning approaches remain especially adept at structured temporal prediction tasks, such as forecasting disease progression, particularly when dealing with longitudinal claims data[25].

The primary objective of this study is to develop a deep learning framework for predicting the progression from MCI to AD. Our approach is distinctive, as it does not rely on imaging or biomarker data but instead exclusively utilizes claims data, which is routinely available for all patients. This distinction is pivotal, as many patients do not undergo brain scans or blood tests, thereby allowing our model to encompass a broader population. The framework leverages retrospective longitudinal claim data spanning a five-year period prior to the initial diagnosis of MCI (index date). We evaluate the model's predictive accuracy across multiple time horizons: one, two, three, and five years from the index date. This approach is innovative in its ability to predict both short-term and long-term progression of MCI to AD using a comprehensive set of readily available clinical variables including, age, gender, diagnoses, medications, and procedures.

2. Materials and methods

2.1. Study cohort

In this study, we used Optum's de-identified Clinformatics® Data Mart Database, which is a database of administrative health claims for members of large commercial and Medicare Advantage health plans. We identified a cohort of 80,931 patients who received their initial MCI diagnosis between January 2007 and March 2020, all of whom were at least 40 years old. Employing Clinformatics®, we conducted comprehensive queries on claim records to extract patient information, including diagnoses, procedures, medications, and demographics both prior to and following the initial MCI diagnosis.

The MCI diagnosis was defined using the International Classification of Diseases, Ninth Revision (ICD-9) code 331.83, and

Tenth Revision (ICD-10) code G31.84[20]. As illustrated in Fig. 1, 16 patients were excluded due to a lack of gender information. Additionally, 785 patients were removed from the study after being deemed ineligible due to the absence of any records before their MCI diagnosis.

For the remaining cohort, patients were classified into case and control groups. Those patients whose MCI progressed to AD were identified using ICD-9 code 331.0 and ICD-10 codes G30.*[21], forming the case group. The control group consisted of MCI patients who did not receive an AD diagnosis during their entire follow-up period. To ensure reliable classification, control patients were required to have a minimum follow-up duration that matched or exceeded the prediction window being analyzed (*i.e.*, at least one year of follow-up for the one-year prediction model, two years for the two-year model, and so on). This criterion ensures adequate observation time to confirm non-conversion status within each prediction timeframe.

2.2. *Observation period and prediction window*

In the final patient cohort, we designated the index date as the date of the first MCI diagnosis for both cases and controls. Our analysis encompassed all accessible patient information within the five-year observation window preceding the first MCI diagnosis. Although not all patients had a full five years of observation data prior to their MCI diagnosis, we utilized all available data within this window, acknowledging variations in observation periods among patients. This observation window constituted the dataset for our AD risk prediction task. For cases, we pinpointed the date of the first AD diagnosis after the index date as the event date, as illustrated in the timeline in Fig. 2.

The predictive modeling of MCI-to-AD progression was assessed across four distinct prediction windows, spanning one, two, three, and five years, all commencing the day following the index date. For each prediction window analysis, we included only those control patients whose follow-up duration met or exceeded the respective prediction window length. This methodology ensured consistency in our control group definition across different prediction timeframe, while also ensuring sufficient observation time to validate non-conversion status.

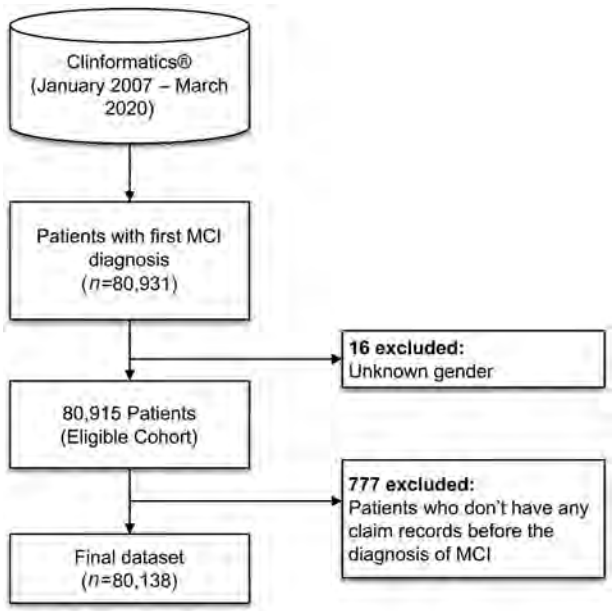

Figure 1 Study cohort construction.

2.3. *Claims data preprocessing and code mapping*

The utilization of large-scale patient claims data holds significant potential for the prediction of MCI-to-AD progression. These datasets contain comprehensive patient information, including vital variables such as demographics, diagnoses, treatments, and procedures. However, medical claim records pose challenges in model development and prediction accuracy due to their large number of variables, data sparsity, and irregular nature.

In the data-driven approach followed in this study, we preprocessed the patient data before employing it as input features in our prediction models. Demographic data, such as age, was categorized into age groups. We included all diagnostic ICD-9 and ICD-10 codes present in the patient records and removed entries containing invalid ICD diagnosis codes, which included incorrectly formatted codes, typographical errors, and codes unrecognized within the ICD system. Additionally, we converted incorrectly labeled ICD-10 diagnosis codes back to their original ICD-9 labels, and *vice versa*. To mitigate the sparsity of the features, we consolidated equivalent ICD-9 and ICD-10 diagnosis codes and mapped them to a reduced set of "phecodes" using the Phenome-wide association studies (PheWAS) method, which analyzes many phenotypes and groups similar diagnoses under unique codes[26,27]. This method reduced the total number of diagnosis codes from 46,434 to 1854.

