Uncovering Important Diagnostic Features for Alzheimer's, Parkinson's and Other Dementias Using Interpretable Association Mining Methods

Kazi Noshin[1*], Mary Regina Boland[3*], Bojian Hou[4], Victoria Lu[1],
Carol Manning[2], Li Shen[4†], Aidong Zhang[1†]

[1]*Department of Computer Science,* [2]*Department of Neurology*
University of Virginia, VA 22903, USA
[3]*Data Science Program, Department of Mathematics,*
Saint Vincent College, Latrobe, PA 15650, USA
[4]*Department of Biostatistics, Epidemiology and Informatics,*
University of Pennsylvania, Philadelphia, PA 19104, USA
E-mail: epw9kz@virginia.edu, mary.boland@stvincent.edu
Bojian.Hou@Pennmedicine.upenn.edu, gbp7sb@virginia.edu
CM4R@uvahealth.org, li.shen@pennmedicine.upenn.edu, aidong@virginia.edu

Alzheimer's Disease and Related Dementias (ADRD) afflict almost 7 million people in the USA alone. The majority of research in ADRD is conducted using post-mortem samples of brain tissue or carefully recruited clinical trial patients. While these resources are excellent, they suffer from lack of sex/gender, and racial/ethnic inclusiveness. Electronic Health Records (EHR) data has the potential to bridge this gap by including real-world ADRD patients treated during routine clinical care. In this study, we utilize EHR data from a cohort of 70,420 ADRD patients diagnosed and treated at Penn Medicine. Our goal is to uncover important risk features leading to three types of Neuro-Degenerative Disorders (NDD), including Alzheimer's Disease (AD), Parkinson's Disease (PD) and Other Dementias (OD). We employ a variety of Machine Learning (ML) Methods, including uni-variate and multivariate ML approaches and compare accuracies across the ML methods. We also investigate the types of features identified by each method, the overlapping features and the unique features to highlight important advantages and disadvantages of each approach specific for certain NDD types. Our study is important for those interested in studying ADRD and NDD in EHRs as it highlights the strengths and limitations of popular approaches employed in the ML community. We found that the uni-variate approach was able to uncover features that were important and rare for specific types of NDD (AD, PD, OD), which is important from a clinical perspective. Features that were found across all methods represent features that are the most robust.

Keywords: Electronic Health Records; Machine Learning; Alzheimers Disease and Related Dementias; Data Mining

*Equal-Contribution First-Authors.
†Equal-Contribution Senior-Authors.

1. Introduction

1.1 Alzheimer's Disease and Related Dementias

ADRD afflicts an estimated 6.9 million people in the United States of America (USA), using current July 2024 statistics.[1] ADRD and dementia collectively kill more patients per year than breast and prostate cancers combined.[1] However, despite its frequency of incidence, not much is known about ADRD patients in community-based settings. This is because the majority of Alzheimer's Disease (AD) research focuses on post-mortem (after the patient has died) samples or patients recruited through expensive clinical trials (that often lack racial/ethnic diversity). In addition, there remains a paucity of research among diverse populations, including investigating sex-disparities[2] and racial disparities[3] in outcomes. Additionally, many state-of-the-art studies on ADRD have limited generalizability because of the almost exclusive use of trials that lack race/ethnicity/socioeconomic inclusiveness,[4] leading to a diversity dearth.[5]

1.2 Electronic Health Records (EHRs)

The recent development and implementation of EHRs now provide a tremendous opportunity to evaluate ADRD patients from community-based settings that includes in-patient and outpatient medical records data obtained through routine clinical care. EHR data contain information on millions of patients from both in-patient and out-patient settings. They often contain more representative patient populations (in terms of race, ethnicity, and socioeconomic inclusiveness) than clinical trials due to their community-based settings. Several studies have used EHR data for AD research. Xu et al.[6] developed a data-driven method to uncover four subphenotypes of AD from EHRs. Their subphenotypes were correlated with common comorbidities of ADRD, including mental health diseases and cardiovascular disease.[6] None of these prior studies (as far as we are able to glean from the reported papers) have incorporated socioeconomic or racial/ethnic disparities into their algorithm development. This is important as not properly capturing these features can lead to biased research results.[7,8]

1.3 Uni-variate Association Mining

While Xu et al.[6] utilized unsupervised Machine Learning (ML) methods to learn types of ADRD (a form of neurodegenerative disorder (NDD)) from the data itself, another common method for uncovering important features or characteristics of a dataset is to utilize association mining. Association mining is used extensively in EHR research through a process called Phenotype-Wide Association Studies (PheWAS) first introduced in 2010 by Denny et al.[9] In their study they held the genetic variant constant while looping over a wide range of clinical EHR-derived phenotypes.[9] This process was then employed by BioBanks throughout the USA and abroad, but also applied to EHR datasets not linked to BioBank data.[9–23] Others used EHR data without genetic information to perform association mining or PheWAS style studies.[18] Boland et al. employed a similar algorithmic approach when exploring the relationship between birth seasonality and later risk of disease through a method first published in 2015[19] and later replicated in several studies.[20–23] The essence of association mining is to test for an association (using some statistical method, e.g., chi-square test, fisher's exact test, or regression) between each phenotype (typically represented as columns in a matrix) and the outcome of interest. In this study, our outcomes are three different NDD types. Therefore, the outcome is set (in this work either AD, PD or OD) *a priori* and then each phenotype

(i.e., covariate/feature/column) is tested for association with that outcome of interest. If one wants to investigate more than one outcome (in this case our different NDD types) then one simply repeats the entire process over again with each outcome. We construct our algorithm such that outcome Y is always the same (in this case a binary indicator variable for whether or not the patient has/had a particular NDD type, e.g., PD). We then have our intercept term (β_0) and our term related to that particular feature that is being tested (or iterated over) is X and the coefficient term related to the feature is represented as β_x. We will loop over all potential features and therefore with each iteration the actual feature in X and the corresponding coefficient β_x will change. A sample regression equation for a binary outcome of interest (NDD type: Parkinson's Disease) is as follows:

$$Y_{(NDD\ type:\ Parkinson'sDisease)} = \beta_0 + \beta_x * X,$$

with β_x indicating the term for each phenotype (or feature) that will be iterated over. Therefore, in our example the first feature would be some clinical or demographic feature, followed by the second feature until all features have been iterated over. Typically, there are a large number of associations explored (often into the thousands) requiring multiple hypothesis correction methods to adjust for multiple comparisons.

