




Research and Applications

Leveraging multi-site electronic health data for characterization of subtypes: a pilot study of dementia in the N3C Clinical Tenant

Suchetha Sharma, MS¹, Jiebei Liu , PhD^{2,*}, Amy Caroline Abramowitz, MD³, Carol Reynolds Geary , PhD, MBA, RN⁴, Karen C. Johnston, MD, MSc⁵, Carol Manning, PhD, ABPP-CN⁵, John Darrell Van Horn, PhD, MEng, FOHBM¹, Andrea Zhou, MS⁶, Alfred J. Anzalone, PhD⁷, Johanna Loomba, MS⁶, Emily Pfaff , PhD, MS⁸, Don Brown, PhD⁹

¹School of Data Science, University of Virginia, Charlottesville, VA 22903, United States, ²Department of Systems Engineering, University of Virginia, Charlottesville, VA 22904, United States, ³Department of Psychiatry, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, United States, ⁴Department of Pathology, Microbiology & Immunology, University of Nebraska Medical Center, Omaha, NE 68198-5900, United States, ⁵Department of Neurology, University of Virginia, Charlottesville, VA 22903, United States, ⁶School of Medicine, University of Virginia, Charlottesville, VA 22903, United States, ⁷Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198, United States, ⁸Department of Medicine, North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States, ⁹School of Data Science, Co-Director integrated Translational Health Research Institute of Virginia (iTHRIV), University of Virginia, Charlottesville, VA 22903, United States

*Corresponding author: Jiebei Liu, PhD, Olsson Hall, 151 Engineer's Way, University of Virginia, Charlottesville, VA 22904, USA (mcu2xn@virginia.edu)
Suchetha Sharma and Jiebei Liu contributed equally.

Abstract

Objectives: To provide a foundational methodology for differentiating comorbidity patterns in subphenotypes through investigation of a multi-site dementia patient dataset.

Materials and Methods: Employing the National Clinical Cohort Collaborative Tenant Pilot (N3C Clinical) dataset, our approach integrates machine learning algorithms—logistic regression and eXtreme Gradient Boosting (XGBoost)—with a diagnostic hierarchical model for nuanced classification of dementia subtypes based on comorbidities and gender. The methodology is enhanced by multi-site EHR data, implementing a hybrid sampling strategy combining 65% Synthetic Minority Over-sampling Technique (SMOTE), 35% Random Under-Sampling (RUS), and Tomek Links for class imbalance. The hierarchical model further refines the analysis, allowing for layered understanding of disease patterns.

Results: The study identified significant comorbidity patterns associated with diagnosis of Alzheimer's, Vascular, and Lewy Body dementia subtypes. The classification models achieved accuracies up to 69% for Alzheimer's/Vascular dementia and highlighted challenges in distinguishing Dementia with Lewy Bodies. The hierarchical model elucidates the complexity of diagnosing Dementia with Lewy Bodies and reveals the potential impact of regional clinical practices on dementia classification.

Conclusion: Our methodology underscores the importance of leveraging multi-site datasets and tailored sampling techniques for dementia research. This framework holds promise for extending to other disease subtypes, offering a pathway to more nuanced and generalizable insights into dementia and its complex interplay with comorbid conditions.

Discussion: This study underscores the critical role of multi-site data analyzes in understanding the relationship between comorbidities and disease subtypes. By utilizing diverse healthcare data, we emphasize the need to consider site-specific differences in clinical practices and patient demographics. Despite challenges like class imbalance and variability in EHR data, our findings highlight the essential contribution of multi-site data to developing accurate and generalizable models for disease classification.

Lay Summary

This study aims to enhance our understanding and classification of dementia subtypes using data from multiple healthcare sites. Dementia includes forms like Alzheimer's, Vascular, and Lewy Body dementia, each with unique health conditions. Researchers analyzed data from 9 US sites using a multi-stage approach with machine learning techniques, specifically logistic regression and eXtreme Gradient Boosting (XGBoost).

The methodology involved 3 steps. First, the dataset was refined to focus on well-represented dementia subtypes. Next, advanced techniques balanced the data for fair representation. Finally, machine learning models classified the dementia types based on comorbidities and gender differences, achieving up to 70% accuracy for Alzheimer's and Vascular dementia, but finding Lewy Body dementia more challenging. A hierarchical model was used to address site-specific variations, revealing disparities among sites and improving generalization across populations.

This study highlights the complexity of diagnosing dementia subtypes and the limitations of single-site studies, which often suffer from biases. By leveraging data from multiple sites, the research underscores the importance of multi-site dataset analysis for better generalization. This approach enhances understanding of dementia and provides a framework applicable to other diseases.

Key words: dementia subtypes; electronic health records; machine learning algorithms; comorbidity patterns; multi-institutional studies.

Received: May 17, 2024; Revised: July 19, 2024; Editorial Decision: July 24, 2024; Accepted: August 1, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

The adoption of electronic health records (EHR) data has become widespread in modern healthcare facilities as they provide a centralized platform to maintain comprehensive patient medical information. A broad cross-section of research demonstrates that the analysis of EHR data is helpful in identifying comorbidities associated with various disease subtypes. Despite these advantages, many observational EHR analyses are frequently constrained to single-institution datasets, which introduces biases specific to institutions or regions. Notably, the choice of EHR vendor can play a significant role in impacting the outcomes and generalizability of these studies. In multi-site research, the maturity of the EDW4R team becomes particularly crucial. These biases emerge from multiple factors, including changes in the International Classification of Diseases (ICD) terminology, inconsistencies in coding practices, discrepancies between clinical and billing diagnoses, and variability in the documentation of hospitalized patients.^{1–5} Such challenges highlight the need for conducting multi-institutional studies to mitigate biases and enhance the generalizability and reliability of EHR-based models.

This work aims to develop a generalizable methodology that leverages machine learning techniques on longitudinal EHR data aggregated from multiple institutions across the United States as part of the National Clinical Cohort Collaborative Tenant Pilot (N3C Clinical) effort. Specifically, our methodology comprises a 3-stage process, employing dementia as a use case to explore comorbidity patterns across subtypes.

Dementia represents an important and challenging use case as it encompasses a broad spectrum of neurological disorders characterized by cognitive decline, including memory loss and impaired reasoning, significantly affecting daily activities.^{6,7} Dementia is classified into various subtypes, such as Alzheimer's, Vascular, Lewy Body, Parkinson's, and Frontotemporal dementia, each with unique comorbidities, progression rates, and pathological causes.⁷ The presentation of these disorders poses considerable diagnostic challenges, as evidenced by a study that found only a 35% accuracy rate in diagnosing the correct subtype in tertiary or specialized healthcare centers.⁸ Alzheimer's disease, the most common form of dementia, currently impacts about 6.5 million Americans aged 65 and older, with projections suggesting this number could rise to 13.8 million by 2060 in the absence of significant medical breakthroughs.⁹

Patients with dementia may experience a range of comorbidities, including chronic vascular, metabolic, and mental health conditions, such as diabetes, hypertension, anxiety, sleep disorders, and depression. Although some of these conditions have been studied as risk factors for a particular dementia subtype,^{10–12} it is crucial to understand how they compare across dementia subtypes. Our proposed methodology aims to help identify comorbidity patterns in a set of diagnosis subtypes, which could be a helpful tool in differential diagnosis. The multistep analytic framework can be reused in other subtype analyses utilizing multi-site electronic health data.