A similar approach was applied to procedure codes, encompassing their various types: ICD-9-CM, ICD-10-PCS, Healthcare Common Procedure Coding System (HCPCS) Level I [also known as Current Procedural Terminology (CPT) codes and maintained by the American Medical Association (AMA)], and HCPCS Level II procedure codes [maintained by the Centers for Medicare & Medicaid Services (CMS)]. We first removed patient records with invalid procedure codes and corrected mislabeled procedure codes. Then, employing the Agency for Healthcare Research and Quality (AHRQ)'s Clinical Classification Software (CCS), we mapped procedure codes into their equivalent single-level procedure categories. Using this method, we reduced the total number of procedure codes from 11,067 to 236. For medications, identified in claim records using National Drug Codes (NDC), we mapped all codes to their corresponding generic medication names, resulting in 2916 unique identifiers.

2.4. *Deep learning model*

In this study, we employed the Bidirectional Gated Recurrent Unit (GRU) model[28] as our deep learning architecture. A Bidirectional GRU, or BiGRU, is a sequence-processing model consisting of two GRUs. The first GRU processes the input sequence in one direction, while the second processes it in the opposite direction. The outputs from both units are combined and used for prediction after processing the entire input sequence. BiGRUs are particularly effective for extracting meaningful representations from sequential data during training because they consider temporal dependencies by processing information bidirectionally, incorporating both preceding and succeeding elements of the input sequence.

We aggregated medical concepts (diagnoses, medications, and procedures) for all patients in 15-day intervals throughout the five-year observation window preceding the MCI index date. This aggregated information served as the input sequence for our

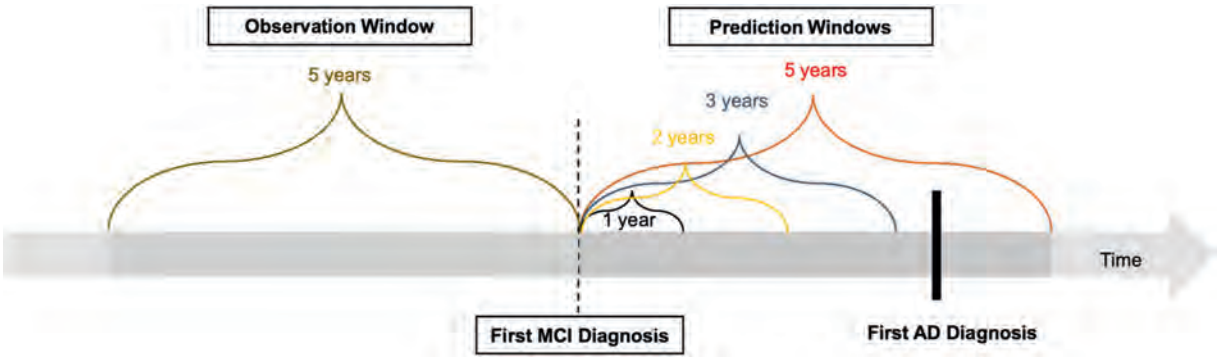

Figure 2 Patient timeline for the prediction task.

proposed deep learning model. Consequently, at each aggregation point, we transformed each patient record into a feature vector containing the aggregated medical concepts, along with demographic details, such as gender and age group.

As depicted in Fig. 3, we implemented this BiGRU architecture through multiple neural network layers. The model processes these 15-day interval inputs (shown as X_{t1} to X_{tn} at the bottom of the figure) through an embedding layer, transforming them into dense vector representations (E_{t1} to E_{tn}) that capture relationships between different medical concepts.

The core of the architecture consists of two stacked BiGRU layers, marked as "1st BiGRU Layer" and "2nd BiGRU Layer" in the figure. Each BiGRU layer contains paired GRU units—forward GRUs (shown in blue) processing the sequence from left to right, and backward GRUs (shown in green) processing from right to left. At each timestep, the outputs from forward and backward GRUs are merged via concatenation operations (depicted by $\oplus$ nodes), enabling the model to capture bidirectional temporal patterns in the patient's medical history.

Unlike traditional machine learning approaches that compress temporal data into a single multi-hot encoded vector aggregated from the five-year observation window, our architecture preserves the temporal sequence of medical events. The model culminates in a final concatenation operation (shown by the topmost $\oplus$ node) that combines the terminal hidden states of the second BiGRU layer's forward and backward paths. This concatenated representation then passes through a sigmoid function to output the probability of MCI-to-AD progression prediction.

For our prediction task, we evaluated the performance of the BiGRU against several established baseline models: Light Gradient Boosting Machine (LGBM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and traditional Logistic Regression (LR). To ensure a fair comparison, we applied consistent data preprocessing across all baseline models. For these models, we aggregated all medical concepts throughout the five-year observation window prior to the MCI index date, counting each unique medical concept only once, regardless of its frequency in the claim records. Subsequently, we constructed each patient's input feature vector by converting these aggregated medical concepts into a multi-hot encoding representation. This method transforms the temporal sequence of medical events into a fixed-length feature vector, where each dimension represents the presence (1) or absence (0) of a specific medical concept during the observation period. While this representation simplifies the complex temporal patterns in the data, it provides a standardized input format for our baseline models. This standardization allows us to effectively evaluate the advantages of our proposed BiGRU's sequential learning approach against these conventional machine learning techniques.

2.5. *Data selection and training*

In each prediction scenario, the patient cohort was randomly divided into two subsets: 80% formed the training set, and 20% constituted the test set. Within each subset, we ensured a balanced distribution of positive cases (MCI-to-AD converters) and negative controls (non-converters) to mitigate bias in model training and evaluation. The test set was reserved exclusively for the final performance assessment of all models.