1.4 Multivariate Association Mining: SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) is a method to explain individual predictions based on Shapley values from cooperative game theory.[24] It assigns each feature an importance value for a particular prediction,[25] aiming to fairly distribute the 'payout' (prediction) among features. SHAP provides both local explanations for individual predictions and global interpretation methods, linking optimal credit allocation with local explanations using Shapley values.[26]

In the context of EHR data, SHAP can be a powerful tool for interpreting the predictions made by the models. This study uses PD, AD, and OD as separate outcomes. Each patient in the dataset can be considered as an instance for which a prediction is made. The features are the 'players' in the game. The 'payout' is the prediction of whether a patient has PD, AD, or OD. For example, if a model predicts a certain patient has a high risk of developing PD, SHAP can help us understand how each feature contributes to this prediction. This can provide valuable insights into which factors are most influential in predicting PD, AD, or OD.

The Shapley value is the average of all the marginal contributions to all possible coalitions.[24] For a set N of n features, the Shapley value $\phi_i(v)$ of feature i is:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{1}$$

where S is a subset of features not including feature i, v is a value function that represents the model's prediction for a subset of features, $v(S)$ is the prediction for subset S, and $|S|$ is the number of features in S.

The SHAP explanation method computes Shapley values. Let g be is the explanation model, $z' \in \{0,1\}^M$ the coalition vector, M the maximum coalition size, and $\phi_j \in \mathbb{R}$ the feature

attribution, i.e., the Shapley values for feature j. SHAP is defined mathematically as follows:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2}$$

In Equation 2, an entry of 1 in the coalition vector indicates that the corresponding feature value is 'present', whereas an entry of 0 signifies that it is 'absent'. Within the framework of SHAP, the Shapley values help us understand each feature's contribution to the prediction.

2. Dataset

2.1 Dataset Description

We obtained de-identified EHR data from Penn Medicine for patients with ADRD using a set of diagnosis codes. The age range of our medical records indicate that the majority of the EHR data was collected between 2002 and 2022 with some diagnosis dates occurring earlier (all the way back to the 1920s indicating manually entered diagnosis information that was pertinent for specific patients). The internal Clinical Data Warehouse at Penn Medicine converted the International Classification of Diseases (ICD) version 9 (ICD-9) codes to version 10 (ICD-10). We have cross-mapped our list of ADRD diagnostic codes using existing resources[27] to provide researchers with our full list of ICD-9 and ICD-10 diagnosis codes for ADRD identification.[28] The EHR data comes in the Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (CDM) format with relevant data broken down into several files corresponding to tables in a SQL database. The dataset contains information on patients' encounters, diagnoses, medications, procedures, vitals, laboratory findings, chemotherapy, and laboratory values. This study was approved by the University of Pennsylvania's Institutional Review Board (IRB) with approval id: 851588. We mapped our entire dataset consisting of 70,420 ADRD patients to their corresponding PheCodes. This allowed us to identify 14,911 patients with AD diagnoses specifically (PheCode:290.11), 16,216 patients with PD diagnoses specifically (PheCode:332) and 14,911 patients with 'Dementias' (PheCode:290.10) called in this paper Other Dementias (OD), which is an unspecified generic dementia category. Demographics are provided in Table 1 and visualized in Figure 2. The Venn diagram in Figure 1 represents the overlap of patients diagnosed with PD, AD, and OD.

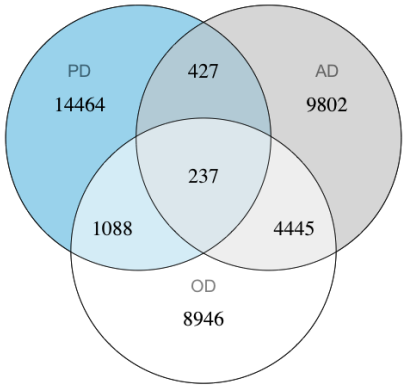

Fig. 1. Venn Diagram of Patients Diagnosed with PD, AD and OD.

Demographic factors differs among the three NDD subtypes. Figure 2 shows the Racial and Sex distributions across the NDD types. The bars represent percentages of different racial groups for four categories: AD, PD, OD, and Overall. White individuals have a higher percentage of PD, whereas Black or African Americans have a higher percentage of OD diagnoses (Figure 2). Figure 2 shows that females have a higher percentage of NDD types that include AD and OD compared to males. On the other hand, males have a higher percentage of PD

compared to females. Overall, across NDD types, there was a higher proportion of females with ADRD diagnoses than males (54.43% vs. 45.56%).