Background

Numerous electronic health record (EHR)-based phenotyping projects have explored the potential of digital phenotyping in clinical conditions, including dementia, using data from single organizations such as the Vanderbilt Synthetic Derivative and the MIMIC-III data from Beth Israel Deaconess Medical

Center.^{13,14} However, many initiatives support decentralized analyses across disparate systems by distributing standardized queries to member sites or extracting a subset of EHR data into an aggregate data warehouse,^{15,16} such as the Greater Plains Collaborative.¹⁷ These large-scale data resources offer extensive patient and facility coverage but are challenged by data heterogeneity and gaps.¹

In dementia research, the utilization of EHR has provided valuable insights into understanding the relevant comorbidities and progression. Machine learning techniques have significantly enhanced diagnostic precision for dementia, especially in initial detection and risk assessment, but their use in classifying dementia subtypes has been limited.^{18–20} The traditional method, which employs rule-based algorithms, continues to be a foundational approach.¹⁹ These traditional rule-based methods rely on predefined criteria and logical constraints developed by clinical experts, utilizing specific diagnostic codes, laboratory results, medications, and clinical notes. However, this dependence on clinical expertise introduces variability in diagnostic accuracy due to differences in individual expertise and site-specific biases, leading to inconsistencies and inaccuracies in diagnosing dementia subtypes. Studies indicate that general practitioners in primary care settings accurately diagnose only about half of mild dementia cases, with undiagnosed cases accounting for 50–66% of all dementia cases, highlighting the limitations of rule-based approaches.²¹ Machine learning approaches, although also based on clinical expertise and diagnostic criteria, offer scalability and precision across heterogeneous sites in post-hoc EHR data analyses. Clinical expertise remains the gold standard, yet machine learning enhances consistency and accuracy by reducing individual and site-specific biases. Recent methodologies such as Natural Language Processing (NLP) and cohort analyses for dementia subtype classification now complement traditional approaches by integrating patient demographics, comorbidities, and medication histories.^{20,22,23} These methods use keywords from cognitive tests or clinical notes and ICD diagnoses of dementia to assess dementia risk.^{24,25} Despite these advancements, challenges persist, including the complexity added by diverse data sources and the risk of inaccuracies due to the varying reliability of billing/insurance records and clinical notes.²⁶ Many studies have focused predominantly on Alzheimer's, with limited information on comorbidities associated with other dementia subtypes.²⁵ Additionally, algorithms applied in limited environments, such as the intensive care unit or a single site, have faced obstacles in capturing the diversity of the wider dementia population and presented challenges in replicability and generalizability.^{20,27}

In our study, we aim to address these limitations by using machine learning to classify dementia subtypes based on comorbidities and combining EHR data from multiple sites. This approach aims to enhance subtype identification precision and investigate unique characteristics, overcoming the constraints of single-location studies for a more comprehensive analysis.

Methods

This section elaborates on our dataset, reproducible set of methods like the feature selection process, our sampling methodology, and the classification models utilized.

Furthermore, we detail the methodology adopted for model evaluation, ensuring a comprehensive understanding of the model.

Dataset

Our data was sourced from the Alzheimer's disease and related dementias (AD/ADRD) Tenant (DUR-A97711D) in the N3C Clinical enclave, which compiles EHR data from 9 US sites, covering 119 946 patients living with dementia (PLWD) up to January 2023.²⁸ Participating institutions transfer longitudinal health data from 2018 to the present using 1 of 4 common data models, which is then harmonized to the Observational Medical Outcomes Partnership (OMOP) common data model²⁹ and added to the N3C Clinical enclave using an established ingestion and harmonization approach.³⁰ Data from institutions participating in the ADRD Tenant—specifically, all records of patients diagnosed with dementia—are analyzed in our study, using de-identified patient information from visits between January 1, 2018, and January 9, 2023. To ensure privacy, ages above 89 are uniformly anonymized. Analyses were conducted within the N3C Clinical enclave using Python, PySpark, and SQL.

Study design

Our research extends the foundational framework of EHR phenotyping, which is a systematic approach to identify patients with specific observable characteristics using their medical records, through 4 principal stages: data preparation, algorithm development, algorithm evaluation, and application.^{31,32} Specially, we concentrate on the subtype classification of dementia. Traditional supervised learning methods commonly employed in EHR phenotyping include random forest, logistic regression, and support vector machine.³¹ For our study, we selected logistic regression and eXtreme Gradient Boosting (XGBoost) for their proficiency in binary and nonlinear classification, aiming to predict the dementia subtype of patients from their existing comorbidities and demographic details.

Generalizable 3-stage multi-institution model

Our generalizable multi-institutional methodology employs 3 stages (Figure 1). The first stage focuses on pruning the list of subtypes based on information insufficiency challenges and sample size constraints. The second stage focuses on combining related subtypes and using sampling techniques to balance the cohort in preparation for a machine learning-based classification model. The final stage focuses on machine learning models for classification. We next describe each stage in detail.

Stage 1: subtype evaluation for sample size and information insufficiency challenges

As illustrated in Figure 1 (stage 1), the first step involves pruning of subtypes. This decision is driven by 2 primary considerations:

- 1) *Information insufficiency*: Interpretability is challenging for some of the dementia subtypes where diagnosis based solely on comorbidities is complicated due to symptoms overlapping with other dementias and psychiatric disorders. For example, Fronto Temporal Dementia (FTD), affecting the frontal and temporal lobes, requires advanced imaging to identify specific

atrophy patterns, crucial for differentiating it from Alzheimer's disease and other dementias.³³ Thus, we leverage help from clinical experts on our team to identify subtypes where there is insufficient information to diagnose solely based on comorbidities.

- 2) *Limited sample size*: We remove subtypes with small sample sizes as they restrict the statistical power and reliability of any findings related to that particular subgroup, making it difficult to draw meaningful conclusions. We choose a threshold of 1200 samples as suggested by clinicians on our team.

By excluding subgroups with small sample sizes and information insufficiency challenges, our analysis aims to enhance the clarity and coherence, focusing on dementia subtypes with well-defined clinical presentations and sufficient sample sizes for robust statistical analysis.

Stage 2: combining subtypes and balancing the cohort

Stage 2(a): sampling methods

Before deploying predictive models, analyzing and rectifying class imbalances in the dataset is crucial. To counter the class imbalance in our dementia dataset, we tested various sampling strategies to attain balanced class distribution. Notably, these sampling methods were applied exclusively to the training dataset, while the test dataset retained its original class distribution without any synthetic data:

- 1) *Undersampling methods*: Applied Random Under-Sampling (RUS) and Iterative Hard Thresholding (IHT) to reduce the size of majority class, prioritizing computational efficiency while acknowledging the risk of losing valuable data.^{34,35}
- 2) *Oversampling methods*: Adopted Synthetic Minority Over-sampling Technique (SMOTE) and its derivatives (Borderline SMOTE, SVM SMOTE, ADASYN), to generate synthetic instances of the minority class, enhancing its representation within the dataset.³⁵⁻³⁹
- 3) *Hybrid/Split methods*: Combined oversampling with data cleaning via SMOTETomek and SMOTEENN (SMOTE + Edited Nearest Neighbors) strategies to remove sample overlap and enhance decision boundaries clarity.^{40,41}

A 65% SMOTE to 35% RUS ratio, further refined by Tomek Links, was selected for achieving optimal class balance in our dataset, significantly marked by class skew. This method excelled in key metrics such as Accuracy, ROC (Receiver Operating Characteristic), which evaluates the true positive rate against the false positive rate at various threshold settings, F1 score, and Recall providing a balanced approach to evaluating model performance and effectively mitigating the risks of data loss and overfitting. The choice of RUS to SMOTE ratio was guided by the need to address pronounced class imbalances, with the imbalanced-learn package⁴² facilitating practical application. Detailed comparisons and results of these techniques are available in the [Supplementary Material](#).

Stage 2(b): combining subtypes

Our analysis begins with binary classification among dementia subtypes with relatively large sample sizes. When the model's F-1 score falls below a predetermined threshold and clinical evidence points to pathological similarities between