To optimize model performance, we implemented a robust cross-validation strategy on the training set. The training process involved two main phases. Initially, we utilized a stratified five-fold cross-validation method for hyperparameter tuning and internal validation. As depicted in Fig. 4, this phase included five iterations, during which the training set was repeatedly split into training folds (80%) and a validation fold (20%). Each iteration featured, the rotation of the validation fold through different segments of the data. The hyperparameters were optimized using the Optuna automatic optimization software[29] (see Supporting Information Table S1 for detailed hyperparameter settings across all prediction scenarios).

Upon determining the most effective model configuration through cross-validation and hyperparameter optimization, we advanced to the second phase. The optimal model was then retrained on the entirely of the training set, representing 80% of the original dataset.

This comprehensive training and evaluation framework, as illustrated in Fig. 4, ensured robust model development. It effectively prevented data leakage and provided unbiased performance estimates, thereby enhancing the reliability of the results.

2.6. *Model evaluation*

We conducted a comprehensive evaluation by comparing the performance of our BiGRU model against several established baseline machine learning models: Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), and Random Forest (RF). The models were assessed using three key performance metrics: the area under the receiver operating curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the F1 score.

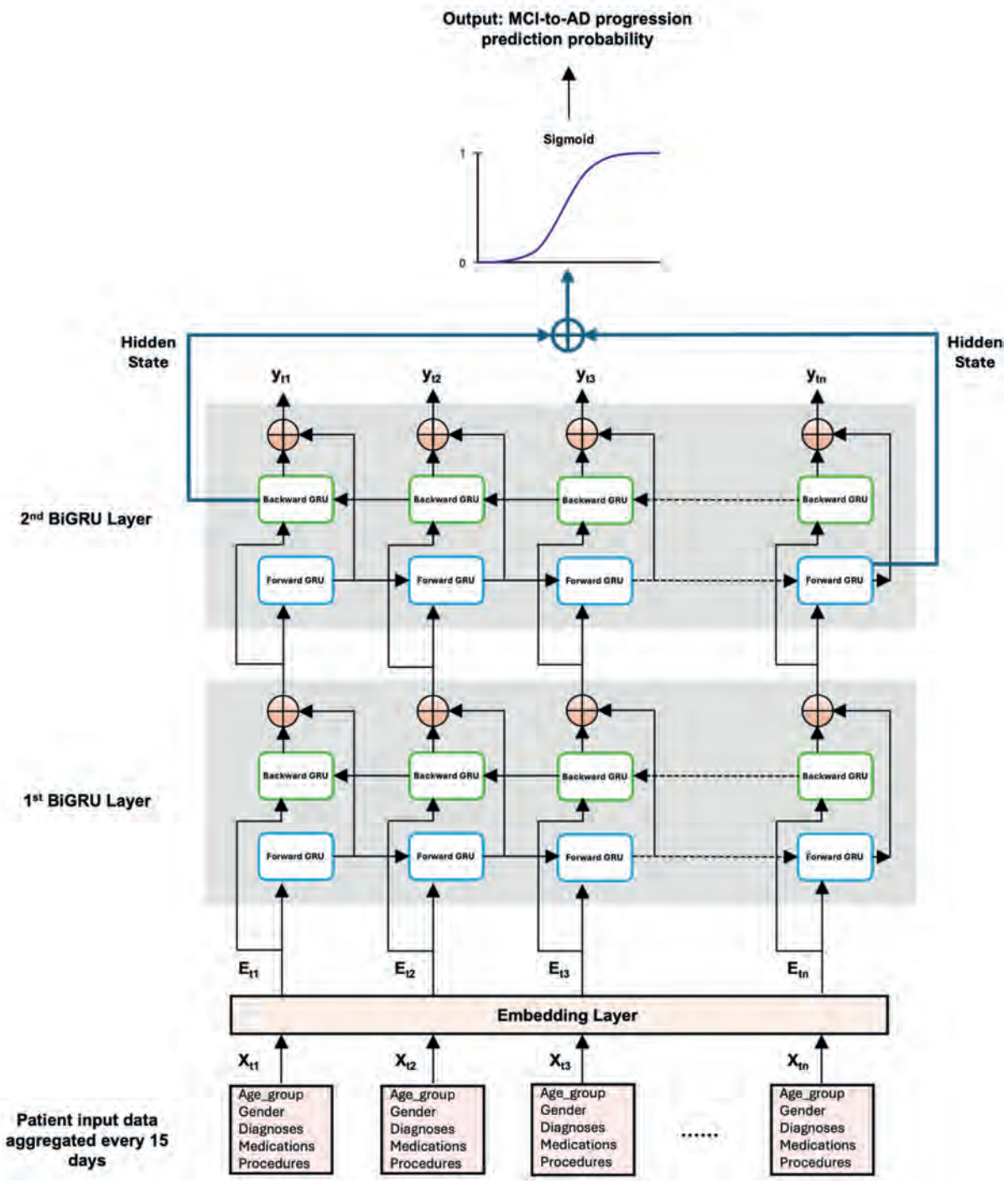

Figure 3 Deep learning model architecture.

