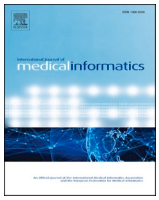
International Journal of Medical Informatics 200 (2025) 105917

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Automated phenotyping of mild cognitive impairment and Alzheimer's disease and related dementias using electronic health records

Ruoqi Wei [a,b,c,d], Stephanie S Buss [a,c], Rebecca Milde [a,c], Marta Fernandes [a,b,c], Daniel Sumsion [a,k], Elijah Davis [a,l], Wan-Yee Kong [a,c], Yiwen Xiong [a,c], Jet Veltink [a,c], Samvrit Rao [a,c], Tara M. Westover [b,c], Lydia Petersen [a,k], Niels Turley [a], Arjun Singh [b,c], Sudeshna Das [a,c], Valdery Moura Junior [a,b,c], Manohar Ghanta [a,b,c], Aditya Gupta [a,b,c], Jennifer Kim [e], Alice D. Lam [b,c], Katie L. Stone [f,m], Emmanuel Mignot [g], Dennis Hwang [h], Lynn Marie Trotti [i], Gari D. Clifford [i,j], Umakanth Katwa [k,c], Robert J Thomas [a,c,1], Shibani Mukerji [a,c,1], Sahar F. Zafar [b,c,1], M. Brandon Westover [a,b,c,*], Haoqi Sun [a,b,c,1]

[a] Beth Israel Deaconess Medical Center, Boston, MA, USA
[b] Massachusetts General Hospital, Boston, MA, USA
[c] Harvard Medical School, Boston, MA, USA
[d] Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, FL, USA
[e] Yale School of Medicine, New Haven, CT, USA
[f] California Pacific Medical Center Research Institute, San Francisco, CA, USA
[g] Stanford University, Palo Alto, CA, USA
[h] Kaiser Permanente, San Bernardino County Sleep Disorders Center. San Bernardino, CA, USA
[i] Emory University School of Medicine, Atlanta, GA, USA
[j] Georgia Institute of Technology, Atlanta, GA, USA
[k] Boston Children's Hospital, Boston, MA, USA
[l] Brigham Young University, Provo, UT 84602, USA
[m] Department of Epidemiology and Biostatistics, University of California, San Francisco, USA

ARTICLE INFO

Keywords:
Electronic health records
Alzheimer disease
Dementia
Mild cognitive impairment
Machine learning

ABSTRACT

Objectives: Unstructured and structured data in electronic health records (EHR) are a rich source of information for research and quality improvement studies. However, extracting accurate information from EHR is labor-intensive. Timely and accurate identification of patients with Alzheimer's Disease, related dementias (ADRD), or mild cognitive impairment (MCI) is critical for improving patient outcomes through early intervention, optimizing care plans, and reducing healthcare system burdens. Here we introduce an automated EHR phenotyping model to streamline this process and enable efficient identification of these conditions.

Methods: We analyzed data from 3,626 outpatients seen at two hospitals between February 2015 and June 2022. Through manual chart review, we established ground truth labels for the presence or absence of MCI/ADRD diagnoses. Our model combined three types of data: (1) unstructured clinical notes, from which we extracted single words, two-word phrases (bigrams), and three-word phrases (trigrams) as features, weighted using Term Frequency-Inverse Document Frequency (TF-IDF) to capture their relative importance, (2) International Classification of Diseases (ICD) codes, and (3) medication prescriptions related to MCI/ADRD. We trained a regularized logistic regression model to predict MCI/ADRD diagnoses and evaluated its performance using standard metrics including area under the receiver operating curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, specificity, precision, recall, and F1 score.

Results: Thirty percent of patients in the cohort carried diagnoses of MCI/ADRD based on manual review. When evaluated on a held-out test set, the best model using clinical notes, ICDs, and medications, achieved an AUROC of 0.98, an AUPRC of 0.98, an accuracy of 0.93, a sensitivity (recall) of 0.91, a specificity of 0.96, a precision of

* Corresponding authors at: Emily Fisher Landau Professor of Neurology, Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, 02215, USA.
E-mail addresses: bwestove@bidmc.harvard.edu (M.B. Westover), hsun3@bidmc.harvard.edu (H. Sun).
[1] Co-senior authors.

https://doi.org/10.1016/j.ijmedinf.2025.105917
Received 7 August 2024; Received in revised form 31 January 2025; Accepted 7 April 2025
Available online 11 April 2025

0.96, and an F1 score of 0.93 The estimated overall accuracy for patients randomly selected from EHRs was 99.88%.
Conclusion: Automated EHR phenotyping accurately identifies patients with MCI/ADRD based on clinical notes, ICD codes, and medication records. This approach holds potential for large-scale MCI/ADRD research utilizing EHR databases.

1. Introduction

Globally, 12 % to 18 % of people aged 60 or older are living with mild cognitive impairment (MCI) [1], and 10 % to 15 % of individuals living with MCI develop dementia each year [2]. About one-third of people living with MCI due to Alzheimer's disease (AD) develop dementia within five years [3]. The number of people in the United States with AD dementia will increase dramatically in the next 30 years due to growth of the population over the age of 65 [4]. This rising prevalence poses a significant burden on healthcare systems and underscores the need for accurate phenotyping to guide research and clinical decision-making.

Observational data from Electronic Health Records (EHRs) are an increasingly important resource for research on risk factors and potential interventions for MCI and AD [5–8]. A key challenge in scaling up EHR-based research is the accurate phenotyping of patients with a diagnosis of MCI and ADRD. Many studies rely on billing codes (International Classification of Diseases: ICD-9, ICD-10) [9,10]; however, these are often inaccurate[11]. Accurate identification of MCI and ADRD is essential, as inaccurate phenotyping may result in misdiagnoses or inappropriate treatment decisions, negatively impacting patient outcomes and leading to inefficient allocation of healthcare resources. Manual review of clinical notes is more accurate [12–16] but labor intensive and impossible to conduct at large scale [17–19].

Automated EHR phenotyping seeks to address these challenges by automatically extracting information from clinical notes and combining this with structured information (e.g., medication prescriptions and diagnostic billing codes) to infer information of interest [20]. In addition, automated EHR phenotyping addresses the limitations of traditional rule-based computable phenotyping methods[21]. These methods rely on predefined algorithms to integrate structured data, such as ICD codes, medications, and laboratory results, providing clear interpretability and ease of implementation. However, they often struggle with data incompleteness and lack the flexibility to capture complex relationships and contextual nuances within clinical language, particularly when unstructured data, such as clinical notes, is involved. In contrast, the phenotyping model presented in this study integrates rule-based methods for structured data, NLP techniques for unstructured data, and machine learning (ML) to dynamically learn patterns and interactions[22,23]. By leveraging the strengths of these approaches, the model adapts to diverse datasets, identifies non-linear relationships, and offers improved accuracy, scalability, flexibility, and interpretability. This hybrid ML-based EHR phenotyping model is designed to automate and enhance the chart review process for MCI/ADRD, providing a robust framework for advancing clinical research and decision-making.We demonstrate that our model combining information from clinical notes, ICD codes, and medications provide accurate MCI/ADRD phenotyping and is thus suitable for large-scale EHR research.