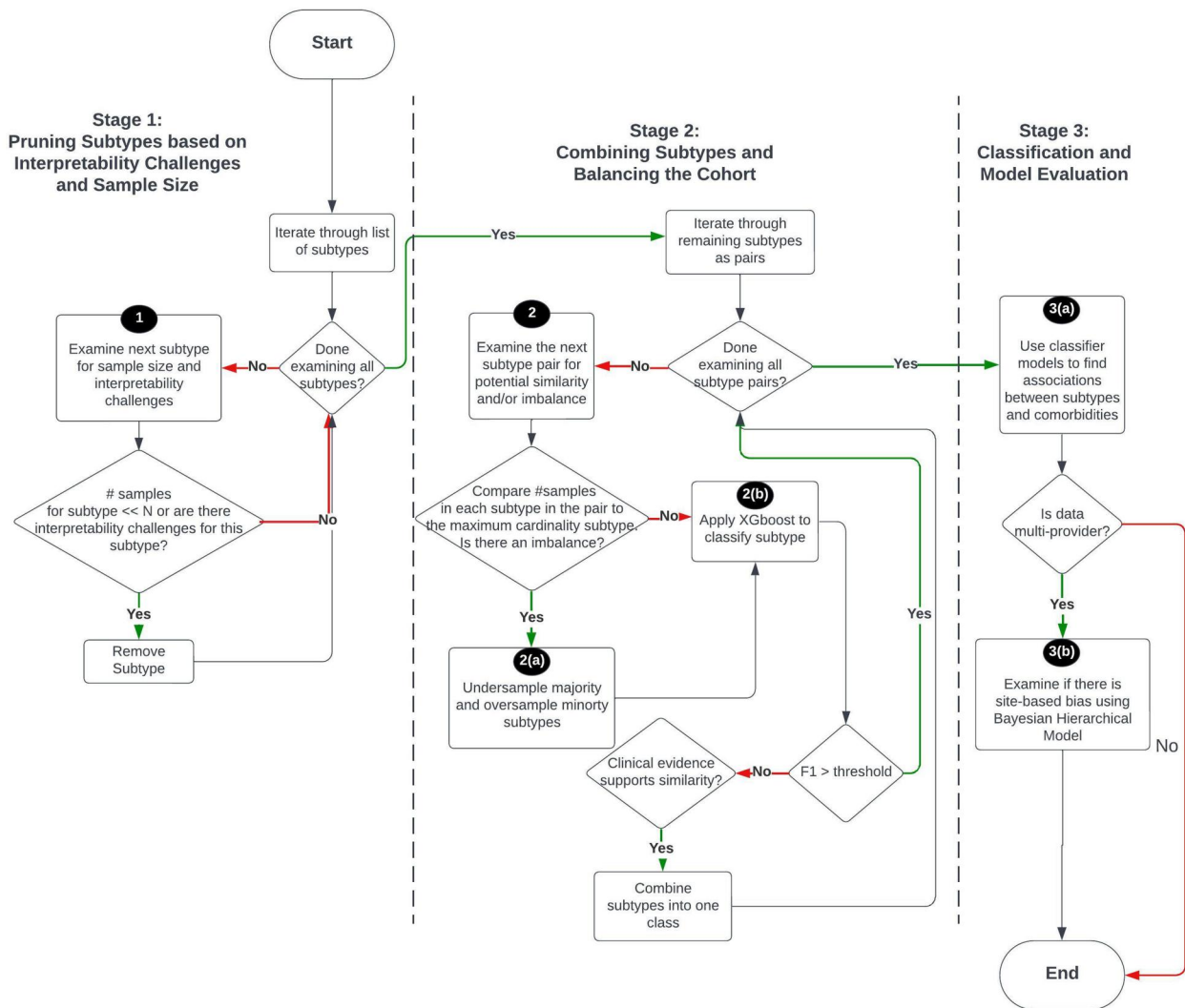


Figure 1. Flowchart of the 3-stage methodology used for the classification of dementia subtypes and the assessment of site-specific biases. (1) Pruning subtypes based on sample size and information insufficiency. (2) Combining and balancing subtypes. (3) Classification, model evaluation, and site bias assessment using Bayesian hierarchical modeling.

subtypes, we combine them. In our experiments, we choose a threshold of 0.55, which is slightly above that of a random classifier. The rationale behind this is that low precision and recall suggest inadequate subtype differentiation. After merging, we proceed to evaluate additional subtypes in sequence, helping to distinguish between key subtypes of dementia.

Stage 3: classification model and diagnostic check

Stage 3(a): classification and model evaluation

Building on our feature selection, we implemented (1) logistic regression⁴³ and (2) eXtreme Gradient Boosting (XGBoost)⁴⁴ classifiers to predict specific dementia subtypes based on patients' comorbidities and demographics, given their effectiveness in binary classifications in medical contexts. Logistic regression served as a straightforward and interpretable baseline, while XGBoost was selected for its advanced capability. XGBoost employs a series of decision trees as base learners, utilizing an induction-based method to incrementally refine the decision trees' accuracy at every step, effectively handling the complex, nonlinear relationships between our features (eg, comorbidities and demographic details) and dementia outcomes.⁴⁴ We validated the models using 5-fold cross-

validation, reporting average accuracy, precision, recall, and F1 scores. To understand the contribution of each feature to the model predictions, we used Shapley⁴⁵ plots, which provide a comprehensive method to attribute the prediction to each feature. These plots revealed the importance of different comorbidities and demographic details, confirming the models' ability to identify key predictors. (The results can be found in [Supplementary Material](#).)

Additionally, to refine XGBoost's parameters, we employed Bayesian hyperparameter tuning,⁴⁶ beginning with a grid search using the scikit-learn⁴⁷ Python package to select optimal hyperparameters from pre-specified ranges. This initial grid search was followed by manual tuning based on its outcomes. The hyperparameter ranges provided to the grid search were as follows: `n_estimators` (100-500, in increments of 50), `learning_rate` (0.001-0.1, log-uniform), `max_depth` (8-24, in increments of 4), `gamma` (8-24, in increments of 4), `verbosity` (fixed at 0), and `objective` (fixed at 'binary'). The Bayesian optimization process identified the best parameters as `colsample_by_tree`: 20.0, `learning_rate`: 0.0153, `max_depth`: 12, `n_estimators`: 250, `objective`: "binary: logistic," and `verbosity`: 0.

Metrics

In evaluating classifiers, it is essential not only to achieve high true positive rates but also to reduce false positives and negatives. This is because the goal of these models is to classify patients that belong to a subtype on the basis of their existing comorbidities, sex, and the data site ID. To examine biases resulting from site-specific data skews, we tested models with and without the site ID. Our evaluation focused on the F1 score, which is a combined measure of Precision and Recall. A high F1 score suggests a balanced rate of true positives and appropriate management of false positives and negatives.⁴⁸

Stage 3(b): Bayesian hierarchical model

Our diagnostic analysis evaluated the impact of site-specific features on dementia subtype classification models using Bayesian hierarchical modeling. This method excels at handling multilevel data structures and mitigating site-related biases, beneficial for datasets with distinct layers or groupings like location.^{49,50} We explored 3 hierarchical modeling strategies:

- Complete Pooling assumes homogeneity across sites.
- No Pooling (Unpooling) treats each site distinctly, ignoring shared patterns.
- Partial Pooling combines site-specific detail with collective trends.⁵¹

These strategies enable an in-depth analysis of multi-site data, distinguishing whether dementia subphenotype classification variations are due to genuine comorbidities or site-specific characteristics. Utilizing the PyMC framework⁵² for Hierarchical Bayesian Model allows for an individualized analysis of each site, accommodating variations in how different predictors—such as demographics and comorbidities—affect dementia diagnosis. This approach enriches our classification efforts and improves our ability to discern the subtle dynamics between site-specific factors and dementia subtypes.

Results

In this section we present the results obtained from our 3-stage methodology.

Stage 1 refined the dataset to include only individuals aged 40 and older, removing entries with incomplete data, such as unknown sex. We further excluded cases of Fronto Temporal Dementia (FTD) and Dementia Not Otherwise Specified (NOS) due to either information insufficiency issues or small sample sizes. The analysis was limited to patients diagnosed with only one type of dementia, narrowing our study cohort to Alzheimer’s, Vascular, and Dementia with Lewy Bodies subtypes. [Supplementary Material](#) provides in-depth dataset details and subtype information. [Figure 2](#) visualizes the cohort construction, highlighting the exclusion of certain subtypes in this initial stage.

In stage 2, we addressed data imbalance using previously described sampling methods and conducted binary classification between Alzheimer’s disease and Vascular dementia, chosen for their larger sample sizes. The feature selection for XGBoost model at this stage was guided by clinical expertise and empirical data density, focusing on comorbidities which are known to affect dementia progression, including hypertension, diabetes, sleep disorders, anxiety, obesity, and depression. The hierarchically clustered heatmap ([Figure 3](#)) visualized comorbidity frequency and co-occurrence, aiding in this selection. The demographics breakdown of the subtypes in the study cohort is described in the below [Table 1](#).

With an accuracy of 59.06%, the XGBoost model struggled to clearly differentiate between Alzheimer’s and Vascular (shown in the [Supplementary Material](#)), which aligns with clinical observations of mixed dementia presentations where individuals exhibit pathological features of both dementia subtypes. To address this, Alzheimer’s and Vascular were combined into a single category to distinguish from Dementia with Lewy Bodies, enhancing our ability to differentiate among these key dementia subtypes. This is shown in the final row of the attrition table.