The AUC-ROC statistic assesses a model's classification performance by depicting the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various threshold settings. A higher AUC-ROC score denotes enhanced discrimination capability between cases and controls. The AUC-PR serves as a critical performance indicator, particularly valuable for assessing classification effectiveness in datasets with class imbalances. It represents the trade-off between precision (positive predictive value) and recall (sensitivity) across different classification thresholds. A higher AUC-PR score reflects better model performance, with a score of 1 signifying an ideal precision-recall trade-off. The F1 score provides a single value performance metric by calculating the harmonic mean of precision and recall, offering a balanced assessment of the model's precision-recall trade-off, and is defined as Eq. (1):

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

For a consistent comparison across all models, including BiGRU and the baseline machine learning models, we calculated

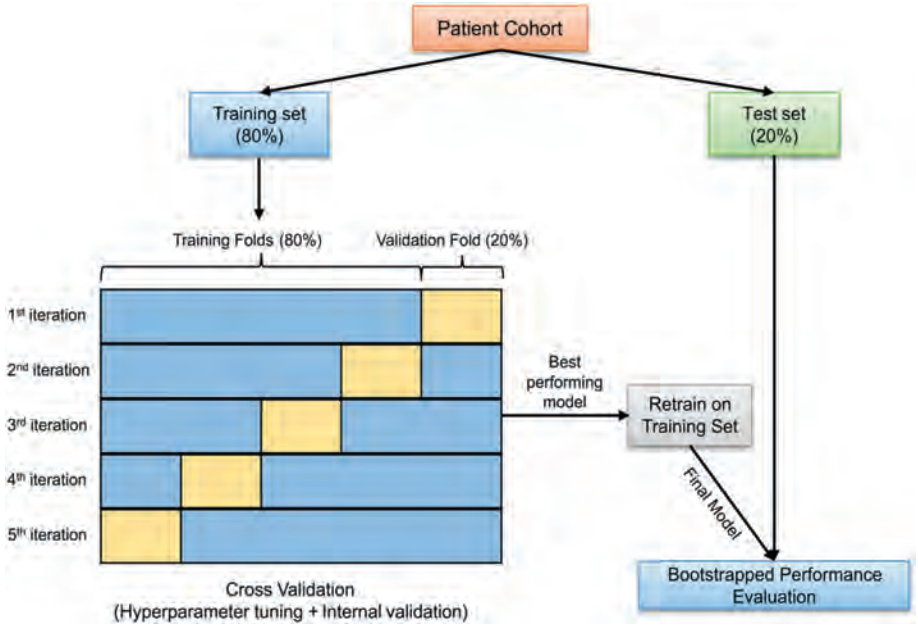

Figure 4 Data selection for the training and testing in all prediction scenarios.

these metrics using a classification threshold of 0.5. To ensure robust evaluation, we employed bootstrapped performance assessment on the held-out test set to generate confidence intervals for our performance metrics. This comprehensive evaluation framework allowed us to rigorously compare the predictive capabilities of our proposed BiGRU architecture against established machine learning approaches.

3. Results

3.1. Patient characteristics

The baseline characteristics of the total study cohort are shown in Table 1. Among patients diagnosed with MCI who progressed to AD (case-patients), females comprised 62.58% ($n = 7811$) and males 37.42% ($n = 4670$). In the control group (patients diagnosed with MCI who did not progress to AD), females represented 56.24% ($n = 38{,}052$) and males 43.76% ($n = 29{,}605$).

The mean age of case patients at MCI diagnosis was 78.49 years (SD: 6.60), and their average age at the AD diagnosis was 80.61 years (SD: 6:65). For control patients, the mean age at the MCI diagnosis was 74.48 years (SD: 10.33). The average time from initial MCI diagnosis to AD conversion was 776.8 days (median: 586 days).

Regarding comorbidities, obesity was present in 21.53% ($n = 2687$) of case patients compared to 31.50% ($n = 21{,}313$) of control patients. Case patients showed different comorbidity patterns compared to controls: hypertension was notably high in both groups but more prevalent in case patients (82.62% *vs* 77.78%), depression (39.99% *vs* 32.42%), and anxiety (28.24% *vs* 24.93%) were more common in case patients. Type 2 diabetes mellitus was observed in 32.67% of case patients compared to 35.04% of controls, while smoking rates were lower in case patients at 39.12% ($n = 4883$) compared to 41.73% ($n = 28{,}232$) in controls. All these differences were statistically significant ($P < 0.0001$).

The number of patients belonging to both the case and control groups varied across the four prediction intervals of our testing

Table 1 Cohort demographic and clinical characteristics of patients.

Characteristic	Case ($n = 12{,}481$)	Control ($n = 67{,}657$)	Overall ($n = 80{,}138$)	*P* value
Age (mean)	MCI mean (SD): 78.49 (6.60) AD mean (SD): 80.61 (6.65)	MCI mean (SD): 74.48 (10.33)	MCI mean: 75.11	<0.0001[a]
Female	7811 (62.58%)	38,052 (56.24%)	45,863 (57.2%)	<0.0001[b]
Obesity	2687 (21.53%)	21,313 (31.50%)	24,000 (29.95%)	<0.0001
Smoking	4883 (39.12%)	28,232 (41.73%)	33,115 (41.32%)	<0.0001
Type 2 diabetes mellitus	4078 (32.67%)	23,711 (35.04%)	27,789 (34.68%)	<0.0001
Hypertension	10,312 (82.62%)	27,118 (77.78%)	52,629 (78.54%)	<0.0001
Anxiety	3525 (28.24%)	16,869 (24.93%)	20,394 (25.45%)	<0.0001
Depression	4991 (39.99%)	21,931 (32.42%)	26,922 (33.59%)	<0.0001
Conversion time (MCI to AD)	Mean: 776.8 days Median: 586 days			

[a]Age *P*-value is calculated with the Kruskal test.
[b]Categorical parameters' *P*-value is calculated with the Chi−Square test.

scenarios. The patient counts for each testing scenario are shown in Fig. 5.

We adopted a methodology analogous to that described in[20] for selecting control patients based on their last claim record. As illustrated in Fig. 6, we established specific criteria for including control patients based on the timeline of their claim records. Panel (a) displays a valid control patient, where the last claim record extends beyond the prediction window (X years) following the MCI diagnosis. In contrast, panel (b) demonstrates scenario where a patient is excluded from the control group because their last claim record falls within the prediction window. This exclusion criterion is essential as it ensures that the follow-up data is complete; without it, we cannot definitively determine whether these patients progressed to AD within the prediction window. This methodology guarantees that all included control patients have adequate follow-up time to confirm their non-conversion status for at least X years following their initial MCI diagnosis.