2. Materials & methods

2.1. Study cohort

This retrospective cohort study aimed to develop and evaluate a ML-based phenotyping model to identify patients with MCI and ADRD by integrating structured data, including ICD codes and medications, with unstructured data from clinical notes. The study utilized data from Massachusetts General Hospital (MGH) and Beth Israel Deaconess Medical Center (BIDMC). Data were collected for visits occurring between January 3, 2012, and November 3, 2017, under Institutional Review Board-approved protocols with waivers of informed consent. A consort diagram is provided in Fig. 1.

Eligibility criteria included patients aged 50 years or older, reflecting the age-associated nature of MCI and ADRD. We randomly selected patients for inclusion using a stratified sampling strategy, to ensure adequate representation of patients with low, medium, and high likelihood of having an MCI/ADRD diagnosis to facilitate subsequent model development. Specifically, we created 4 groups based on the presence or absence of computable criteria (i.e. criteria that do not depend on analysis of unstructured text data): 'MED-ICD-', 'MED + ICD+', 'MED + ICD-', 'MED-ICD+', denoting groups of patients with and without ICD codes and medications (MED) associated with MCI/ADRD. Notes for patients within each of these groups were subsequently manually reviewed to determine which patients had an MCI/ADRD diagnosis (described below).

2.2. Study data

Study data included unstructured (i.e., free text) clinical notes and structured data. Structured data included International Classification of Diseases (ICD) codes [24] for MCI/ADRD, and dementia-related medications (see below). Clinical notes included office visit notes, admission notes, progress notes, discharge notes, and correspondence from all medical specialties in the MGH and BIDMC systems. Clinical notes contain a wide range of information such as chief complaint; history of present illness; physician examinations, observations, assessments, and treatment plans; active problems; and current and past medications. All patients included in the study had at least one clinical note. If a patient had any MED or ICD records, only notes recorded after the first appearance of a MED or ICD record were selected for analysis. We removed notes with fewer than 500 words as these were generally administrative notes without significant medical content.

2.3. Ground truth labels: Manual chart review

The primary outcome was a confirmed diagnosis of MCI or ADRD, determined through manual chart review. We used a web-based tool developed in house that highlights keywords within notes from a

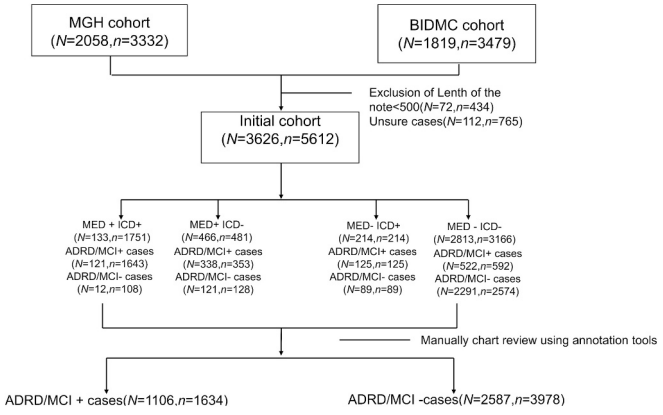


Fig. 1. CONSORT diagram. *N* is the number of unique patients. *n* is the number of visits.

provided list. A neurologist (MBW) performed manual review of all notes in each of the 'MED + ICD+', 'MED + ICD-', 'MED- ICD+', and 'MED-ICD-' groups, and assigned a final yes/no label regarding MCI/ADRD status. Cases marked as 'uncertain' were excluded.

In the 'MED + ICD+' group, patients were prescribed medications typically associated with ADRD/MCI ('MED+') and had ADRD/MCI-related ICD codes ('ICD+'). In the 'MED + ICD-' group, patients were prescribed ADRD/MCI-related medication ('MED+'), but no ADRD/MCI-related ICD codes ('ICD-') were present in their medical records. In the 'MED- ICD+' group, patients were not prescribed ADRD/MCI-related medications ('MED-'), yet their medical records contained ADRD/MCI-related ICD codes ('ICD+'). 'MED-ICD-' cases involved neither ADRD-related medication ('MED-') nor ADRD-related ICD codes ('ICD-') in the medical records.

2.4. Predictors included in the model

Predictors included structured and unstructured data. The entire method is depicted in Fig. 2. Structured data consisted of 9 predefined groups of dementia-related medications and 6 categories of ICD codes relevant to MCI and ADRD. Unstructured data comprised text features extracted from clinical notes, such as unigrams, bigrams, and trigrams, weighted using Term Frequency-Inverse Document Frequency (TF-IDF). These features served as inputs to the machine learning model, enabling the integration of diverse data sources for phenotyping.

ICD code groupings: ICD groupings and medications were defined a priori by three neurologists (MBW, SZ, SM), including: "Alzheimer's disease" – ICD-10 F00, G30.0, G30.1, G30.8, G30.9 and ICD-9 290.0, 290.2x, 290.3, 331.0; "Vascular Dementia" – ICD-10 F01.X and ICD-9 290.4X; "Lewy Body Dementia" – ICD-10 G31.83 and ICD-9 331.82; "Frontotemporal Dementia" – ICD-10 G31.0, G31.01, G31.09 and ICD-9 311.11, 331.19; "Unspecified Dementias" – ICD-10 F02.8x, F03.9x and ICD-9 294.1x, 294.2x; "Mild Cognitive Impairment" ICD-10 F06.7 and ICD-9 331.83. One day before and after the visit was considered for assignment of an ICD code to account for prior or delayed data entry.

Medications: Aricept, donepezil, Exelon, rivastigmine, memantine, Namenda, Namzaric, Razadyne, and galantamine [25].

Text features: Keywords, phrases, and word patterns were extracted from notes. We converted the text in each note to lowercase, removed stop words and special characters, and applied lemmatization. Subsequently, we extracted unique words from each note. We also extracted unique bigrams (two consecutive words) and trigrams (three consecutive words) to identify potentially discriminative features. For each note, we created a vector representation of the information within the note using TF-IDF weighting [26,27]. TF-IDF quantifies the importance of a word within a note, relative to its prevalence across a collection of notes. Specifically, we created vectors containing TF-IDF values for each of the candidate features identified above. TF-IDF features were assembled into a single overall vector which serves as input to the classification model.