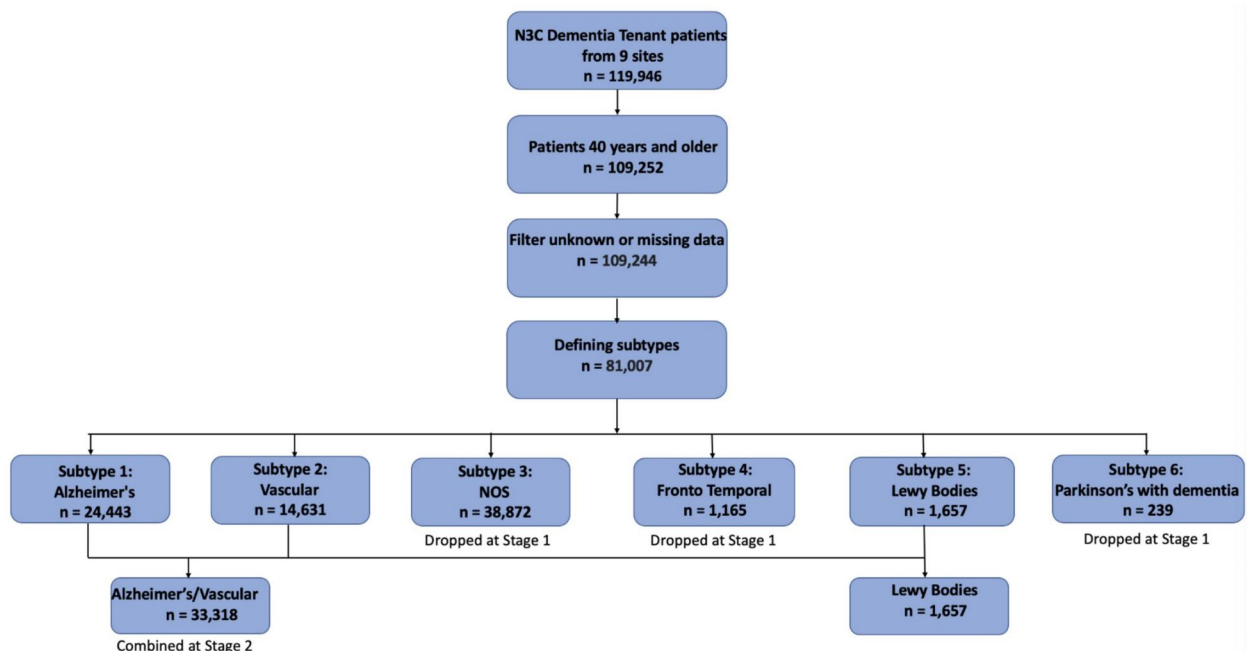


Figure 2. Attrition table describing the study cohort construction.

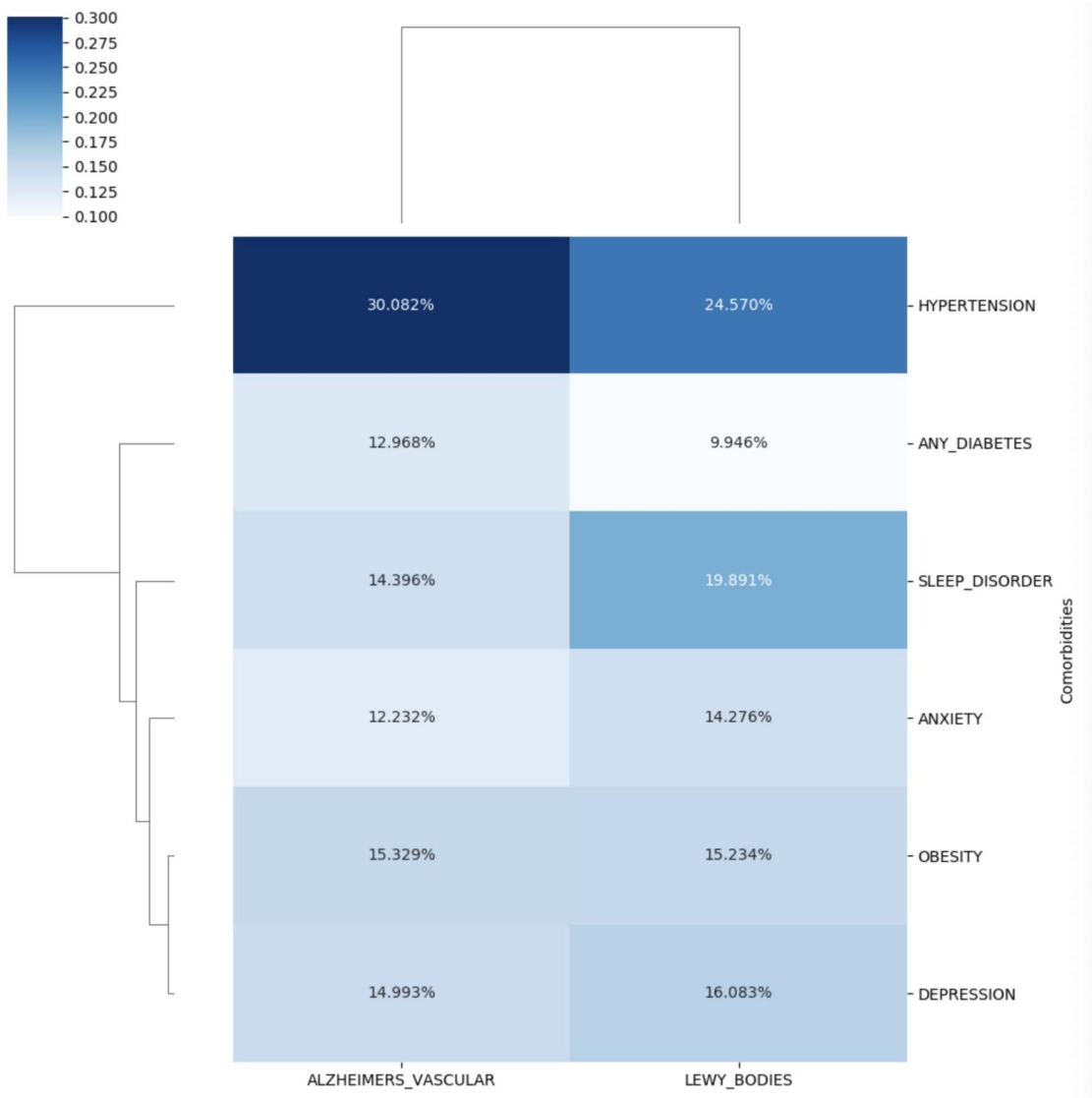


Figure 3. Clustered heatmap of subtypes and comorbidities. Note that darker regions show high density and the tree on the left shows a hierarchy of clusters obtained based on patient comorbidities. The default Euclidean distance is employed as the metric used for the dendrograms as we are interested in the relative strengths of the comorbidities (ie, number of patients affected by it) in determining clusters rather than associations/correlations among the comorbidities.

Model Performance Analysis in Stage 3 compared logistic regression and XGBoost classifiers, with outcomes detailed in Table 2 revealing several critical insights.

The logistic regression model achieved an overall accuracy of 67%, slightly surpassing the performance of a random classifier, suggesting its limited efficacy in accurately classifying all patients within our dataset when based solely on comorbidities and sex. The model demonstrated high precision of 97% for Alzheimer's/Vascular dementia, indicating a strong ability to identify true positive cases within this category. However, it faced significant challenges in classifying Dementia with Lewy Bodies, as evidenced by a much lower precision of around 10%. This poor performance is likely due to the relative underrepresentation of Dementia with Lewy Bodies cases in the dataset. The recall rates for both dementia types are about 60%, pointing to a substantial number of false negatives for both conditions.

The XGBoost model displayed performance with a 70% accuracy and a 71% recall for the Alzheimer's/Vascular class,

suggesting better overall predictive strength compared to logistic regression. Nevertheless, it still struggles to classify Dementia with Lewy Bodies accurately, mirroring the challenges seen with the logistic regression model. The stable precision for Alzheimer's/Vascular across different sampling methods suggests that the challenge in model's class differentiation transcends base rate influences. The persistence of high false negative rates for the Dementia with Lewy Bodies class resulted in an overall low F1 score, underscoring the need for further model refinement.