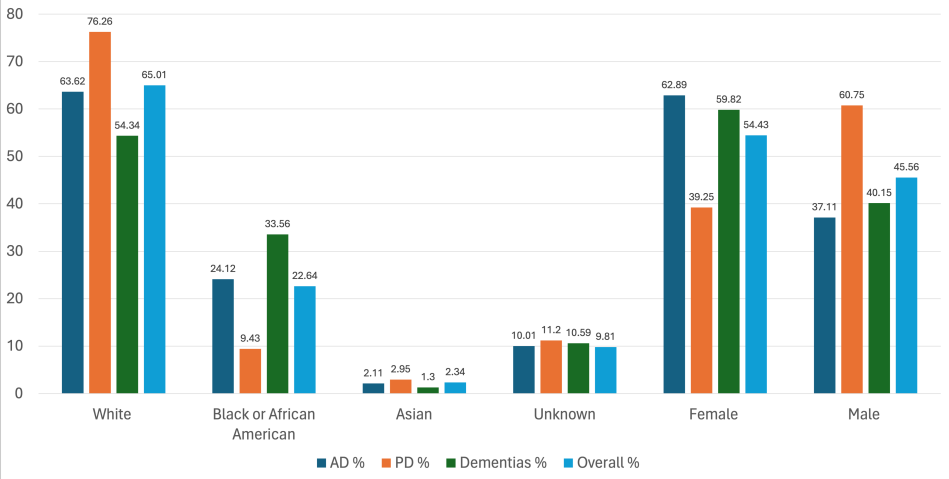

Fig. 2. Racial and Sex Distribution by NDD type.

Table 1. Demographics of ADRD Patients by NDD type.

Attribute	Value	AD(%) (N = 14911)	PD(%) (N = 16216)	OD(%) (N = 14911)	Overall(%) (N = 70420)
Race	White	63.62	76.26	54.34	65.01
	BAA	24.12	9.43	33.56	22.64
	Asian	2.11	2.95	1.3	2.34
	NHOPI	0.1	0.09	0.16	0.12
	AIAN	0.05	0.07	0.04	0.08
	Unknown	10.01	11.2	10.59	9.81
Gender	Female	62.89	39.25	59.82	54.43
	Male	37.11	60.75	40.15	45.56
	Missing	0.01	0	0.03	0.01

AD: Alzheimer's Disease, PD: Parkinson's Disease, OD: Other Dementias, BAA: Black or African American, NHOPI: Native Hawaiian or Other Pacific Islande, AIAN: American Indian or Alaska Native.

2.2 Dataset Preprocessing

Our raw data consisted of diagnosis code information in both ICD version 9 (ICD-9) and version 10 (ICD-10). We mapped these codes to their respective PheCodes.[29] These PheCodes were used for each terminology system (ICD-9 and ICD-10), aligning on the 'code_system' and 'code' fields. This also allowed us to collapse results to the PheCode level rather than using individual ICD-9 and ICD-10 codes. To enable quantitative analysis, we used one-hot encoding of these categorical data to transform those data into binary format with one column per unique PheCode. For each unique phenotype (PheCode) identified, we created a new column in the diagnosis data and assigned binary values indicating the presence (1) or absence (0) of the phenotype (PheCode) for each patient. The final dataset comprised of patient identifiers,

demographic information, and binary-encoded phenotypes, providing a structured and analyzable representation of the patient diagnosis data. We also transformed the Race variable using one-hot encoding with Race_White, Race_Black, Race_Asian and Race_Other with each corresponding to a binary relationship with the race variable. We also transformed the Hispanic and Sex_Male columns to binary variables. We also included features pertaining to the type of hospital-visit, including: chemotherapy, emergency_visit, inpatient_visit, ambulatory_visit, and other_unknown_visit. Like the demographic features, each of these was binary indicating that a patient had at least one occurrence of that particular type of visit or chemotherapy.

For the multi-variate analysis no missing data was allowed, and therefore the missing data for Hispanic and Sex_Male were coded with -1 to indicate that those values were missing. We decided not to use imputation methods because that could result in other biases. For the univariate analysis, this was not needed as each feature was assessed one at a time and therefore, if there was missing data for a feature then those rows would be dropped automatically from the analysis via the glm() function in R.

3. Methodology

3.1 Uni-variate Logistic Regression Association Mining

We utilize traditional EHR association mining methods.[18] To do this, we evaluate each NDD type as an outcome separately to compare the features association with that particular type of NDD. This allows us to identify features that are strongly associated with a particular NDD type, and also features that are only associated with one NDD type. For each outcome (AD, PD, OD), we test each feature for its association with the outcome. Each feature (N=1796) was tested for association with each outcome (hence uni-variate association mining). The majority of features were conditions/diseases represented by PheCodes. The non-PheCode features included demographic features: Race_White, Race_Black, Race_Asian, Race_Other, Hispanic and Sex_Male. Hospital-visit characteristic features were also explored including: chemotherapy, emergency visit, ambulatory visit, inpatient visit, and other unknown visit. Each of these was binary indicating that a patient had at least one occurrence of that particular type of visit or chemotherapy. Once all features were tested for association with each NDD type, we then removed the intercept terms from our model results and corrected for multiple hypothesis testing using the Bonferroni adjustment method, defined as:

$$\text{corrected p-value} = \alpha/N = 0.05/1796$$

where alpha represents our significance cutoff (0.05 in this case) and N represents the number of tests (1796 in this case).

We used Logistic Regression (LR) to test for the association between each feature and the NDD type, given that the outcome variables are binary. This analysis was performed in the statistical programming language R using the glm() function with the statistical family set to binomial (i.e., to perform LR). Importantly, while we tested 1796 features for association with each NDD type, in the Venn Diagrams we only show 1794 features because we removed the features that consist of the NDD types themselves (AD, PD, and OD).