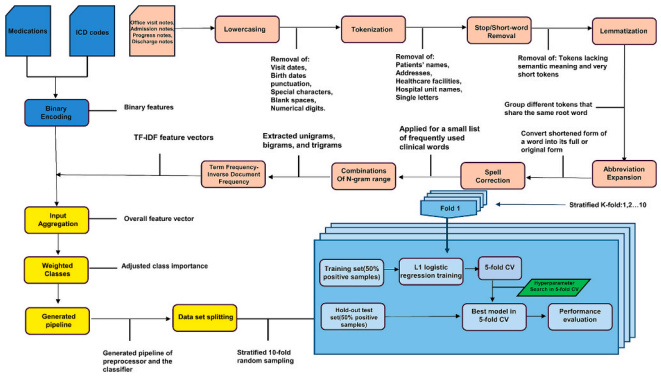


Fig. 2. Method flowchart.

2.5. Classification model training and testing

We trained four machine learning models—logistic regression, Random Forest, Naïve Bayes, and Multi-Layer Perceptron (MLP) to assign a probability to each note, representing the likelihood of the clinical note indicating MCI/ADRD. To deal with imbalanced numbers of positive and negative subsamples, we used class weights inversely proportional to subsample size. To evaluate the performance of our model and assess generalizability, we implemented 5-fold stratified nested cross-validation [28]. Hyperparameter optimization for all models was conducted using a grid search [29] approach with internal cross-validation. To evaluate model performance and assess generalizability, 5-fold stratified nested cross-validation was implemented. For logistic regression, the regularization strength was tuned to balance model complexity and predictive performance, with L1 regularization applied for automated feature selection. For Random Forest, the optimization process involved varying the number of trees, the depth of each tree, the minimum samples required to split a node, and the minimum samples required at a leaf node to identify the optimal parameter combination. For Naïve Bayes, the feature smoothing parameter was adjusted to improve predictive accuracy while ensuring alignment with the model's probabilistic assumptions. For Multi-Layer Perceptron, the hyperparameters tuned included the architecture of the hidden layers, the activation functions, the learning rate strategy, the optimization algorithm, and the strength of regularization to achieve optimal performance…To compare the informativeness of different data types, we trained models with the following input combinations: (1) clinical notes, ICDs, and medications combined, (2) clinical notes only; (3) ICDs only; and (4) medications only. Feature importance analysis was performed to determine which variables had the most significant impact on the model's predictions. LASSO regularization [30] was employed in logistic regression for automated feature selection, with the relative importance of each feature assessed based on the magnitude of the resulting regression coefficients.

2.6. Model performance metrics

Model performance was evaluated using accuracy, precision, recall, specificity, F1-score, area under the AUROC [31], and area under the AUPRC [32]. For each metric, we present micro-average performance metrics for positive and negative diagnoses of MCI/ADRD. We conducted 1000 iterations of bootstrapping [33] to obtain the 95 % confidence intervals (CI). Additionally, confusion matrices [34] for various training and testing datasets further illustrate the model's performance across different data splits. Error analysis was conducted to identify the primary sources of misclassification.

2.7. Generalizability experiments

To evaluate the model's generalizability across institutions and to enhance robustness by incorporating data from both, we conducted five experiments: 1) MGH as Training Set, BIDMC as Testing Set: We trained the model exclusively with data from MGH and tested the model on data from BIDMC. 2) BIDMC as Training Set, MGH as Testing Set: We trained the model exclusively with BIDMC and tested the model on MGH data. 3) MGH + BIDMC Training Set, MGH + BIDMC Testing Set: Training data came from both MGH and BIDMC, as did testing data. 4) MGH + BIDMC Training Set, MGH Testing Set: Training data came from both MGH and BIDMC, testing data from MGH only. 5) MGH + BIDMC Training Set, BIDMC Testing Set: Training data from both MGH and BIDMC, testing data from BIDMC only.

2.8. Performance in unselected / random EHR samples

The sample utilized for training and testing is, by construction, enriched for "positive" cases, i.e. there are more MED+/ICD+, MED+/ICD-, and MED-/ICD+, and fewer MED-/ICD- cases than would be present in a random sample. Thus, the performance statistics calculated for our cohort do not represent the performance that we would expect in a general, unselected hospital population. To obtain an unbiased estimate of model performance, we first estimated the error rates within each of the 4 groups (P_e++, P_e+-, P_e-+, P_e–), where P_e++ represents the error rate for MED+/ICD + cases, P_e+- represents the error rate for MED+/ICD − cases, P_e-+ represents the error rate for MED−/ICD + cases, and P_e– represents the error rate for MED−/ICD − cases. Next, we estimated the prevalence (p) of each group in a general hospital population by randomly sampling 500 patients from BIDMC and 500 patients from MGH (N = 1000). The prevalence estimates were calculated as the proportion of patients falling into each group, where p++ is the proportion of the population classified as MED+/ICD+, p+- is the proportion classified as MED+/ICD−, p-+ is the proportion classified as MED−/ICD+, and p– is the proportion classified as MED−/ICD − . The prevalences and error rates are then combined to give an overall expected error *P[E]* rate using the equation(1) [35]:

$$P[E] = (P_e{+}{+} \times p{+}{+}) + (P_e{+}{-} \times p{+}{-}) + (P_e{-}{+} \times p{-}{+}) + (P_e{-}{-} \times p{-}{-}) \quad (1)$$

3. Results

3.1. Patient population

Fig. 1 presents the CONSORT diagram illustrating the cohort selection process. The MGH cohort comprised 2,058 patients with 3,332 visits, while the BIDMC cohort included 1,819 patients with 3,479 visits. A total of 112 cases, accounting for 765 visits, were excluded due to uncertain manual annotations. After categorizing patients into four sampling groups, the final counts were as follows: 3,626 patients with 5,612 visits. Specifically, the 'MED + ICD+' group had 133 patients with 1,751 visits; the 'MED + ICD-' group included 466 patients with 481 visits; the 'MED- ICD+' group consisted of 214 patients with 214 visits; and the 'MED- ICD-' group contained 2,813 patients with 3,166 visits. In the subgroup analysis, the 'MED + ICD+' group had 121 ADRD/MCI-positive patients with 1,643 visits and 12 MCI/ADRD-negative patients with 108 visits. The 'MED + ICD-' group included 338 ADRD/MCI-positive patients with 353 visits and 121 MCI/ADRD-negative patients with 128 visits. The 'MED- ICD+' subgroup comprised 125 ADRD/MCI-positive patients with 125 visits and 89 MCI/ADRD-negative patients with 89 visits. The largest group, 'MED- ICD-', included 522 ADRD/MCI-positive patients with 592 visits and 2,291 MCI/ADRD-negative patients with 2,574 visits. The cohort was selected to ensure sufficient patients with MCI/ADRD diagnoses for model training. Consequently, the final MGH and BIDMC cohort included 1,106 (30.5 %) MCI/ADRD-positive patients from 1,634 visits and 2,587 MCI/ADRD-negative patients from 3,978 visits.