Site-specific analysis

To evaluate the influence of site-specific factors, we first tested our model trained on multi-site data on individual sites and compared these results with models trained on single-site data. Despite site variability, the results underscore the necessity of multi-site data to enhance model generalization. Models trained on data from all sites combined consistently outperformed those trained on single-site data when tested

Table 1. Summary of the demographics characteristics of the final study cohort after combining subtypes (the proportion is too small to quantitatively).

Demographics	Subtypes of dementia	
	Alzheimer/Vascular	Lewy Bodies
N = Number of patients per subtype	33 318	1657
Average age (range: 40-88)	78.19	77.11
Age 89 or older	8516 (25.5)	214 (12.9)
Sex		
Male	12 550 (37.67)	993 (59.92)
Female	20 768 (63.33)	664 (40.08)
Other/Missing/Unknown	0	0
Race ethnicity		
American Indian or Alaska Native Non-Hispanic	<340 (1.01)	<20(-)
Asian Non-Hispanic	1216 (3.65)	<30(-)
Black or African American Non-Hispanic	6262 (18.79)	155 (9.35)
Hispanic or Latino Any Race	2018 (6.06)	80 (4.83)
Native Hawaiian or Other Pacific Islander Non-Hispanic	51 (0.15)	0
Other Non-Hispanic	49 (0.15)	0
Unknown	1108(3.33)	53(3.2)
White Non-Hispanic	22 277 (66.86)	1338 (80.75)

Table 2. Detailed statistical measures for all classifier models: (a) Logistic Regression and (b) XGBOOST.

Model		Accuracy	Precision	Recall	F1
Logistic regression	Alzheimer’s/Vascular	0.67 (95% CI, 0.6638-0.6859)	0.97 (95% CI, 0.9645-0.9742)	0.67 (95% CI, 0.6656-0.6752)	0.80 (95% CI, 0.7912-0.8076)
	Lewy Bodies		0.08 (95% CI, 0.0690-0.0913)	0.56 (95% CI, 0.5080-0.6172)	0.14 (95% CI, 0.1227-0.1585)
XGBoost	Alzheimer’s/Vascular	0.70 (95% CI, 0.6901-0.7105)	0.97 (95% CI, 0.9644-0.9742)	0.71 (95% CI, 0.6974-0.7183)	0.82 (95% CI, 0.8108-0.8257)
	Lewy Bodies		0.08 (95% CI, 0.6638-0.6859)	0.55 (95% CI, 0.4910-0.6029)	0.14 (95% CI, 0.1270-0.1677)

The numbers in brackets represent the 95% confidence interval (CI) obtained from bootstrapping.

individually. The detailed comparison can be found in the [Supplementary Material](#).

To determine if subtype overlap could be attributed to data anomalies from specific sites, we added site ID to our analysis. Incorporating site ID into our binary classification models highlighted the role of site-specific factors in diagnosing dementia subtypes, especially noted in sites with high prevalence of vascular dementia. To further understand the influence of site-specific factors on our models, we conducted separate analyses using data from each site. Conducting site-wise analyses allowed us to gauge each site’s unique impact on model performance, revealing variations shown in [Table 3](#). Models based on data from individual sites often outperformed those using the aggregated dataset, indicating the potential benefits of a site-specific approach. A hierarchical model was then applied to integrate site-specific intercepts and covariate effects, offering a more nuanced understanding of site influences on predictions. Yet, model accuracy was largely unaffected. The Shapley plots also confirmed this, indicating a consistent pattern of comorbidities influencing model predictions across both experiments.

The Partial Pooling Hierarchical Model’s diagnostics, depicted in [Figure 4](#), demonstrated successful MCMC convergence. “Mean_beta” represents the average effect size across all sites, providing a baseline for the site-specific intercepts, while “sigma_beta” captures the variability or

standard deviation of these intercepts, indicating the extent of differences among site-specific intercepts. Both hyperparameters exhibited stable convergence, evidenced by consistent central tendencies. Density and trace plots for “beta_sites” revealed site-specific intercept variability, affirming the model’s ability to identify unique intercepts for different sites. [Figure 5](#) complements this analysis by providing a Forest Plot of site-specific effects, further delineating how covariates such as sex and comorbidities interact with binary outcomes. The 95% Highest Posterior Density intervals and point estimates for each parameter are meticulously detailed, showcasing the heterogeneity across sites.

Discussion

We developed a multi-stage approach to classify dementia subtypes based on comorbidities using data from multiple healthcare institutions, highlighting the limitations of single-site studies, which are often susceptible to institution-specific biases. By integrating data from diverse sites, we explored the association of common comorbidities with specific dementia subtypes through steps like subtypes pruning, combination, group balancing, and machine learning classification. Unlike prior research primarily targeting Alzheimer’s disease risk prediction or subtype clustering, our approach, leveraging multiple “off-the-shelf” machine learning algorithms and sampling methods on multi-site EHR data, including patient

Table 3. Detailed statistical measures for all site-specific models.

Site	Accuracy	ROC	Precision	Recall	F1	
1	Alzheimer's/Vascular Lewy Bodies	0.71 (95% CI, 0.6554-0.7542)	0.40	0.97 (95% CI, 0.9500-0.9918) 0.02 (95% CI, 0.0000-0.0515)	0.72 (95% CI, 0.6696-0.7652) 0.22 (95% CI, 0.0000-0.5008)	0.83 (95% CI, 0.7917-0.8581) 0.04 (95% CI, 0.0000-0.0901)
2	Alzheimer's/Vascular Lewy Bodies	0.66 (95% CI, 0.6147-0.6991)	0.68	0.96 (95% CI, 0.9393-0.9832) 0.10 (95% CI, 0.0610-0.1528)	0.66 (95% CI, 0.6166-0.7041) 0.61 (95% CI, 0.4443-0.7917)	0.78 (95% CI, 0.7511-0.8139) 0.18 (95% CI, 0.1087-0.2513)
3	Alzheimer's/Vascular Lewy Bodies	0.69 (95% CI, 0.6461-0.7263)	0.76	0.97 (95% CI, 0.9536-0.9878) 0.09 (95% CI, 0.0506-0.1411)	0.69 (95% CI, 0.6487-0.7295) 0.62 (95% CI, 0.4210-0.8126)	0.81 (95% CI, 0.7765-0.8352) 0.16 (95% CI, 0.0919-0.2353)
4	Alzheimer's/Vascular Lewy Bodies	0.68 (95% CI, 0.6461-0.7145)	0.66	0.96 (95% CI, 0.9419-0.9761) 0.09 (95% CI, 0.0546-0.1261)	0.69 (95% CI, 0.6563-0.7224) 0.51 (95% CI, 0.3589-0.6667)	0.80 (95% CI, 0.7780-0.8270) 0.15 (95% CI, 0.0952-0.2090)
5	Alzheimer's/Vascular Lewy Bodies	0.71 (95% CI, 0.6453-0.7564)	0.53	0.96 (95% CI, 0.9341-0.9880) 0.05 (95% CI, 0.0000-0.0984)	0.72 (95% CI, 0.6607-0.7768) 0.33 (95% CI, 0.0000-0.6667)	0.82 (95% CI, 0.7810-0.8606) 0.08 (95% CI, 0.0000-0.1644)
6	Alzheimer's/Vascular Lewy Bodies	0.62 (95% CI, 0.5976-0.6508)	0.66	0.96 (95% CI, 0.9490-0.9751) 0.10 (95% CI, 0.0735-0.1290)	0.64 (95% CI, 0.5971-0.6523) 0.63 (95% CI, 0.5211-0.7361)	0.76 (95% CI, 0.7353-0.7780) 0.17 (95% CI, 0.1299-0.2175)
7	Alzheimer's/Vascular Lewy Bodies	0.60 (95% CI, 0.5764-0.6189)	0.71	0.98 (95% CI, 0.9675-0.9856) 0.08 (95% CI, 0.0576-0.0942)	0.59 (95% CI, 0.5703-0.6147) 0.70 (95% CI, 0.6067-0.7959)	0.74 (95% CI, 0.7197-0.7550) 0.14 (95% CI, 0.1058-0.1669)
8	Alzheimer's/Vascular Lewy Bodies	0.67 (95% CI, 0.6395-0.6948)	0.61	0.97 (95% CI, 0.9614-0.9854) 0.06 (95% CI, 0.0344-0.0846)	0.67 (95% CI, 0.6434-0.7011) 0.53 (95% CI, 0.3704-0.6774)	0.80 (95% CI, 0.7743-0.8161) 0.10 (95% CI, 0.0630-0.1483)
9	Alzheimer's/Vascular Lewy Bodies	0.67 (95% CI, 0.6205-0.7098)	0.65	0.98 (95% CI, 0.9623-0.9934) 0.07 (95% CI, 0.0274-0.1111)	0.67 (95% CI, 0.6227-0.7133) 0.62 (95% CI, 0.3845-0.8571)	0.80 (95% CI, 0.7596-0.8265) 0.12 (95% CI, 0.0519-0.1928)
<i>Average</i>		0.67	0.97	0.67	0.79	
<i>Logistic Regression</i>		0.65 (95% CI, 0.6409-0.6620)	0.66	0.97 (95% CI, 0.9642-0.9740) 0.08 (95% CI, 0.0655-0.0873)	0.65 (95% CI, 0.6428-0.6654) 0.59 (95% CI, 0.5333-0.6448)	0.78 (95% CI, 0.7755-0.7922) 0.14 (95% CI, 0.1170-0.1530)