3.2 Machine Learning (ML) Methods

We employed three distinct models: LR, Ridge Regression (RR), and a Residual Network

(ResNet) based Neural Network, to predict the occurrence of PD, AD and OD separately. The dataset, after preprocessing, consisted of a feature set of 1796 features per NDD type model. The data was split into training and testing sets in an 80:20 ratio. We used 'LogisticRegression' and 'Ridge' from Python package 'sklearn'.[30] We implemented a ResNet model using Keras, starting with an input layer for feature vectors, followed by a dense layer with 64 units, batch normalization, ReLU activation, and a dropout layer (rate 0.5) to prevent over-fitting. The model's core has five ResNet blocks, each comprising two dense layers with batch normalization, ReLU activation, and dropout (rate 0.5). The output of the second dense layer was added to the block's input tensor, followed by ReLU activation. The final output was generated by a dense layer with a single unit and sigmoid activation. We used the Adam optimizer (learning rate 0.001), binary cross-entropy loss, and accuracy as the metric. Early stopping with a patience of 5 epochs was employed to mitigate over-fitting. The model was trained for up to 50 epochs with a batch size of 32, using 20% of the training data for validation.

We performed 5-fold cross-validation on the training set for all the above-mentioned models. The models were then trained on the entire training dataset. A bootstrapping procedure generated multiple bootstrap samples from the test data, evaluated the model's accuracy on each sample, and used those accuracies to compute the 95% confidence intervals.

3.3 Analysis with SHAP

We aim to identify factors contributing to the progression of AD, PD, and OD using the SHAP method. To do this, we construct separate models for each NDD type, using patient attributes as predictors. The target variable was defined as the presence or absence of AD. Similarly, separate models were constructed for PD and OD. The value of the target variable is 1 if the targeted event happened to the subject during the whole project and 0 otherwise. For each outcome (AD, PD, OD), we train the models mentioned in subsection 3.2 to determine the presence or absence of the disease. The SHAP method from the 'shap'[25] Python package was used to identify significant features using LR and RR. We used LinearExplainer for both LR and RR models. [a]

3.4 Feature Selection and Top 5% Subset

For methods that used multi-variate approaches, we selected features as being important if the mean shapley value for that feature was greater than or equal to the overall mean shapley value for that NDD type.[31] For the uni-variate approach, we selected features as important if their Bonferroni adjusted p-value was statistically significant. For the 5% subset, we selected the top 5% of features from each method and each NDD type. The top 5% of features amounts to 90 features from our overall feature set. The features are ranked based on their mean shapley value if a multi-variate method, or their p-value if a uni-variate method.

4. Results

4.1 Uni-Variate Assoication Mining Results

We found 340 significant associations with AD, 590 significant associations with PD and

[a] For the ResNet-based Neural Network, we encountered significant computational constraints using KernelExplainer due to its inherent complexity and the large size of the dataset. Consequently, to maintain consistency in our analysis of feature importance, we proceeded without considering the contributions derived from the neural network model.

583 significant associations with OD using uni-variate LR. Table 2 reports findings based on nominal significance, Bonferroni adjusted significance and a combination of Bonferroni significance and Odds Ratio ≥ 2. We visualized the uni-variate LR results using Manhattan plots for each NDD type: AD (in Figure 3A), PD (in Figure 3B) and OD (in Figure 3C). One can see that there are many Bonferroni significant results spread across the NDD types, but that AD has only a few significant results (see Figure 3A).

Table 2. Number of Association Mining Results

Results	AD	PD	OD
Number of Nominal Significant Results	723	1017	971
Number of Bonferroni Significant Results	340	590	583
Number of Bonferroni Significant Results and OR ≥ 2	3	16	278

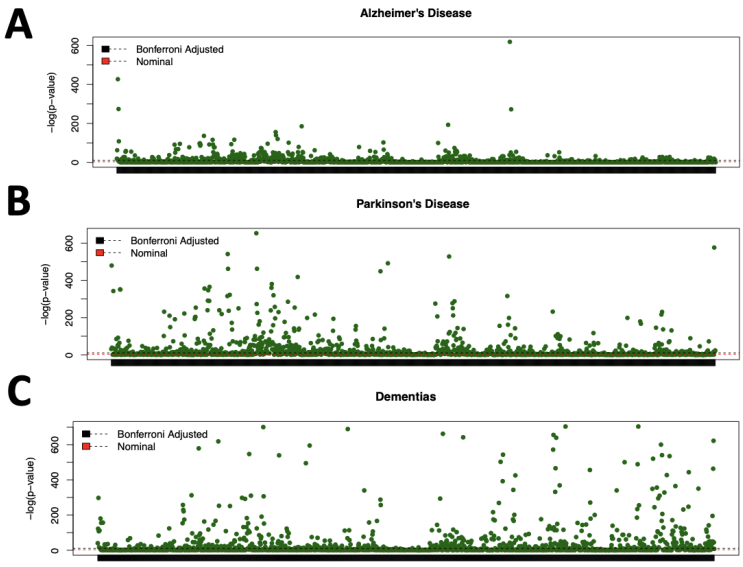

Fig. 3. Manhattan Plot for NDD type: AD, PD, OD

4.2 Performance of Multi-variate Methods for: AD, PD, OD

Table 3. Accuracy Performance of ML Methods by NDD type (All Features, N=1796).