3.2. Prediction performance comparison across models

Table 2 presents the prediction performance metrics of our BiGRU deep learning model alongside four baseline machine learning models: LGBM, XGBoost, LR, and RF. Performance was assessed across four prediction scenarios using the testing set.

A comparative analysis of the AUC-ROC, AUC-PR, and the F1-Score values reveals distinct performance patterns across different prediction windows. Notably, the BiGRU model exhibited higher performance for longer prediction intervals, particularly in the 3- and 5-year scenarios. For the 5-year prediction window, BiGRU achieved an AUC-ROC of 0.833 (95% CI: 0.822−0.843), surpassing LGBM at 0.818 (95% CI: 0.806−0.830), XGBoost at 0.818 (95% CI: 0.807−0.823), LR at 0.816 (95% CI: 0.804−0.828), and RF at 0.801 (95% CI: 0.790−0.813). In this scenario, the BiGRU model also attained an AUC-PR of 0.856 (95% CI: 0.845−0.867) and an F1-Score of 0.71 (95% CI: 0.694−0.724), marking the best overall performance.

For the 2-year prediction window, all models showed comparable performance levels. However, in the shorter 1-year interval, the baseline models, particularly LGBM with an AUC-ROC of 0.694 (95% CI: 0.678−0.709), exceeded the performance of the BiGRU model, which posted an AUC-ROC of 0.679 (95% CI: 0.662−0.695). This pattern suggests that while traditional machine learning approaches may suffice for short-term predictions, the BiGRU's capacity to integrate temporal patterns proves more advantageous for predicting for longer-term progression prediction.

4. Discussion

This research systematically investigated the application of sequence-based deep learning in predicting the progression from MCI to AD, focusing on the Bidirectional Gated Recurrent Unit (BiGRU) architecture while comparing its performance against multiple baseline machine learning models. The study leveraged patient data gathered before the initial MCI diagnosis and involved the analysis of over 80,000 patients diagnosed with MCI. The effectiveness of the models in forecasting both short-term and long-term progression from MCI to AD was evaluated across different prediction windows of one, two, three, and five years. This predictive capability was enhanced by utilizing extensive patient claims data, which provided comprehensive information including demographics, diagnoses, medications, and procedures.

The clinical significance of this research is underscored by current progression patterns and treatment options. Notably, the conversion rate from MCI to AD over five years varies from 30% to 50%, depending on the studied cohort[30], indicating that a significant portion of MCI patients do not progress to AD during this period. With the advent of disease-modifying therapies (DMTs), which are most effective in the earliest stages of MCI and AD,

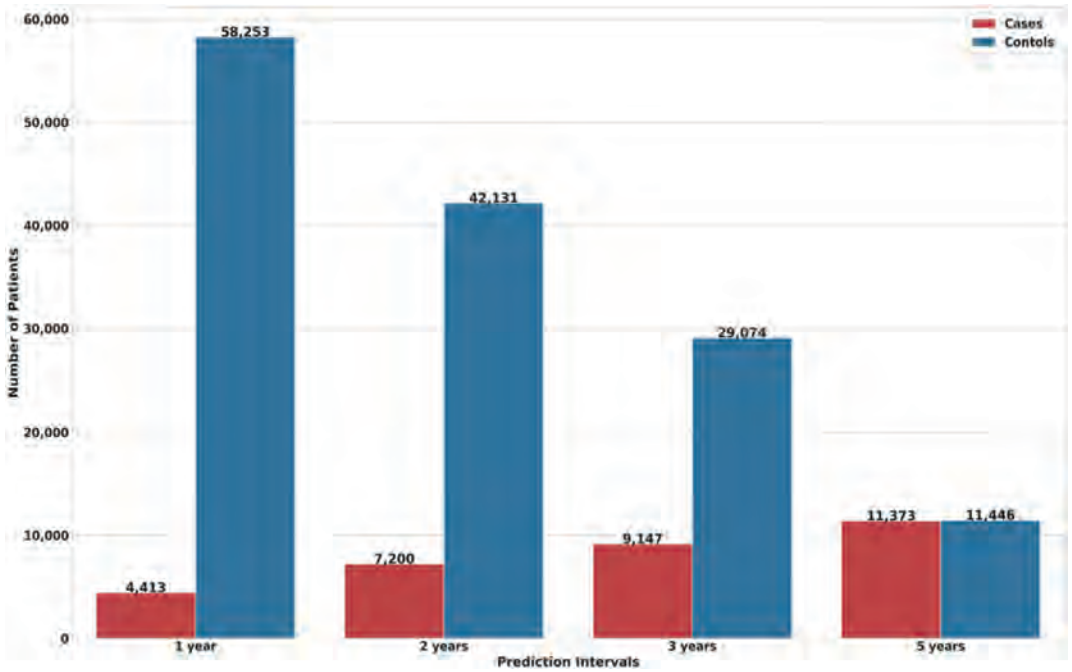

Figure 5 Patient count for each prediction interval.