The numbers in brackets represent the 95% confidence interval (CI) obtained from bootstrapping. The ROC (Receiver Operating Characteristic) value quantifies the overall performance of a binary classifier, with higher values indicating better discriminative ability.

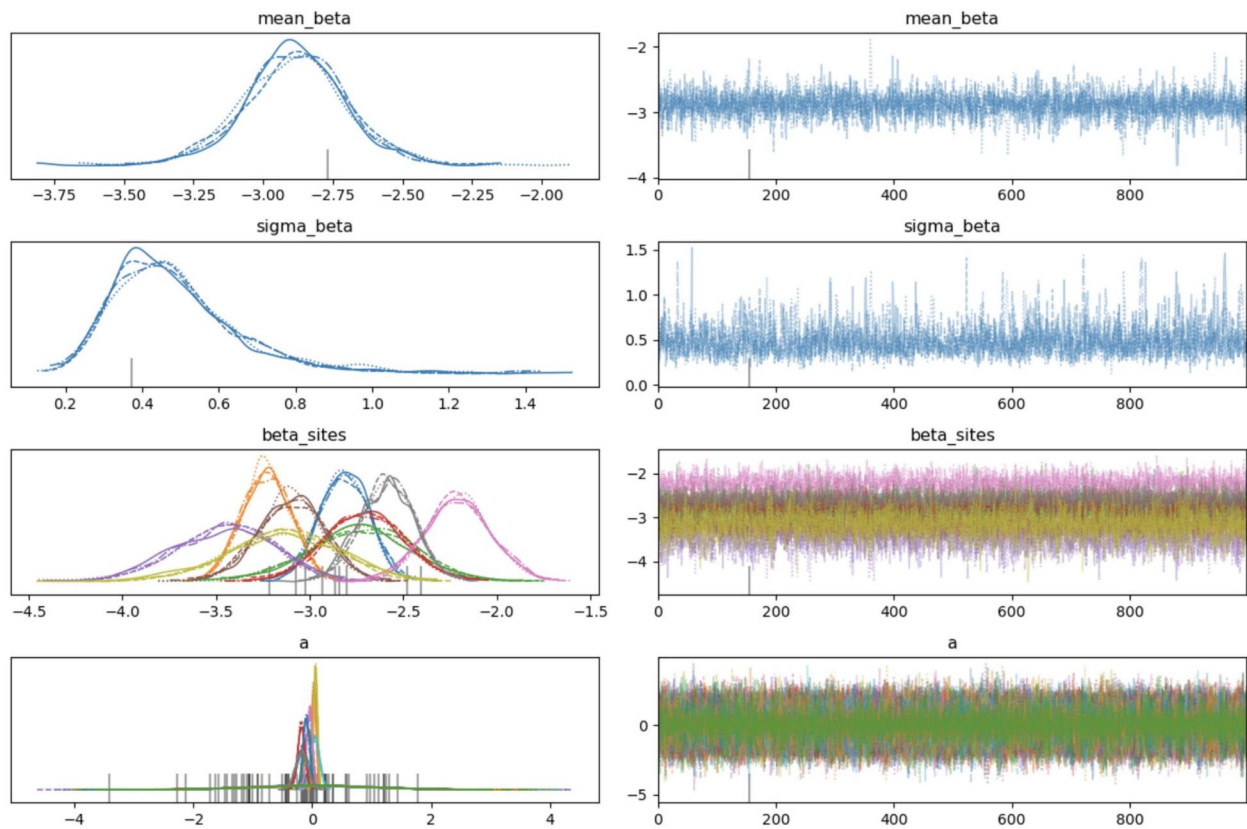


Figure 4. Diagnostic plots for Bayesian hierarchical model parameters fitted via MCMC. Top-left: Density plot representing the average effect across sites. Top-right: Trace plot illustrating sampling convergence. Middle-left: Density plot for site-specific effects, with each color corresponding to a different site. Middle-right: Trace plot displaying sampling paths per site. Bottom-left: Histogram showing the standard deviation of site effects. Bottom-right: Trace plot demonstrating chain convergence and mixing.

gender and comorbidities, specifically targets classification among a broader range of patients living with dementia.

Our multi-institutional approach significantly advances dementia research. It leverages the N3C enclave to address the disparate nature of the data in the US healthcare system, enabling comprehensive multi-site data analyses. This methodology unveils biases that single-site studies may overlook, enhancing our understanding of the association between comorbidities and dementia subtypes. By utilizing harmonized data, our approach ensures high data quality and supports both broad and region-specific insights, making our findings more generalizable and relevant.

Utilizing comprehensive multi-site data allows for a more representative analysis of the general population and highlights site-specific variability, emphasizing the need for caution in multi-provider research. Our binary classification, particularly the binary classification between Alzheimer's and Vascular dementia, and the hierarchical model, highlights site variability, pointing to the potential impact of unique clinical practices or patient demographics on dementia diagnoses across different sites. Moreover, our reliance on EHR diagnoses assumes uniform EHR implementation across contributing sites, which could introduce inaccuracies. Despite the challenges of site variability and potential inaccuracies inherent in EHR, the imperative for model replication and validation remains, bolstered by the expansive and diverse dataset of the N3C enclave. To mitigate site-specific differences, hierarchical models prove beneficial, allowing for sophisticated differentiation that considers site-specific practices and patient populations. However, these models come with

increased complexity and computational demands and require a deeper understanding of the underlying disease processes to accurately structure the relationships among subtypes. To address site variability and enhance the generalizability of models across varied healthcare settings, current approaches must allow models to learn generalized representations that are not overly specific to any single site's data distribution.

Our study confronts challenges in EHR datasets, including variability and incompleteness, necessitating the exclusion of certain dementia subtypes due to small sample sizes and information insufficiency challenges. To mitigate class imbalance, we implemented a 65% SMOTE, 35% RUS, and Tomek Links strategy, demonstrating the importance of customized sampling methods in healthcare data analysis. This approach, with a tailored 35:65 undersampling to oversampling ratio, emphasizes the importance of empirical testing in optimizing sampling balances. Future work will aim to incorporate a broader range of variables and a more representative sample to enhance model robustness and refine models for classifying dementia subtypes. Furthermore, our exploration of multi-site data revealed site differences that may account for the occurrences of certain dementia subtypes. Notably, Parkinson's was reported by only a few sites, and one site showed a higher prevalence of vascular dementia, underscoring the need to consider site-specific factors for improved generalizability. Despite challenges like class imbalance and site variability, employing machine learning and hierarchical modeling, combined with strategic sampling and multi-site data exploration, offers promising avenues for advancing dementia classification models.