Out-come	Logistic Regression (LR)				Ridge Regression (RR)				Neural Net (ResNet)			
	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI(95%) [lower, upper]
AD	78.48	80.1	79.44	[78.74, 80.08]	78.8	79.85	79.43	[78.73, 80.07]	79.03	79.87	79.26	[78.59, 79.89]
PD	85.07	86.51	85.17	[84.58, 85.77]	83.34	84.29	83.21	[82.58, 83.83]	84.83	88.49	84.89	[84.32, 85.5]
OD	84.91	86.37	84.5	[83.91, 85.11]	84.37	85.5	84.15	[83.56, 84.75]	86.49	88.51	86.34	[85.76, 86.92]

Table 4. Accuracy Performance of ML Methods by NDD type (Intersection Features).

Outcome	Logistic Regression (LR)				Ridge Regression (RR)				Neural Net (ResNet)			
	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI(95%) [lower, upper]	cv (mean)	Train	Test	CI (95%) [lower, upper]
AD (N=180)	78.99	79.2	79.54	[78.88, 80.18]	78.99	79.03	79.47	[78.76, 80.13]	79.95	80.25	80.32	[79.67, 80.94]
PD (N=225)	85.54	85.76	85.05	[84.46, 85.63]	83.28	83.47	83.2	[82.56, 83.84]	86.15	87.87	86.16	[85.57, 86.71]
OD (N=205)	85.06	85.29	84.8	[84.24, 85.41]	84.45	84.56	84.29	[83.73, 84.88]	88.22	89.95	87.98	[87.44, 88.51]

Tables 3 and 4 present the performance metrics of patients with NDD types of PD, AD, and OD, assessed using different ML models: LR, RR, and ResNet. For each NDD type, both of the tables display cross-validation mean accuracy (cv mean), the training and testing accuracy percentages alongside the 95% confidence intervals (CIs) for both the lower and upper bounds with respect to testing accuracy. The test accuracies are obtained using a held-out independent test set. The main difference between the two tables is the number of features used for training the models. In Table 3, all features were used, while in Table 4, only selected intersectional subsets of features mentioned in Section 5 were used. In both tables, we additionally included PD and AD as features with OD as the outcome, PD and OD as features with AD as the outcome, and AD and OD as features with PD as the outcome while evaluating the models' performances.

Results for all 1796 features shown in Table 3. LR, RR and ResNet models show slight variations in the mean cv, training, testing accuracies and CIs. In Table 4, the models show comparable results, highlighting the contribution of the reduced feature sets of 180 features for AD, 225 for PD, and 205 for OD. Tables 3 and 4 demonstrate that the use of selected features, as opposed to all features, does not significantly degrade model performance. Specifically, the slight differences in test accuracy, e.g., 79.44% vs. 79.54% for AD classification using LR, indicate that the models maintain robust performance even with reduced feature sets.

4.3 Overlap of Features Across Methods per NDD type

Characteristics of important features are given in Table 5. Non-overlapping features represent those that are unique to one method. We show the results for the 5% subset and the entire set of important features. AD had the lowest amount of non-overlapping features at 30.7% indicating that many of the features found by methods when applied to AD were similar Table 5. However, both PD and OD had higher amounts of non-overlapping features (i.e., unique) with 48.3% and 51.3% respectively in Table 5. Depending on the particular use case, some researchers may want to use only the top important features, which is our rationale for the top 5% feature subset from each method. This results in the same number of features being selected per method (i.e., 90 features). We found that for the 2 NDD types with less overlap (i.e., PD and OD) there were fewer non-overlapping features in the top 5% subset with 35.8% vs. 48.3% for PD, and 41.8% vs. 51.3% for OD. However, for AD, which already had a high agreement across methods, the top 5% of features actually had more non-overlapping features with 38.6% vs. 30.7% in the 5% subset in Table 5.

Table 5. Characteristics of Important Features

Results	AD	PD	OD
All Important Features			
Total Number of Important Features Across Methods	567	681	708
Number of Non-Overlapping Features	174	329	363
Percentage of Non-Overlapping Important Features	30.7%	48.3%	51.3%
Top 5% Feature Subset			
Total Number of Important Features Across Methods	140	137	146
Number of Non-Overlapping Features	54	49	61
Percentage of Non-Overlapping Important Features	38.6%	35.8%	41.8%

Feature penetrance indicates how often a feature was determined to be important by one of the three methods used for each NDD type. We also calculated penetrance across all NDD types, therefore a feature could have a maximum of 9 to indicate that it was found across all 3 NDD types and methods.[28] However, in some situations differences across methods maybe important. The Venn diagrams show the results for all important features and for only the top 5% of features for AD Figure 4A, for PD Figure 4B and for OD Figure 4C. Results for the intersections appear similar across the NDD types with OD having a larger number of features overall, mainly resulting from the large number of OD results generated by the uni-variate LR approach. Figure 4C.

Many interesting unique features were identified using the uni-variate LR method, including the association between Creutzfeldt-Jakob Disease or (CJD) and OD with a large reported Odds Ratio (OR=51.87, 95% CI: 18.88, 214.32). Note that CJD is listed in the PheCodes as Jakob Creutzfeldt Disease (PheCode:324.1). The percentage of individuals with CJD and OD was 93.18% versus 4.55% with AD and 2.27% with PD (Figure 5).