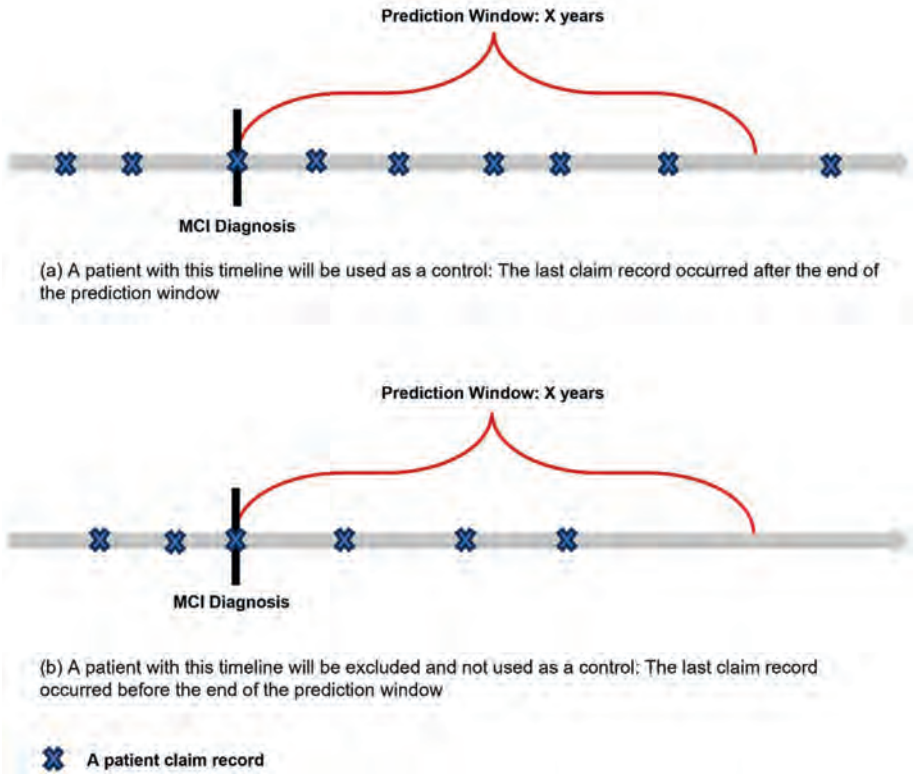

Figure 6 Control patient inclusion and exclusion criteria.

Table 2 Comparison of model performance between the bidirectional gated recurrent units (BiGRU) deep learning model and baseline machine learning models (LGBM, XGBoost, LR, RF) using the area under the receiver operating characteristic curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the F1-Score.

Performance measure	Prediction scenarios			
	1−365 days (1-year)	1−730 days (2-years)	1−1095 days (3-years)	1−1825 days (5-years)
BiGRU				
AUC-ROC	0.679 (0.662,0.695)	0.690 (0.674, 0.702)	0.751 (0.741,0.763)	0.833 (0.822,0.843)
AUC-PR	0.122 (0.112,0.136)	0.274 (0.251,0.295)	0.554 (0.533,0.575)	0.856 (0.845,0.867)
F1 (Binary)	0.182 (0.169,0.191)	0.326 (0.309,0.341)	0.492 (0.473,0.510)	0.71 (0.694,0.724)
LGBM				
AUC-ROC	0.694 (0.678,0.709)	0.692 (0.679,0.707)	0.736 (0.721,0.749)	0.818 (0.806,0.830)
AUC-PR	0.131 (0.119,0.145)	0.272 (0.252,0.295)	0.495 (0.473,0.521)	0.838 (0.824,0.851)
F1 (Binary)	0.195 (0.186,0.215)	0.327 (0.316,0.349)	0.481 (0.465,0.499)	0.716 (0.701,0.731)
XGBoost				
AUC-ROC	0.676 (0.659,0.693)	0.698 (0.684,0.711)	0.731 (0.718, 0.745)	0.818 (0.807,0.823)
AUC-PR	0.127 (0.115,0.143)	0.295 (0.276,0.311)	0.493 (0.469,0.517)	0.838 (0.825,0.849)
F1 (Binary)	0.188 (0.176,0.201)	0.334 (0.320,0.349)	0.484 (0.468,0.501)	0.718 (0.704,0.732)
LR				
AUC-ROC	0.676 (0.659, 0.692)	0.695 (0.682,0.709)	0.729 (0.717,0.741)	0.816 (0.804,0.828)
AUC-PR	0.123 (0.111,0.138)	0.268 (0.251,0.288)	0.486 (0.463,0.509)	0.836 (0.823,0.849)
F1 (Binary)	0.191 (0.178,0.205)	0.340 (0.325,0.356)	0.479 (0.461, 0.494)	0.711 (0.695, 0.727)
RF				
AUC-ROC	0.662 (0.645,0.678)	0.681 (0.667,0.695)	0.714 (0.701,0.727)	0.801 (0.790,0.813)
AUC-PR	0.117 (0.107,0.130)	0.271 (0.252,0.292)	0.463 (0.435,0.485)	0.821 (0.806,0.834)
F1 (Binary)	0.186 (0.173,0.199)	0.328 (0.311,0.343)	0.464 (0.447,0.482)	0.693 (0.678,0.708)

early prediction of progression risk has become crucial. However, given that DMTs can entail significant side effects, clinicians may prefer to reserve these treatments for patients most likely to progress. This research addresses this clinical need by developing a deep learning approach that can predict MCI-to-AD progression across different time windows, thus aiding clinicians in more informed discussions with their patients regarding individual risk-benefit ratios when considering DMT initiation.

It is important to highlight that predicting MCI-to-AD progression presents notably challenging endeavors compared to

identifying initial Alzheimer's disease and associated dementias. MCI patients often share similar diagnoses and medications, complicating the distinction between those who will progress to AD and those who will not. In contrast, distinguishing patients who have never experienced any stage of AD from those who have is generally more straightforward, as the clinical differences are well-defined.

Our study sourced its data from Clinformatics®, ensuring a comprehensive representation that extends beyond the confines of a single hospital or member hospitals. Despite this broad scope, accurately forecasting the progression from MCI to AD remains challenging due to the multitude of variables within the claims data. Furthermore, the sparsity, irregularity, and temporality of the data impede the seamless creation of models, adding complexity to the prediction process.