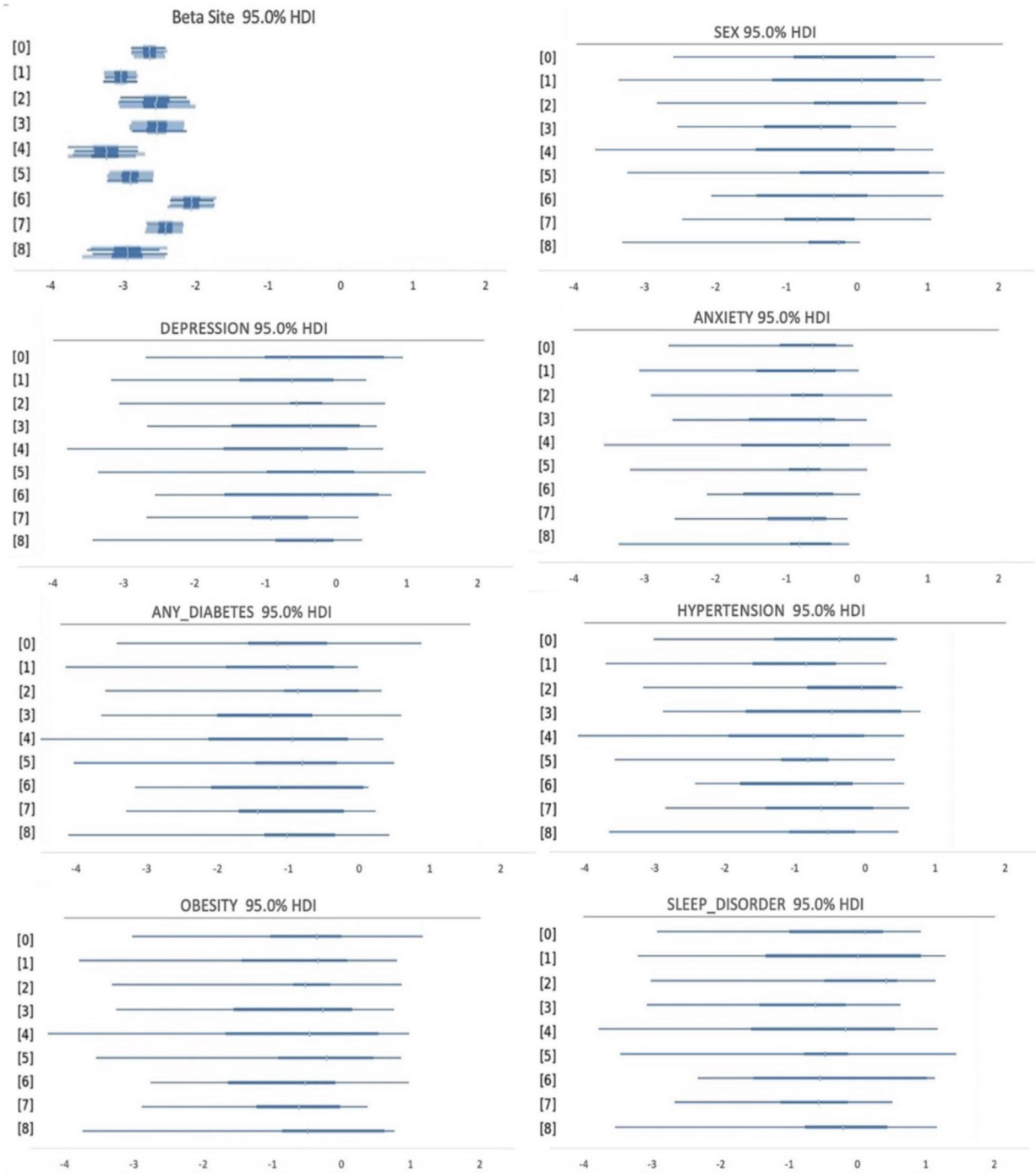


Figure 5. Forest Plot of the hierarchical model: summary of site-specific effects and interactions of covariates such as sex and comorbidities on binary outcomes, with each parameter’s 95% highest posterior density interval and point estimate shown.

Conclusion

Our study offers a foundational methodology for understanding dementia by employing a multi-site dataset to examine comorbidity patterns across various dementia subtypes. Integrating a 65% SMOTE, 35% RUS, and Tomek Links strategy combats class imbalance, highlighting the importance of tailored sampling in healthcare analytics. The hierarchical model reveals more information about the relationships between dementia subtypes and their comorbidities across multiple sites, enhancing our comprehension of

these complex interactions. Extending this methodology to other disease subtype analyses—such as cardiovascular diseases and chronic inflammatory disorders—promises to unearth valuable insights into their inherent heterogeneity.

Acknowledgments

The analyses described in this journal were conducted with data or tools accessed through the NCATS N3C Data Enclave <https://covid.cd2h.org> and N3C Attribution &

Publication Policy (<https://zenodo.org/records/10698093>) supported by NCATS Contract No. 75N95023D00001, Axle Informatics Subcontract: NCATS-P00438-B, and also funded by the NIH award UL1TR003015 for the integrated Translational Health Research Institute of Virginia. This research was possible because of the patients whose information is included within the data and the organizations [covid.cd2h.org/duas] and scientists who have contributed to the ongoing development of this community resource.

Author contributions

Suchetha Sharma and Jiebei Liu contributed equally to this paper. Both authors were involved in the conception and design of the study, designed and performed the data analysis, and implemented the algorithms. They led in writing and revising the manuscript.

Amy Caroline Abramowitz, Carol Manning, Carol Reynolds Geary, Karen C. Johnston, and John Darrell Van Horn were involved in the study conception, design of the dementia subtypes and concept sets, and provided other clinical insights relevant to the study. They all helped in the writing/revising the manuscript.

Johanna Loomba, Emily Pfaff, Andrea Zhou, Alfred J. Anzalone, and Don Brown were involved in the study conception that contributed to the data preprocessing and implementation of machine learning algorithms. They were also involved in drafting/revising the manuscript.

All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work, ensuring its accuracy and integrity.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This research was funded by the NIH National Center for Advancing Translational Sciences (NCATS) through award UL1TR003015 (awarded to the integrated Translational Health Research Institute of Virginia at the University of Virginia) and award UM1TR004406 (awarded to the University of North Carolina). The project was also supported by the National Institute of General Medical Sciences (NIGMS) award U54GM115458 (awarded to the University of Nebraska). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH or the NIGMS.

Conflicts of interest

None declared.

Data availability

As per the N3C Clinical Tenant Pilot Data Usage Agreement (DUA), the source data cannot be shared publicly. It can only be accessed and used by individuals covered by the pilot DUA. This pilot is now complete.

References

- Glynn EF, Hoffman MA. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. *JAMIA Open*. 2019;2(4):554-561. <https://doi.org/10.1093/jamiaopen/ooz035>
- Zozus MN, Richesson RL, Walden A, et al. Research reproducibility in longitudinal multi-center studies using data from electronic health records. *AMIA Jt Summits Transl Sci Proc*. 2016;2016:279-285. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001777/>
- Burrows EK, Razzaghi H, Utidjian L, et al. Standardizing clinical diagnoses: evaluating alternate terminology selection. *AMIA Jt Summits Transl Sci Proc*. 2020;2020:71-79. <https://pubmed.ncbi.nlm.nih.gov/32477625/>
- Rahman N, Wang DD, Ng SHX, et al. Processing of electronic medical records for health services research in an academic medical center: methods and validation. *JMIR Med Inform*. 2018;6(4):e10933. <https://doi.org/10.2196/10933>
- Malhotra K, Hobson TC, Valkova S, et al. Sequential pattern mining of electronic healthcare reimbursement claims: experiences and challenges in uncovering how patients are treated by physicians. *IEEE Int Conf Big Data (Big Data)*. IEEE; 2015:2670-2679. <https://doi.org/10.1109/BigData.2015.7364067>
- Chertkow H, Feldman HH, Jacova C, et al. Definitions of dementia and predementia states in Alzheimer's disease and vascular cognitive impairment: consensus from the Canadian conference on diagnosis of dementia. *Alzheimers Res Ther*. 2013;5(Suppl 1):S2. <https://doi.org/10.1186/alzrt198>
- Duong S, Patel T, Chang F. Dementia. *Can Pharm J*. 2017;150(2):118-129. <https://doi.org/10.1177/1715163517690745>
- Beach TG, Monsell SE, Phillips LE, et al. Accuracy of the clinical diagnosis of Alzheimer disease at national institute on aging Alzheimer disease centers, 2005–2010. *J Neuropathol Exp Neurol*. 2012;71(4):266-273. <https://doi.org/10.1097/nen.0b013e31824b211b>
- Alzheimer's Association. 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dement*. 2022;18(4):700-789. <https://doi.org/10.1002/alz.12638>
- Sanderson M, Wang J, Davis DR, et al. Co-morbidity associated with dementia. *Am J Alzheimers Dis Other Dement*. 2002;17(2):73-78. <https://doi.org/10.1177/153331750201700210>
- Formiga F, Fort I, Robles MJ, et al. Comorbidity and clinical features in elderly patients with dementia: differences according to dementia severity. *J Nutr Health Aging*. 2009;13(5):423-427. <https://doi.org/10.1007/s12603-009-0078-x>
- Khondoker M, Macgregor A, Bachmann MO, et al. Multimorbidity pattern and risk of dementia in later life: an 11-year follow-up study using a large community cohort and linked electronic health records. *J Epidemiol Community Health*. 2023;77(5):285-292.
- Teixeira PL, Wei WQ, Cronin RM, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J Am Med Inform Assoc*. 2016;24(1):162-171. <https://doi.org/10.1093/jamia/ocw071>
- Johnson AEW, Pollard TJ, Shen LU, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. <https://doi.org/10.1038/sdata.2016.35>
- Waitman LR, Aaronson LS, Nadkarni PM, et al. The greater plains collaborative: a PCORnet clinical research data network. *J Am Med Inform Assoc*. 2014;21(4):637-641. <https://doi.org/10.1136/amiajnl-2014-002756>
- DeShazo JP, Hoffman MA. A comparison of a multistate inpatient EHR database to the HCUP nationwide inpatient sample. *BMC Health Serv Res*. 2015;15(1):384. <https://doi.org/10.1186/s12913-015-1025-7>
- Raman SR, Kreda DA, Huang SC, et al. The greater plains collaborative: a national research network to improve health outcomes for