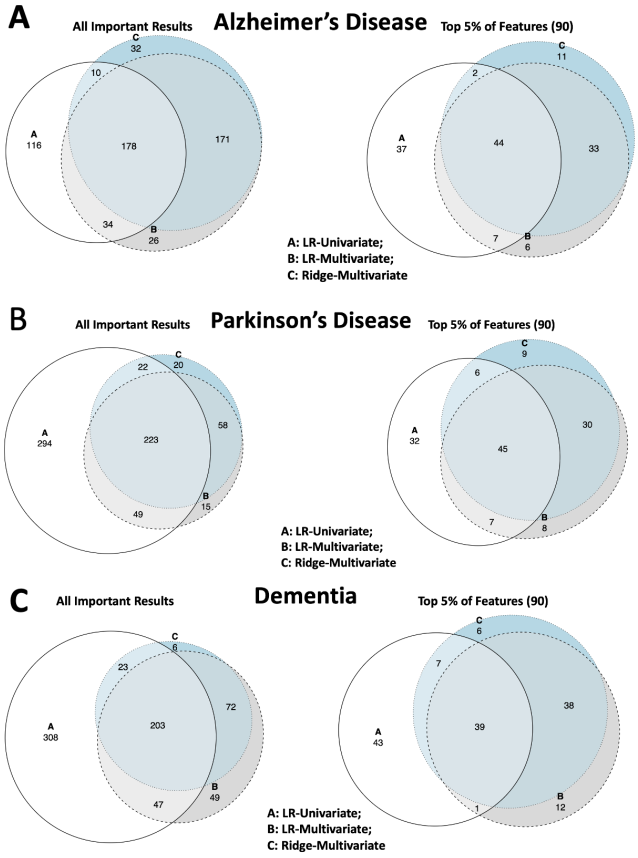

Fig. 4. Venn Diagram of Important Features for NDD type: AD, PD, and OD.

5. Discussion

5.1 Overview of Study

Overall, our study found that it is possible to identify important features for different NDD types, specifically AD, PD and OD. We found that the performance obtained using the specific method (LR, RR or ResNet) in terms of accuracy varied somewhat by NDD type, with all achieving similar performance. We also found that while methods achieved similar performance overall, there were substantial differences in 'important' features revealed by each

method. We identified features that were common (i.e., found by each method) and also features that were unique to one particular method. Therefore, our findings are important for others using EHR data for ADRD analyses because the 'important' features identified not only varies by statistical method used, but also by NDD type. Because of the heterogeneity of EHR data, the exact prevalence of each NDD type may vary by site to site, making this finding of importance for those utilizing EHR data for ADRD analyses.

The features found in the intersection of all 3 methods, namely uni-variate LR, multivariate LR (identified using shapley values), and multi-variate RR (identified via shapley values) may represent the most significant and robust features. These features are of particular interest because they are important across multiple methods, suggesting they are less likely to be influenced by confounding factors. In contrast, features identified by only one method may be less reliable and could be artifacts of the specific analytical approach used. Therefore, focusing on the intersecting features provides a more comprehensive and reliable understanding of the key predictors in the dataset. However, we will describe below circumstances that illustrate the advantages and disadvantages of various methods and features identified by the methods indicating that the intersection features may include only a subset of the truly 'important' features.

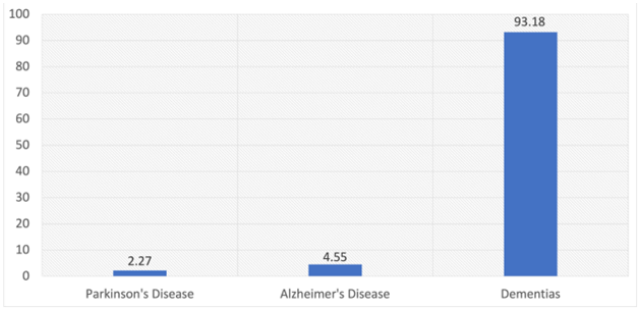

Fig. 5. Distribution of Creutzfeldt-Jakob disease (CJD) by NDD types.

5.2 Uni-variate versus Multivariate Models

CJD Disease Identified via Uni-variate Method Alone. There are some findings that were only uncovered via the uni-variate LR approach. It was the only method that revealed that CJD was significant in OD (one of the NDD types), and clearly there is a dramatic difference in our dataset for the prevalence observed among those with CJD with the majority of individuals having OD (see Figure 5). CJD is established as a rare cause of dementia[32] and therefore, this finding is of clinical significance and would be missed in multivariate approaches due to the overall rarity of this disease. However, there have been studies that found that CJD could be mistaken for AD,[33] indicating that clinically distinguishing these various diseases can be challenging in different circumstances. CJD is an example of one of the 308 features identified for OD that were only identified using the uni-variate LR approach (see Figure 4C). There were 6 OD features uniquely identified via the Multi-variate RR approach, but these features were odd, and included 'late pregnancy and failed induction' along with 'genital prolapse', which indicates that perhaps these findings were associated with a lower chance of OD. However, our population only includes those who are 65 and older and therefore, these features existing in our cohort remains somewhat odd. Features unique to the multi-variate LR approach also appeared somewhat unusual, including 'elevated Prostate Specific Antigen'(PSA test). This is an unusual finding given that our OD patients were predominantly female.

Sleep Apnea. Interestingly, the PheCode for Sleep Apnea (PheCode:327.30) was found

to be significant for AD across all 3 methods, including both multi-variate and uni-variate approaches. However, the uni-variate LR approach also identified another related PheCode for Obstructive Sleep Apnea (PheCode:327.32) as being significantly important for the AD type. The 'Obstructive Sleep Apnea' or OSA PheCode was **not** identified as being important by the other multi-variate approaches, and indicates a finding unique to the uni-variate method for the AD type, one of the 116 unique features in Figure 4A. This also represents a clinical relevant finding as OSA has been linked with AD specifically in a number of studies[34,35] indicating its importance in the AD type.