Table 1 summarizes the baseline characteristics of the cohort. The average age was 67.6 years, with 47 % of patients being male and 53 % female. Racial distribution included 16.0 % Black or African American, 1.5 % Asian, 70.5 % White, and 11.94 % categorized as 'Other'. The MCI/ADRD diagnosis rate was 30.5 %. Among the four sampling groups, the 'MED + ICD+' group had 10.94 %, the 'MED + ICD-' group had 30.56 %, the 'MED- ICD+' group had 11.3 %, and the 'MED- ICD-' group had 47.2 %.

Table 1
Cohort characteristics.

Characteristic	Value(N = 3,626)
Age [(a)] (years, mean (SD))	67.6 (16.7)
Sex, n (%)	
Male	1704 (47 %)
Female	1922 (53 %)
Race, n (%)	
Black or African American	580 (16.0 %)
Asian	55 (1.5 %)
White	2558 (70.5 %)
Other [(b)]	433 (11.94 %)
MCI/ADRD diagnosis, n (%)	1106 (30.5 %)
MED + ICD + group	121 (10.94 %)
MED + ICD- group	338 (30.56 %)
MED-ICD + group	125 (11.3 %)
MED-ICD- group	522 (47.2 %)

(a) Age at baseline for the first visit in the study period. (b) 'Other' includes 'unknown', 'declined to answer', 'American Indian or Alaska Native' and 'Native Hawaiian or other Pacific Islander'.

3.2. Model performance

Performance results for predicting MCI/ADRD chart diagnoses are presented in Table 2, which varies the model inputs, and Table 3, which varies the training and testing cohorts. Table 2 presents the average performance for logistic regression models using ICD Only, Med Only, Note Only, and ICD + MED + Note inputs in the MGH + BIDMC training sets and MGH + BIDMC testing sets. The findings indicate a clear pattern in the performance of logistic regression models based on different input data types. Models that incorporate textual note data, either alone or in combination with ICD codes and medication data, consistently outperform models using only ICD codes or only medication data across all performance metrics, with an accuracy of 0.89, specificity of 0.90, AUROC of 0.95, and AUPRC of 0.95. In Table 3, the highest performance was observed when the MGH + BIDMC training set was tested on the MGH set, achieving an AUROC of 0.98, an AUPRC of 0.98, an accuracy of 0.93, a specificity of 0.96, a precision of 0.96, an F1 score of 0.93 and a recall of 0.91.

Fig. 3 provides a comparative analysis of various training and testing approaches using ROC curves (left panel) and Precision-Recall (PR) curves (right panel). The highest ROC observed is 0.99 for the MGH training and MGH testing set. The Precision-Recall curves demonstrate the trade-off between precision and recall, with the highest Precision-Recall AUC being 0.98 for multiple model configurations, including MGH training tested on the MGH set and MGH + BIDMC training tested on the MGH set.

Table 4 presents the average performance, and 95 % confidence intervals of machine learning models trained on combined ICD codes, medications, and clinical notes from the MGH and BIDMC datasets, evaluating metrics such as AUROC, AUPRC, accuracy, precision, recall, and specificity. Logistic regression demonstrated the highest overall performance, achieving an AUROC of 0.97 and an AUPRC of 0.97, reflecting exceptional discriminative ability and a strong balance between precision and recall. It also achieved high accuracy (0.91), specificity (0.94), and precision (0.92), highlighting its reliability in both correctly identifying negative cases and minimizing false positives. Its recall of 0.88 further underscored its robustness in capturing true positives, contributing to a balanced F1 score of 0.90. Random Forest demonstrated robust discriminative ability, with an AUROC of 0.94 and an AUPRC of 0.90, though its lower recall (0.67) indicated limitations in identifying positive cases. The Multilayer Perceptron performed competitively, achieving an AUROC and AUPRC of 0.88, though with reduced recall (0.68) and an F1 score of 0.76. In contrast, Naïve Bayes showed the lowest performance, with an AUROC of 0.63 and an AUPRC of 0.56, highlighting challenges in effectively handling the dataset. Overall, logistic regression consistently outperformed other models,

Table 2
Average performance and [95 % confidence intervals] for logistic regression models using ICD Only, Med Only, Note Only, and ICD + MED + Note in the MGH + BIDMC training sets and MGH + BIDMC testing sets.

Model Input	Accuracy	Specificity	F1-score	Recall	Precision	AUROC	AUPRC
ICD Only	0.54	0.37	0.60	0.7	0.53	0.54	0.69
	[0.52–0.55]	[0.35–0.38]	[0.59–0.62]	[0.69–0.72]	[0.51–0.55]	[0.52–0.55]	[0.68–0.70]
Med Only	0.56	0.43	0.61	0.69	0.55	0.56	0.70
	[0.55–0.58]	[0.41–0.45]	[0.60–0.63]	[0.68–0.71]	[0.53–0.57]	[0.55–0.58]	[0.69–0.71]
Note Only	0.88	0.90	0.88	0.87	0.88	0.94	0.94
	[0.87–0.91]	[0.89–0.93]	[0.86–0.90]	[0.85–0.88]	[0.87–0.91]	[0.93–0.95]	[0.93–0.95]
ICD + MED + Note	0.91	0.94	0.9	0.88	0.92	0.97	0.97
	[0.90–0.92]	[0.92–0.95]	[0.89–0.92]	[0.89–0.92]	[0.91–0.94]	[0.96–0.98]	[0.96–0.98]

AUROC: Area Under the Receiver Operating Characteristic curve, shows model's ability to distinguish between classes.
AUPRC: Area Under the Precision-Recall Curve, summarizes the precision and recall across different thresholds.
Inputs: ICD Only: Models using only International Classification of Diseases codes. Med Only: Models using only medication data.
Note Only: Models using only textual note data. ICD + MED + Note: Models combining ICD codes, medication data, and textual note data.
Data Sets: MGH + BIDMC: Data derived from Massachusetts General Hospital and Beth Israel Deaconess Medical Center.

Table 3
Average performance and [95% confidence intervals] for logistic regression model using all features in the different testing sets.