- patients with diverse healthcare needs. *J Am Med Inform Assoc.* 2022;29(4):660-671. <https://doi.org/10.1093/jamia/ocac028>
18. Fiest KM, Jetté N, Roberts JL, et al. The prevalence and incidence of dementia: a systematic review and meta-analysis. *Can J Neurol Sci.* 2016;43 Suppl 1(S1):S3-S50. <https://doi.org/10.1017/cjn.2016.18>
 19. Pengo M, Alberici A, Libri I, et al. Sex influences clinical phenotype in frontotemporal dementia. *Neurol Sci.* 2022;43(9):5281-5287. <https://doi.org/10.1007/s10072-022-06185-7>
 20. Xu J, Wang F, Xu Z, et al. Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn Health Syst.* 2020;4(4):e10246. <https://doi.org/10.1002/lrh2>
 21. Haendel MA, Chute CG, Bennett TD, N3C Consortium, et al. The National COVID cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc.* 2020;28(3):427-443. <https://doi.org/10.1093/jamia/ocaa19>
 22. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc.* 2014;21(2):221-230. <https://doi.org/10.1136/amiainjnl-2013-001935>
 23. Barnes DE, Zhou J, Walker RL, et al. Development and validation of eRADAR: a tool using EHR data to detect unrecognized dementia. *J Am Geriatr Soc.* 2019;68(1):103-111. <https://doi.org/10.1111/jgs.16182>
 24. Wei WQ, Teixeira PL, Mo H, et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc.* 2016;23(e1):e20-e27. <https://doi.org/10.1093/jamia/ocv130>
 25. Harding BN, Floyd JS, Scherrer JF, et al. Methods to identify dementia in the electronic health record: comparing cognitive test scores with dementia algorithms. *Healthc (Amst).* 2020;8(2):100430. <https://doi.org/10.1016/j.hjdsi.2020.100430>
 26. Haut ER, Pronovost PJ, Schneider EB. Limitations of administrative databases. *JAMA.* 2012;307(24):2589; author reply 2589-2590. <https://doi.org/10.1001/jama.2012.6626>
 27. Walling AM, Pevnick JM, Bennett AV, et al. Dementia and electronic health record phenotypes: a scoping review of available phenotypes and opportunities for future research. *J Am Med Inform Assoc.* 2023;30(7):1333-1348. <https://doi.org/10.1093/jamia/ocad086>
 28. National COVID Cohort Collaborative. Alzheimer's Phenotype. GitHub. Published May 30, 2023. Accessed March 24, 2024. <https://github.com/National-COVID-Cohort-Collaborative/tenant-pilot/wiki/Alzheimer>
 29. OMOP Common Data Model—OHDSI. Ohdsi.OM. 2019. Accessed July 31, 2024. <https://www.ohdsi.org/data-standardization/the-common-data-model/>
 30. Boonyasai RT, Song L, Bandeen-Roche K, et al. Measures of quality and outcomes that matter to patients and caregivers of older adults with multiple chronic conditions. *J Am Med Inform Assoc.* 2022;29(4):609-621. <https://doi.org/10.1093/jamia/ocab278>
 31. Yang S, Varghese P, Stephenson E, et al. Machine learning approaches for electronic health records phenotyping: a methodical review. *J Am Med Inform Assoc.* 2022;30(2):367-381. <https://doi.org/10.1093/jamia/ocac216>
 32. Almowil ZA, Zhou SM, Brophy S. Concept libraries for automatic electronic health record based phenotyping: a review. *Int J Popul Data Sci.* 2021;6(1):1362. <https://doi.org/10.23889/ijpds.v5i1.1362>
 33. Boccardi M, Laakso MP, Bresciani L, et al. The MRI pattern of frontal and temporal brain atrophy in fronto-temporal dementia. *Neurobiol Aging.* 2003;24(1):95-103. [https://doi.org/10.1016/s0197-4580\(02\)00045-3](https://doi.org/10.1016/s0197-4580(02)00045-3)
 34. Smith MR, Martinez T, Giraud-Carrier C. An instance level analysis of data complexity. *Mach Learn.* 2013;95(2):225-256. <https://doi.org/10.1007/s10994-013-5422-z>
 35. Wongvorachan T, He S, Bulut O. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information.* 2023;14(1):54. <https://doi.org/10.3390/info14010054>
 36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res.* 2002;16(16):321-357. <https://doi.org/10.1613/jair.953>
 37. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Lect Notes Comput Sci.* 2005;3644:878-887. https://doi.org/10.1007/11538059_91
 38. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. *IJKESDP.* 2011;3(1):4. <https://doi.org/10.1504/ijkesdp.2011.039875>
 39. He H, Bai Y, Garcia EA, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks.* Hong Kong: IEEE; 2008:1322-1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
 40. Batista G, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor Newsl.* 2004;6(1):20-29. <https://doi.org/10.1145/1007730.1007735>
 41. Ramentol E, Caballero Y, Bello R, et al. SMOTE-RSB: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowl Inf Syst.* 2011;33(2):245-265. <https://doi.org/10.1007/s10115-011-0465-6>
 42. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res.* 2016;18(1):559-563. <https://dl.acm.org/doi/10.5555/3122009.3122026>
 43. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Methodol.* 1958;20(2):215-232. <https://www.jstor.org/stable/2983890>
 44. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'16.* New York: Association for Computing Machinery; 2016:785-794. <https://doi.org/10.1145/2939672.2939785>
 45. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *NeurIPS.* 2017. Accessed August 1, 2024. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
 46. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. *NeurIPS.* Published 2012. Accessed July 23, 2023. <https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html>
 47. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12(85):2825-2830. <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>
 48. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *Adv Inform Retrieval.* 2005;3408:345-359. https://doi.org/10.1007/978-3-540-31865-1_25
 49. Allenby GM, Rossi PE, McCulloch RE. Hierarchical bayes models: a practitioners guide. Social Science Research Network; 2005. <https://doi.org/10.2139/ssrn.655541>
 50. Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. *J Am Stat Assoc.* 1985;80(391):513-524. <https://doi.org/10.1080/01621459.1985.10478148>
 51. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Verlag: Cambridge University Press; 2018.
 52. Abril-Pla O, Andreani V, Carroll C, et al. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Comput Sci.* 2023;9:e1516. <https://doi.org/10.7717/peerj-cs.1516>

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
JAMA Open, 2024, 7, 1–12
<https://doi.org/10.1093/jamiaopen/ooae076>
Research and Applications