Overall Advantages of Uni-variate Alone. Overall, these findings highlight a main advantage of uni-variate LR (sometimes referred to as a 'traditional approach' for ML) in that it enables one to calculate Odds Ratios (OR) and to determine whether a finding increases or decreases the risk of diagnosis for each NDD type. Shapley values on the other hand provide the importance of the feature without the directionality of the finding, which in some cases makes them more difficult to interpret, and might be the reason for some of these results identified as unique to the multi-variate approaches. Overall, our findings suggest that uni-variate LR may be better at detecting NDD-type-specific differences, even with smaller sample sizes like we observed with CJD. Features supported across the methods appear to be more robust than features identified by just one method - unless that method was a uni-variate approach (again due to the advantages of ORs and directionality of the result).

5.3 Performance Varies by NDD type: AD, PD, and OD

Based on the models' performance presented in Tables 3 and 4, it is evident that the performance of different ML methods varies depending on both the NDD type and the method used. When using LR with all features, the highest accuracy was achieved when detecting PD with a test accuracy of 85.17%, while the lowest accuracy was for AD with a test accuracy of 79.44%. In contrast, when using RR and ResNet, the highest accuracy was for OD with a test accuracy of 84.15% and 86.34%, but the lowest accuracy was again for AD with a test accuracy of 79.43% and 79.26%. The performances using the subset of the features also demonstrate similar pattern. RR had the lowest test accuracy for all 3 NDD types across all features (in table 3) and the intersectional subsets of features (in table 4). While ResNet demonstrated the highest test accuracy using all features in classifying OD, LR had the highest test accuracy for AD and PD. On the other hand, in Table 4 the ResNet model significantly outperformed the regression models' test accuracies for all NDD types. The ResNet's superior performance implies the possibility that this increased performance is due to the neural network's capacity to learn complex, non-linear relationships, which might be more important for certain NDD types.

5.4 Comparison of Feature Results Across Methods

Interestingly, when applying Neural Net with selected features for prediction purposes, there was an increase in accuracy across all NDD types compared to using all features as presented in Tables 3 and 4. Both AD and OD predictions yield higher test accuracies with intersection features rather than the full feature set with all the 3 models. Although PD results demonstrate a slight decrease in test accuracies for both LR and RR while using instersectional features, the train accuracies also decrease. These observations suggest that

our feature selection enhances model performance by reducing noise and focusing on the most relevant information. The use of intersection features, which encapsulate the most critical and discriminative attributes, facilitates better generalization across models, reducing overfitting and improving robustness.

5.5 Implications of Our Findings on Other ADRD ML Studies

Spectrum bias[36,37] occurs when a test is studied among a population that is not representative of the intended target population. For example, if a study is conducted on an ADRD population using EHR data in Florida with a large population of ADRD patients having OD and then that method was applied in a population from Delaware where the majority of ADRD patients have PD, that could result in spectrum bias. Therefore, it is important to understand the important disease features that are unique to each NDD type: AD, PD and OD because the case-mix distribution of patients among ADRD patients may vary across the USA. Therefore, to develop robust ML models, we must understand the relationships between these features and each NDD type to understand if models (ours and others) will validate adequately at other locations in the USA treating ADRD patients.

6. Conclusion

In conclusion, we utilized a large (70,420 patients) ADRD cohort derived from EHR data collected during routine clinical care. Our cohort is an order of magnitude larger in size (70k versus 7k) than another recent ML ADRD study using EHR data.[6] Using this large and comprehensive dataset, we aimed to identify important diagnostic features for the NDD types using a variety of ML methods. Our study demonstrates the strengths and weakness of univariate and multivariate ML methods in detecting features specific to certain NDD types, namely, AD, PD and OD. We report accuracies of these methods and report what NDD types where each method worked best. We also identified features that were found across all methods, and features that were unique to a particular method. We share these findings with the research community with the goal of mitigating spectrum bias in ADRD studies as the NDD types vary from site to site across the USA and could therefore introduce biases if not accounted for.