Training set	Testing set	Accuracy	Specificity	F1-score	Recall	Precision	AUROC	AUPRC
MGH+BIDMC	BIDMC	0.86	0.86	0.84	0.85	0.82	0.92	0.91
		[0.84–0.87]	[0.83–0.87]	[0.82–0.86]	[0.83–0.88]	[0.80–0.84]	[0.91–0.94]	[0.90–0.93]
MGH	BIDMC	0.90	0.92	0.83	0.88	0.79	0.95	0.90
		[0.89–0.92]	[0.91–0.93]	[0.81–0.85]	[0.86–0.90]	[0.77–0.83]	[0.93–0.96]	[0.89–0.92]
BIDMC	MGH	0.94	0.96	0.91	0.90	0.91	0.98	0.98
		[0.94–0.95]	[0.96–0.97]	[0.89–0.92]	[0.88–0.91]	[0.91–0.93]	[0.97–0.98]	[0.97–0.98]
MGH + BIDMC	MGH	0.93	0.96	0.93	0.91	0.96	0.98	0.98
		[0.92–0.95]	[0.95–0.98]	[0.92–0.95]	[0.88–0.92]	[0.94–0.99]	[0.97–0.99]	[0.98–0.99]
MGH + BIDMC	MGH + BIDMC	0.91	0.94	0.9	0.88	0.92	0.97	0.97
		[0.90–0.92]	[0.92–0.95]	[0.89–0.92]	[0.89–0.92]	[0.91–0.94]	[0.96–0.98]	[0.96–0.98]

The bootstrapping results in 95 % confidence intervals are in parenthesis. ACC – accuracy, Spec –specificity, AP – average precision, AUROC – Area under the receiver operating characteristic curve, AUPRC – area under the precision-recall curve. Data Sets: MGH: Data derived from Massachusetts General Hospital. BI: Data derived from Beth Israel Deaconess Medical Center. MGH + BIDMC: Data derived from Massachusetts General Hospital and Beth Israel Deaconess Medical Center.

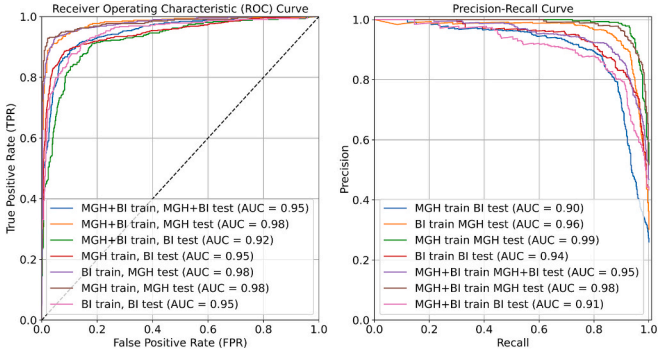


Fig. 3. Left panel: Comparative analysis of all training/testing approaches using Receiver Operating Characteristic (ROC) curves. Right panel: Comparative analysis of all training/testing approaches using Precision-Recall curves.

demonstrating robust and reliable performance across all evaluated metrics.

Fig. 4 shows the confusion matrices representing various training/testing experiments in the context of predicting MCI/ADRD. The columns correspond to predicted MCI/ADRD status, while rows represent the ground truth classification based on chart review. The model trained and tested on MGH data shows the highest accuracy, with 98.17 % for negative and 91.48 % for positive predictions. Conversely, models trained on one dataset and tested on another exhibit lower performance, particularly in positive predictions. Combining both datasets for training (MGH + BI) and testing on the combined or individual datasets yields intermediate performance.

Fig. 5 compares the performance of the model across the four sampling groups, 'MED-ICD-', 'MED-ICD+', 'MED + ICD-', and 'MED + ICD+'. Performance is consistently higher for all metrics except recall in the patients with congruent ICD and medication information ('MED-ICD-', 'MED + ICD+' subgroups) across most metrics, and lower for patients with 'mixed' information ('MED + ICD-', 'MED-ICD+'). In terms of accuracy, the 'MED + ICD+' group performed best (0.94) and 'MED-ICD+' performed worst (0.81). Precision was highest for 'MED-ICD-' (0.86) and lowest for 'MED + ICD-' (0.73). Recall was best for 'MED-ICD+' (0.96) with little difference among the other groups. The F1 Score was highest for 'MED-ICD-' (0.89) and lowest for 'MED + ICD-' (0.81). AUROC was the same for 'MED-ICD-', 'MED + ICD-', and 'MED + ICD+' (0.97), while 'MED-ICD+' was the lowest (0.88). AUPRC was highest for 'MED-ICD-' (0.95) and lowest for 'MED-ICD+' (0.90).

3.3. Features importance

Fig. 6 shows the coefficient values of the top 15 features selected during model training. Notably, the presence of the word "dementia" in a note emerged as the most informative feature, followed closely by the prescription of MCI/ADRD-related medications, including donepezil, aricept, rivastigmine, and memantine. Other top 15 MCI/ADRD-related keywords included "cognitive impairment", "Alzheimer", "MCI", "memory", "cognitive", and "decline".

3.4. Performance in unselected / random EHR samples

We calculated the expected error rate in a general hospital population following the procedure described in the Methods section. In the random sample selected (N = 1000, 500 from MGH, 500 from BIDMC), the proportion (and numbers) falling within each of the 4 groups were: MED + ICD+, P_e++ = 0.1% (n = 1); MED + ICD−, P_e+-0.4% (n = 4); MED − ICD+, P_e-+ = 0.3 % (n = 3); MED − ICD−, P_e– = 0.2 % (n = 2). Combining these with the error rates in each group (p++, p+-, p-+, p–),

Table 4
Average performance and [95 % confidence intervals] for different machine learning models using ICD + MED + Note in the MGH + BIDMC training sets and MGH + BIDMC testing sets.

Model	Accuracy	Specificity	F1-score	Recall	Precision	AUROC	AUPRC
Random Forest	0.82 [0.81–0.83]	0.95 [0.94–0.96]	0.78 [0.76–0.79]	0.67 [0.65–0.69]	0.92 [0.91–0.94]	0.94 [0.93–0.95]	0.90 [0.87–0.92]
Naïve Bayes	0.6 [0.59–0.62]	0.61 [0.59–0.64]	0.58 [0.57–0.59]	0.59 [0.57–0.60]	0.57 [0.55–0.60]	0.63 [0.61–0.64]	0.56 [0.54–0.59]
Multilayer Perceptron	0.80 [0.80–0.82]	0.91 [0.90–0.93]	0.76 [0.75–0.78]	0.68 [0.67–0.69]	0.88 [0.85–0.90]	0.88 [0.86–0.89]	0.88 [0.86–0.90]
Logistic Regression	0.91 [0.90–0.92]	0.94 [0.92–0.95]	0.9 [0.89–0.92]	0.88 [0.89–0.92]	0.92 [0.91–0.94]	0.97 [0.96–0.98]	0.97 [0.96–0.98]

AUROC: Area Under the Receiver Operating Characteristic curve, shows model's ability to distinguish between classes.
AUPRC: Area Under the Precision-Recall Curve, summarizes the precision and recall across different thresholds.
Data Sets: MGH + BIDMC: Data derived from Massachusetts General Hospital and Beth Israel Deaconess Medical Center.