To address these complexities, advanced analytics techniques, including machine learning (ML) algorithms, prove invaluable in analyzing these claims data. These algorithms can handle nonlinear data and unveil hidden correlations between clinical variables and desired outcomes. Deep learning algorithms, particularly effective with large datasets, eliminate the need for extensive feature selection expertise. The temporal nature of medical claim data encodes valuable information that can be particularly well-captured through deep learning sequence modeling techniques. Our findings showed that sequence-based models' ability to process sequential information bidirectionally made them suitable for capturing temporal patterns in claims data. While other architectures like vanilla RNNs or different LSTM variants could potentially offer additional insights, our focus was on establishing the effectiveness of sequence-based deep learning approaches versus traditional ML methods that rely on engineered features. We chose BiGRU over LSTM as it offers simpler and computationally less intensive architecture with fewer parameters.

Our model contributes to the field by addressing both short-term and long-term progression from MCI to AD. Using longitudinal patient claim records before MCI diagnosis, which include demographic and clinical information, our study demonstrates that deep learning models can effectively predict MCI-to-AD progression with promising accuracy. Through the analysis of our patient data, we confirmed previously established high-risk factors for AD, including age, gender, hypertension, anxiety, and depression. Notably, one study[31] suggests that mood disorders not only serve as potential risk factors for AD but may also manifest as presenting symptoms of the disease, a correlation that aligns seamlessly with our research outcomes.

The robustness of this study is emphasized by its reliance on a sizable and diverse patient cohort, which ensures a comprehensive representation of individuals transitioning from MCI to AD. The BiGRU model employed leverages all accessible patient data as predictors for the progression from MCI to AD. This approach obviates the need for expert-driven feature selection and thereby streamlining the analytical process. Additionally, the model's ability to handle temporal sequences of varying lengths is particularly valuable in healthcare settings where patient visit patterns can be irregular. Noteworthy is the model's capacity to forecast the short-term progression from MCI to AD and its effectiveness over an extended period, providing valuable insights into the long-term trajectory of the disease, with performance notably improving in longer prediction windows.

Another distinctive advantage of this study is the utilization of claims data, a universally available resource for virtually every patient. The model's versatility extends beyond the confines of a single hospital, demonstrating applicability across various healthcare institutions rather than being restricted to a specific facility or group of member hospitals. The model efficiently processes raw claims data, incorporating a diverse range of predictors including demographics, diagnoses, medications, and procedures. This approach provides a comprehensive view of patient health status without the need for extensive data preprocessing.

In contrast, with the limited availability of recorded images and the potential challenges associated with obtaining genetic information, which may not be as widespread unless necessary for specific medical requirements. The study's reliance on this widely accessible claims data enhances its practicality and generalizability in real-world healthcare settings.

It is important, however, to recognize certain limitations to this analysis. The accuracy of diagnoses, medications, or procedural codes may be influenced by the complexities of billing practices and clinic workflows, introducing a potential source of imprecision. Misclassification of these codes could introduce bias into our results. Additionally, claims data lack information about important factors such as lifestyle, socioeconomic status, and other social determinants of health that might influence disease progression. The potential impact of missing data or incomplete follow-up in claims databases presents another limitation. Moreover, since our predictive model is based on historical data, future alterations in clinical practices and treatments could affect disease progression patterns.

Another limitation stems from inherent class imbalance in our dataset, where the number of patients who progress from MCI to AD is considerably smaller than those who do not progress. To address this issue, we implemented a class weighting strategy during model training to mitigate its impact. Although this adjustment has improved model performance, the underlying issue of data imbalance remains and warrants further investigation.

Despite recognizing the importance of evaluating our model's generalizability across diverse datasets, external validation was not conducted due to limited accessibility to such data. To mitigate this constraint, our study design incorporated two tiers of internal validation, to provide a robust assessment. Because the model relies on standard coded variables, it can be readily adapted to comparable datasets, potentially validated without retraining. This adaptability highlights a practical avenue for future research and validation efforts, ensuring that similar coding structures can facilitate the assessment and enhancement of the model's predictive reliability.

Additionally, while our study utilized a proprietary claims dataset which may affect result reproducibility, we have made our complete implementation publicly available through a GitHub repository, https://github.com/Tao-AI-group/MCI-to-AD-Progression-Model, with comprehensive documentation. This ensures that researchers can apply our methodology to any EHR or claims dataset, fostering reproducibility and wider applicability of our approach.

5. Conclusions

This study explores the potential of deep learning models in predicting the progression from MCI to AD using longitudinal claims data. Our results demonstrate that the BiGRU model can effectively predict progression risk, particularly over longer intervals following initial MCI diagnosis. By leveraging comprehensive patient data without the necessity for expert-driven feature

selection, the model provides a pragmatic approach to risk assessment. This could assist clinicians in identifying patients at elevated risk for AD progression, potentially facilitating more informed decisions about monitoring and intervention strategies. The findings underscore that machine learning approaches, when applied to readily available healthcare data, can contribute to early detection strategies in neurocognitive disorders. While further validation is needed, this research provides a foundation for developing automated tools that can augment existing clinical assessment methods. Future research could explore the integration of such predictive models into clinical workflows, potentially enhancing the timeliness and efficiency of AD risk assessment in clinical practice. This integration promises to advance clinical practice by enabling more proactive and personalized patient care.

Acknowledgments

The study was partially funded by the National Institute on Aging (NIH/NIA) [grant numbers R01AG084236, R01AG083039, RF1AG072799 and R56AG074604, USA].

Author contributions

Cui Tao and Ahmed Abdelhameed designed the study; Ahmed Abdelhameed prepared and analyzed the data with the help of Jingna Feng, Xinyue Hu, and Sori Lundin. Ahmed Abdelhameed designed the models and generated the results and the figures. Fang Li helped with the design of the machine learning model; Paul Schulz reviewed and evaluated the clinical findings of the study. Cui Tao provided institutional support and guided the data analysis; All authors discussed the results and read and approved the final manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2025.01.027.