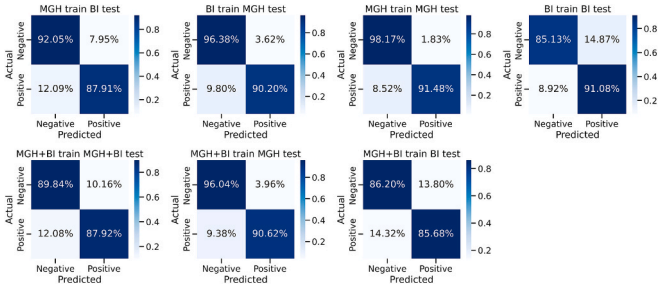


Fig. 4. Comparative analysis of all training/testing approaches using confusion matrix.

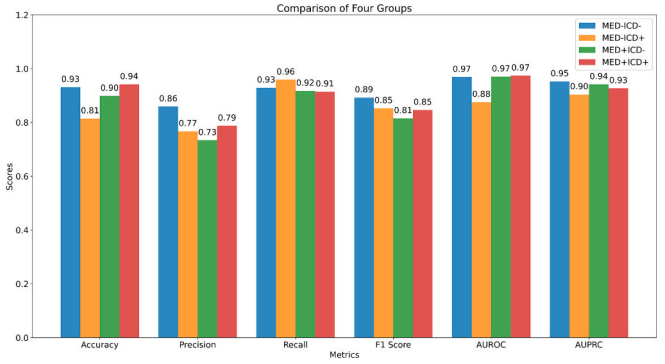


Fig. 5. Performance Evaluation of Different Subgroups Across Multiple Metrics.

we obtain as an estimate of the overall (unbiased) error rate:

$$P[E] = (P_e{+}{+} \times p{+}{+}) + (P_e{+}{-} \times p + {-}) + (P_e{-}{+} \times p{-}{+}) + (P_e{-} \times p{-})$$
$$= (0.06 \times 0.001) + (0.1 \times 0.004) + (0.19 \times 0.003) + (0.07 \times 0.002)$$
$$= 0.00006 + 0.0004 + 0.00057 + 0.00014 = 0.00117 \approx 0.12\ \%.$$

The overall error rate in the general hospital population is estimated to be 0.12 %, i.e. accuracy of 99.88 %.

3.5. Error analysis

We conducted a manual review of cases to gain qualitative insights into reasons for model errors. False positives arose primarily from clinical notes describing symptoms resembling MCI/ADRD, such as memory loss and cognitive decline attributable to alternative causes such as depression and anxiety. Conversely, false negatives arose primarily from notes from specialists seeing patients with MCI/ADRD for specialized care in other areas of medicine, such as nephrology or gynecology, who commented sparsely on issues related to the MCI/ADRD diagnosis in their notes.

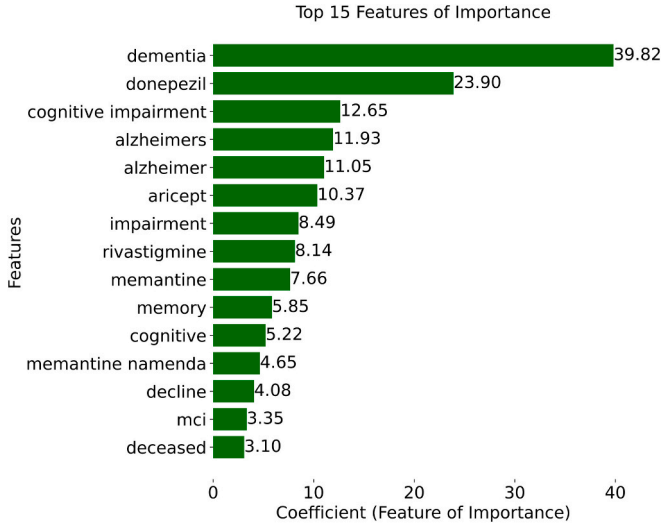


Fig. 6. Top 15 important features based on model coefficients.

4. Discussion

Our machine learning-based automated EHR phenotyping model accurately identifies patients diagnosed with MCI/ADRD using unstructured clinical notes, ICD codes, and MCI/ADRD medications. Models incorporating textual notes, either alone or combined with ICD codes and medication data, consistently outperformed those relying solely on ICD codes or medication data. The ICD + MED + Note model achieved the highest performance, underscoring the importance of integrating diverse data sources for enhanced accuracy and reliability. As shown in Table 2, models using only ICD codes exhibit lower specificity (i.e., higher false positive rate). In contrast, models based solely on clinical notes demonstrate superior performance, with notable differences in AUROC (0.94 vs. 0.54), AUPRC (0.69 vs. 0.95), and F1-score (0.78 vs. 0.60).

Our analysis revealed an important finding that the information in clinical notes leads to much better performance than using ICD codes alone [36,37]. For example, the ICDs may be triggered by a broad range of cognitive symptoms that are not necessarily due to MCI or dementia. This is more likely in the older population where other disease conditions are common, including medication side effects or interactions, depression /mood difficulties [38], hypothyroidism, substance use (such as alcohol and marijuana), and sleep difficulties [39]. The accurate diagnosis of MCI or dementia requires comprehensive testing [40–42], including cognitive testing, physical examination, and often neuroimaging [43]. The clinical notes often contain more detailed descriptions and therefore more information about the ground truth.

The performance indicates good generalizability across sites.

Nevertheless, there are notable site differences. Performance on MGH test data was better compared to performance on BIDMC test data, regardless of the source (BIDMC or MGH) of the training data, suggesting that the MGH dataset may present fewer complexities or challenges. Conversely, adding data from BIDMC to the training dataset resulted in a slight decline in performance on tests conducted at BIDMC. Specifically, the AUROC decreased from 0.98 to 0.92, and the AUPRC decreased from 0.98 to 0.91. These changes represent a minimal impact on the model's performance. The alignment of the features identified for retention within the model with existing medical knowledge and their consistency with an established understanding of dementia and MCI/ADRD medications suggests that the model has learned a reasonable, interpretable pattern. In the future, we hope to apply this model to identify MCI/ADRD patients from electronic health records at scale, thus creating opportunities for large-scale EHR-based studies.

A strength of our approach is the use of data across two health networks (MGH and BIDMC); most published studies focus on single-site data [36,44–46]. This multi-site comparison allowed for a broader validation of our model, showing consistency in performance metrics such as accuracy, specificity, and AUC across different institutional datasets. Our model achieved high AUROC (0.98), demonstrating robust discrimination across testing scenarios. We also emphasize the importance of the highly observed AUPRC, particularly as it relates to the clinical relevance in contexts where class imbalance is pronounced. The high AUPRC indicates that our model effectively identifies "rare" events, such as ADRD/MCI diagnoses, from large healthcare datasets. This strong performance in correctly identifying true positive cases is crucial for accurately diagnosing fewer common conditions with significant clinical implications.

Additionally, our study included a larger patient cohort than those typically reported [36,44–46], which enhances the generalizability of our findings. References to other studies comparing site performance are scarce, making our contributions significant for future multicentric studies. The testing performance was generally better on MGH data than on BIDMC data, regardless of the site of origin of the training data. This suggests that there may be a larger proportion of difficult or ambiguous cases in the BIDMC test set. Nevertheless, test performance was excellent for both sites, suggesting generalizability across notes written within different medical institutions.

Our study has important limitations. While our experiments included two medical centers, these are located in the same geographic region (Boston, United States), and may thus not be representative of other US and non-US populations. Thus, future studies that utilize our model across different hospitals and EHR systems should check for performance biases that might arise due to different demographics, larger sample sizes, bias in data collection, and EHR data stored formats.

An additional limitation is that our model does not identify specific subtypes of ADRD and provides no information about the severity of ADRD or MCI, which are clinically relevant aspects. While the exclusion of uncertain manual annotations reduced noise in the training data, it may have biased the model toward clearer cases, potentially limiting its applicability to more ambiguous cases in real-world scenarios. Moreover, incorporating advanced large language models (LLMs) like GPT may enhance the feature extraction and interpretation of clinical notes by handling complex medical language and context-specific nuances more effectively – an approach we did not use. Differences in model performance between MGH and BIDMC datasets suggest potential overfitting to institutional-specific documentation patterns, highlighting the need for further generalizability testing across varied datasets. Overall, this study represents an important step towards unlocking the vast potential of EHR data to advance our understanding of mild cognitive impairment and dementia and enables various downstream studies.

In conclusion, our model combining the clinical notes, ICD codes, and medications from the EHR system provides accurate MCI/ADRD phenotyping. This approach enables large-scale EHR-based studies that can improve our understanding of MCI/ADRD, facilitate early diagnosis, and support the development of targeted interventions. In the future, this work will enable important downstream large-scale analyses to understand various aspects of MCI/ADRD.

Funding.

This work was supported by grants from the National Institutes of Health (NIH): R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, and National Science Foundation (NSF):2014431.

Data Availability

All data and computer code used to produce the figures and tables will be available at the time of publication here: (https://github.com/rockey1006/Automated_ADRD-MCI) and here (https://bdsp.io/projects/jpCoV5N1MsB9a5QsgprQ/).

Declarations

Ethics approval and consent to participate

EHR data was extracted under protocols approved by the Massachusetts General Hospital (MGH) and Beth Israel Deaconess Medical Center (BIDMC) Institutional Review Boards (IRB) with waivers of informed consent. The relevant IRB protocol numbers are 2013P001024 and 2023P000487 for MGH, and 2022P000417 for BIDMC.

Consent for publication

All authors have approved the manuscript for publication.

CRediT authorship contribution statement

Ruoqi Wei: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation. **Stephanie S Buss:** Writing – review & editing, Data curation. **Rebecca Milde:** Data curation, Conceptualization. **Marta Fernandes:** Writing – review & editing, Methodology. **Daniel Sumsion:** Writing – review & editing. **Elijah Davis:** Writing – review & editing. **Wan-Yee Kong:** Writing – review & editing. **Yiwen Xiong:** Writing – review & editing. **Jet Veltink:** Writing – review & editing. **Samvrit Rao:** Writing – review & editing. **Tara M. Westover:** Writing – review & editing. **Lydia Petersen:** Writing – review & editing. **Niels Turley:** Writing – review & editing. **Arjun Singh:** Writing – review & editing, Methodology, Conceptualization. **Sudeshna Das:** Writing – review & editing, Supervision. **Manohar Ghanta:** Writing – review & editing, Conceptualization. **Aditya Gupta:** Data curation. **Jennifer Kim:** Writing – review & editing, Conceptualization. **Emmanuel Mignot:** Writing – review & editing, Supervision. **Dennis Hwang:** Conceptualization. **Gari D. Clifford:** Writing – review & editing, Supervision. **Umakanth Katwa:** Writing – review & editing. **Robert J Thomas:** Writing – review & editing. **Shibani Mukerji:** Writing – review & editing, Supervision, Conceptualization. **Haoqi Sun:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Bai, P. Chen, H. Cai, et al., Worldwide prevalence of mild cognitive impairment among community dwellers aged 50 years and older: a meta-analysis and systematic review of epidemiology studies, *Age Ageing* 51 (2022).

[2] K.B. Rajan, J. Weuve, L.L. Barnes, E.A. McAninch, R.S. Wilson, D.A. Evans, Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060), *Alzheimers Dement.* 17 (12) (2021) 1966–1975.

[3] J. Liss, S. Seleri Assunção, J. Cummings, 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.* 290 (2) (2021) 310–334.