References

1. Davis M, O'Connell T, Johnson S, Cline S, Merikle E, Martenyi F, et al. Estimating Alzheimer's disease progression rates from normal cognition through mild cognitive impairment and stages of dementia. *Curr Alzheimer Res* 2018;**15**:777−88.
2. Alzheimer's disease facts and figures. *Alzheimers Dement* 2023;**19**: 1598−695.
3. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;**7**:280−92.
4. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;**7**:270−9.
5. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;**7**:263−9.
6. Monfared AAT, Byrnes MJ, White LA, Zhang Q. Alzheimer's disease: epidemiology and clinical progression. *Neurol Ther* 2022;**11**:553−69.
7. Abrol A, Bhattarai M, Fedorov A, Du Y, Plis S, Calhoun V. Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease. *J Neurosci Methods* 2020;**339**:108701.
8. Lin Y, Huang K, Xu H, Qiao Z, Cai S, Wang Y, et al. Predicting the progression of mild cognitive impairment to Alzheimer's disease by longitudinal magnetic resonance imaging-based dictionary learning. *Clin Neurophysiol* 2020;**131**:2429−39.
9. Zhang T, Liao Q, Zhang D, Zhang C, Yan J, Ngetich R, et al. Predicting MCI to AD conversation using integrated SMRI and RS-FMRI: machine learning and graph theory approach. *Front Aging Neurosci* 2021;**13**:688926.
10. Cabral C, Morgado PM, Costa DC, Silveira M. Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Comput Biol Med* 2015;**58**:101−9.
11. Fang Y, Du N, Xing L, Duo Y, Zheng L. Evaluation of hippocampal volume and serum brain-derived neurotrophic factor as potential diagnostic markers of conversion from amnestic mild cognitive impairment to Alzheimer disease. *Medicine* 2019;**98**:e16604.
12. Pusil S, Dimitriadis SI, López ME, Pereda E, Maestú F. Aberrant MEG multi-frequency phase temporal synchronization predicts conversion from mild cognitive impairment-to-Alzheimer's disease. *Neuroimage Clin* 2019;**24**:101972.
13. Xu M, Sanz DL, Garces P, Maestu F, Li Q, Pantazis D. A Graph Gaussian embedding method for predicting Alzheimer's disease progression with MEG brain networks. *IEEE Trans Biomed Eng* 2021;**68**: 1579−88.
14. Grueso S, Viejo-Sobera R. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. *Alzheimers Res Ther* 2021;**13**:162.
15. Burkett BJ, Johnson DR, Lowe VJ. Evaluation of neurodegenerative disorders with Amyloid-β, Tau, and dopaminergic PET imaging: interpretation pitfalls. *J Nucl Med* 2024;**65**:829−37.
16. Bouwman FH, Frisoni GB, Johnson SC, Chen X, Engelborghs S, Ikeuchi T, et al. Clinical application of CSF biomarkers for Alzheimer's disease: from rationale to ratios. *Alzheimers Dement (Amst)* 2022;**14**:e12314.
17. Palmqvist S, Tideman P, Mattsson-Carlgren N, Schindler SE, Smith R, Ossenkoppele R, et al. Blood biomarkers to detect Alzheimer disease in primary care and secondary care. *JAMA* 2024;**332**:1245.
18. Spasov S, Passamonti L, Duggento A, Liò P, Toschi N. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage* 2019;**189**: 276−87.
19. Fouladvand S, Mielke MM, Vassilaki M, Sauver JS, Petersen RC, Sohn S. Deep learning prediction of mild cognitive impairment using electronic health records. *Proc (IEEE Int Conf Bioinforma Biomed* 2019;**2019**:799−806.
20. Gao XR, Chiariglione MP, Nuytemans K, Li Y, Martin ER. Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction. *Alzheimers Dement* 2023;**19**:e063653.
21. Mao C, Xu J, Rasmussen L, Li Y, Adekkanattu P, Pacheco J, et al. AD-BERT: using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *J Biomed Inform* 2023;**144**:104442.
22. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *NPJ Digit Med* 2022;**5**:194.
23. Yu H, Fan L, Li L, Zhou J, Ma Z, Xian L, et al. Large language models in biomedical and health informatics: a review with bibliometric analysis. *arXivOrg* 2024. Available from: https://arxiv.org/abs/2403.16303.
24. Du X, Novoa-Laurentiev J, Plasaek JM, Chuang YW, Wang L, Marshall G, et al. Enhancing early detection of cognitive decline in the

elderly: a comparative study utilizing large language models in clinical notes. *medRxiv* 2024. Available from: https://doi.org/10.1101/2024.04.03.24305298.
25. Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. *J Biomed Inform* 2021;**126**:103980.
26. Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017;**12**:e0175508.
27. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to Phecodes: workflow development and initial evaluation. *JMIR Med Inform* 2019;**7**:e14325.
28. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXivOrg* 2014. Available from: https://arxiv.org/abs/1409.0473.
29. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna. In: *Proceedings of the 25th ACM SIGKDD international Conference on knowledge discovery; data mining*; 2019. Available from: https://doi.org/10.1145/3292500.3330701.
30. Liss JL, Assunção SS, Cummings J, Atri A, Geldmacher DS, Candela SF, et al. Practical recommendations for timely, accurate diagnosis of symptomatic Alzheimer's disease (MCI and dementia) in primary care: a review and synthesis. *J Intern Med* 2021;**290**:310−34.
31. Mendez MF. Degenerative dementias: alterations of emotions and mood disorders. *Handb Clin Neurol* 2021;**183**:261−81.