[4] L.E. Hebert, J. Weuve, P.A. Scherr, D.A. Evans, Alzheimer disease in the United States (2010–2050) estimated using the 2010 census, *Neurology* 80 (19) (2013) 1778–1783.
[5] B. Rothman, J.C. Leonard, M.M. Vigoda, Future of electronic health records: implications for decision support, *Mount Sinai J. Med.: A J. Translat. Persona. Med.* 79 (6) (2012) 757–768.
[6] J. Sun, J. Hu, D. Luo, et al., Combining knowledge and data driven insights for identifying risk factors using electronic health records. Paper presented at: AMIA Annual Symposium Proceedings2012.
[7] H. Hemingway, F.W. Asselbergs, J. Danesh, et al., Big data from electronic health records for early and late translational cardiovascular research: challenges and potential, *Eur. Heart J.* 39 (16) (2018) 1481–1495.
[8] X. He, R. Wei, Y. Huang, et al., Develop and validate a computable phenotype for the identification of Alzheimer's disease patients using electronic health record data, *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring.* 16 (3) (2024) e12613.
[9] K.E. Innes, U. Sambamoorthi, The potential contribution of chronic pain and common chronic pain conditions to subsequent cognitive decline, new onset cognitive impairment, and incident dementia: A systematic review and conceptual model for future research, *J. Alzheimers Dis.* 78 (3) (2020) 1177–1195.
[10] E. Jutkowitz, J.P. Bynum, S.L. Mitchell, et al., Diagnosed prevalence of Alzheimer's disease and related dementias in Medicare Advantage plans, *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring.* 12 (1) (2020) e12048.
[11] T.B. Sanders, F.M. Bowens, W. Pierce, B. Stasher-Booker, E.Q. Thompson, W. A. Jones, The road to ICD-10-CM/PCS implementation: forecasting the transition for providers, payers, and other healthcare organizations, *Perspectives in health information management/AHIMA, American Health Information Management Association.* 9 (Winter) (2012).
[12] E. Ford, M. Oswald, L. Hassan, K. Bozentko, G. Nenadic, J. Cassell, Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK, *J. Med. Ethics* 46 (6) (2020) 367–377.
[13] H.-J. Kong, Managing unstructured big data in healthcare system, *Healthcare Informatics Research.* 25 (1) (2019) 1–2.
[14] D.B. Preen, C.A.J. Holman, D.M. Lawrence, N.J. Baynham, J.B. Semmens, Hospital chart review provided more accurate comorbidity information than data from a general practitioner survey or an administrative database, *J. Clin. Epidemiol.* 57 (12) (2004) 1295–1304.
[15] H. Kaur, S. Sohn, C.-I. Wi, et al., Automated chart review utilizing natural language processing algorithm for asthma predictive index, *BMC Pulm. Med.* 18 (1) (2018) 1–9.
[16] A. Noori, C. Magdamo, X. Liu, et al., Development and evaluation of a natural language processing annotation tool to facilitate phenotyping of cognitive status in electronic health records: diagnostic study, *J. Med. Internet Res.* 24 (8) (2022) e40384.
[17] E. Ford, A. Nicholson, R. Koeling, et al., Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? *BMC Med. Res. Method.* 13 (1) (2013) 1–12.
[18] S.J. Price, S.A. Stapley, E. Shephard, K. Barraclough, W.T. Hamilton, Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case–control study, *BMJ Open* 6 (5) (2016).
[19] Daniel Capurro M, Erik van Eaton M, Black R. Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment.
[20] Y. Wang, L. Wang, M. Rastegar-Mojarad, et al., Clinical information extraction applications: a literature review, *J. Biomed. Inform.* 77 (2018) 34–49.
[21] T. He, A. Belouali, J. Patricoski, et al., Trends and opportunities in computable clinical phenotyping: a scoping review, *J. Biomed. Inform.* 140 (2023) 104335.
[22] S. Yang, P. Varghese, E. Stephenson, K. Tu, J. Gronsbell, Machine learning approaches for electronic health records phenotyping: a methodical review, *J. Am. Med. Inform. Assoc.* 30 (2) (2023) 367–381.
[23] C. Shivade, P. Raghavan, E. Fosler-Lussier, et al., A review of approaches to identifying patient phenotype cohorts using electronic health records, *J. Am. Med. Inform. Assoc.* 21 (2) (2014) 221–230.
[24] S.J. Steindel, International classification of diseases, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets, *J. Am. Med. Inform. Assoc.* 17 (3) (2010) 274–282.
[25] E. Ye, H. Sun, M.J. Leone, et al., Association of sleep electroencephalography-based brain age index with dementia, *JAMA Netw. Open* 3 (9) (2020) e2017357–e.
[26] A. Rajaraman, J.D. Ullman, *Mining of massive datasets*, Autoedicion (2011).
[27] J.K. Sparck, A statistical interpretation of term specificity and its application in retrieval, *J. Doc.* 28 (1) (1972) 11–21.
[28] J. Wainer, G. Cawley, Nested cross-validation when selecting classifiers is overzealous for most practical applications, *Expert Syst. Appl.* 182 (2021) 115222.
[29] P. Liashchynskyi, P. Liashchynskyi, Grid search, random search, genetic algorithm: a big comparison for NAS, 2019. arXiv preprint arXiv:191206059.
[30] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Stat Methodol.* 58 (1) (1996) 267–288.
[31] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recogn.* 30 (7) (1997) 1145–1159.
[32] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS One* 10 (3) (2015) e0118432.
[33] A.R. Henderson, The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data, *Clin. Chim. Acta* 359 (1–2) (2005) 1–26.
[34] I. Düntsch, G. Gediga, Confusion matrices and rough set data analysis. Paper presented at, J. Phys. Conf. Ser. (2019).
[35] M. Fernandes, A. Cardall, J. Jing, et al., Identification of patients with epilepsy using automated electronic health records phenotyping, *Epilepsia* 64 (6) (2023) 1472–1481.
[36] C.A. Hane, V.S. Nori, W.H. Crown, D.M. Sanghavi, P. Bleicher, Predicting onset of dementia using clinical notes and machine learning: case-control study, *JMIR Med. Inform.* 8 (6) (2020) e17819.
[37] E. Ford, J.A. Carroll, H.E. Smith, D. Scott, J.A. Cassell, Extracting information from the text of electronic medical records to improve case detection: a systematic review, *J. Am. Med. Inform. Assoc.* 23 (5) (2016) 1007–1015.
[38] R.W. Lam, S.H. Kennedy, R.S. McIntyre, A. Khullar, Cognitive dysfunction in major depressive disorder: effects on psychosocial functioning and implications for treatment, *Can. J. Psychiatry* 59 (12) (2014) 649–654.
[39] R.C. Hamdy, A. Kinser, K. Dickerson, et al., Insomnia and mild cognitive impairment, *Gerontol. Geriatr. Med.* 4 (2018) 2333721418778421.
[40] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, E.M. Stadlan, Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group* under the auspices of department of health and human services task force on Alzheimer's disease, *Neurology* 34 (7) (1984) 939.
[41] G.M. McKhann, D.S. Knopman, H. Chertkow, 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.* 7 (3) (2011) 263–269.
[42] C.R. Jack Jr, D.A. Bennett, K. Blennow, et al., NIA-AA research framework: toward a biological definition of Alzheimer's disease, *Alzheimers Dement.* 14 (4) (2018) 535–562.
[43] C.R. Jack Jr, D.A. Bennett, K. Blennow, et al., A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers, *Neurology* 87 (5) (2016) 539–547.
[44] D.B. Reuben, A.S. Hackbarth, N.S. Wenger, Z.S. Tan, L.A. Jennings, An Automated approach to identifying patients with dementia using electronic medical records, *J. Am. Geriatr. Soc.* 65 (3) (2017) 658–659.
[45] W.-Q. Wei, P.L. Teixeira, H. Mo, R.M. Cronin, J.L. Warner, J.C. Denny, Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance, *J. Am. Med. Inform. Assoc.* 23 (e1) (2016) e20–e27.
[46] J. Xu, F. Wang, Z. Xu, et al., Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records, *Learn. Health. Syst.* 4 (4) (2020) e10246.