7. Acknowledgments and Appendices

Research reported in this publication/ presentation was supported by the National Institute On Aging of the National Institutes of Health under Award Number P30AG073105. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. ALZ.org, Alzheimer's disease facts and figures, *Accessed in July 2024* **https://www.alz.org/alzheimers-dementia/facts-figures** (2024).
2. A. S. Tang, T. Oskotsky, S. Havaldar, W. G. Mantyh, M. Bicak, C. W. Solsberg, S. Woldemariam, B. Zeng, Z. Hu, B. Oskotsky *et al.*, Deep phenotyping of alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations, *Nature communications* **13** (2022).
3. G. M. Babulal, Y. T. Quiroz, B. C. Albensi, E. Arenaza-Urquijo, A. J. Astell, C. Babiloni, A. Bahar-Fuchs, J. Bell, G. L. Bowman, A. M. Brickman *et al.*, Perspectives on ethnic and racial disparities in alzheimer's disease and related dementias: update and areas of immediate need, *Alzheimer's & Dementia* **15**, 292 (2019).
4. A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo and I. S. Kohane, Genetic misdiagnoses and the potential for health disparities, *New England Journal of Medicine* **375**, 655 (2016).
5. A. L. Chin, S. Negash and R. Hamilton, Diversity and disparity in dementia: the impact of ethnoracial differences in alzheimer disease, *Alzheimer Disease & Associated Disorders* **25** (2011).
6. J. Xu, F. Wang, Z. Xu, P. Adekkanattu, P. Brandt, G. Jiang, R. C. Kiefer, Y. Luo, C. Mao, J. A. Pacheco *et al.*, Data-driven discovery of probable alzheimer's disease and related dementia subphenotypes using electronic health records, *Learning Health Systems* **4**, p. e10246 (2020).
7. D. S. Char, N. H. Shah and D. Magnus, Implementing machine learning in health care—addressing ethical challenges, *New England Journal of Medicine* **378**, 981 (2018).
8. I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman and M. Ghassemi, Ethical machine learning in healthcare, *Annual review of biomedical data science* **4**, 123 (2021).
9. J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden and D. C. Crawford, Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations, *Bioinformatics* **26**, 1205 (2010).
10. M. Saad, A. El-Menyar, K. Kunji, E. Ullah, J. Al Suwaidi and I. J. Kullo, Validation of polygenic risk scores for coronary heart disease in a middle eastern cohort using whole genome sequencing, *Circulation: Genomic and Precision Medicine* **15**, p. e003712 (2022).
11. A. Verma, A. Lucas, S. S. Verma, Y. Zhang, N. Josyula, A. Khan, D. N. Hartzel, D. R. Lavage, J. Leader, M. D. Ritchie *et al.*, Phewas and beyond: the landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from geisinger, *The American Journal of Human Genetics* **102**, 592 (2018).
12. X. Li, X. Meng, A. Spiliopoulou, M. Timofeeva, W.-Q. Wei, A. Gifford, X. Shen, Y. He, T. Varley, P. McKeigue *et al.*, Mr-phewas: exploring the causal effect of sua level on multiple disease outcomes by using genetic instruments in uk biobank, *Annals of the rheumatic diseases* **77** (2018).
13. J. Pathak, R. C. Kiefer, S. J. Bielinski and C. G. Chute, Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank, *Journal of biomedical semantics* **3**, 1 (2012).
14. R. W. Read, K. A. Schlauch, G. Elhanan, W. J. Metcalf, A. D. Slonim, R. Aweti, R. Borkowski and J. J. Grzymski, Gwas and phewas of red blood cell components in a northern nevadan cohort, *PLoS One* **14**, p. e0218078 (2019).
15. B. Namjou, K. Marsolo, R. J. Caroll, J. C. Denny, M. D. Ritchie, S. S. Verma, T. Lingren, A. Porollo, B. L. Cobb, C. Perry *et al.*, Phenome-wide association study (phewas) in emr-linked pediatric cohorts, genetically links plcl1 to speech language development and il5-il13 to eosinophilic esophagitis, *Frontiers in genetics* **5**, p. 401 (2014).
16. X. Chang, M. March, F. Mentch, H. Qu, Y. Liu, J. Glessner, P. Sleiman and H. Hakonarson, Genetic architecture of asthma in african american patients, *Journal of Allergy and Clinical*

Immunology **151**, 1132 (2023).
17. Y.-C. A. Feng, I. B. Stanaway, J. J. Connolly, J. C. Denny, Y. Luo, C. Weng, W.-Q. Wei, S. T. Weiss, E. W. Karlson and J. W. Smoller, Psychiatric manifestations of rare variation in medically actionable genes: a phewas approach, *BMC genomics* **23**, p. 385 (2022).
18. M. R. Boland, S. Alur-Gupta, L. Levine, P. Gabriel and G. Gonzalez-Hernandez, Disease associations depend on visit type: results from a visit-wide association study, *BioData Mining* **12**, 1 (2019).
19. M. R. Boland, Z. Shahn, D. Madigan, G. Hripcsak and N. P. Tatonetti, Birth month affects lifetime disease risk: a phenome-wide method, *Journal of the American Medical Informatics Association* **22**, 1042 (2015).
20. L. Li, M. R. Boland, R. Miotto, N. P. Tatonetti and J. T. Dudley, Replicating cardiovascular condition-birth month associations, *Scientific reports* **6**, p. 33166 (2016).
21. M. R. Boland, M. Fieder, L. H. John, P. R. Rijnbeek and S. Huber, Female reproductive performance and maternal birth month: a comprehensive meta-analysis exploring multiple seasonal mechanisms, *Scientific Reports* **10**, p. 555 (2020).
22. M. R. Boland, P. Parhi, L. Li, R. Miotto, R. Carroll, U. Iqbal, P.-A. Nguyen, M. Schuemie, S. C. You, D. Smith *et al.*, Uncovering exposures responsible for birth season–disease effects: a global study, *Journal of the American Medical Informatics Association* **25**, 275 (2018).
23. M. R. Boland, M. S. Kraus, E. Dziuk and A. R. Gelzer, Cardiovascular disease risk varies by birth month in canines, *Scientific Reports* **8**, 1 (2018).
24. C. Molnar, *Interpretable machine learning : a guide for making Black Box Models interpretable* (Lulu, 2019).
25. S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, *Advances in neural information processing systems* **30** (2017).
26. L. S. Shapley and A. E. Roth, *The Shapley value : essays in honor of Lloyd S. Shapley* (Cambridge University Press, 1988).
27. ICDCodes, Icd10 to icd9 code coverter, *Accessed in July 2024* **https://icd.codes/convert/icd10-to-icd9-cm** (2024).
28. M. R. Boland, Boland lab github: Alzheimer's disease and related dementias (adrd) project, *Accessed in July 2024* **https://github.com/bolandlab/AlzheimersDiseaseandRelatedDementias** (2024).
29. PheWAS, Phewas - phenome wide association studies.
30. L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt and G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.
31. L. Bloch, C. M. Friedrich and A. D. N. Initiative, Data analysis with shapley values for automatic subject selection in alzheimer's disease data sets using interpretable machine learning, *Alzheimer's Research & Therapy* **13**, 1 (2021).
32. R. Knight, Creutzfeldt-jakob disease: a rare cause of dementia in elderly persons, *Clinical infectious diseases* **43**, 340 (2006).
33. H. J. Tschampa, M. Neumann, I. Zerr, K. Henkel, A. Schröter, W. J. Schulz-Schaeffer, B. Steinhoff, H. A. Kretzschmar and S. Poser, Patients with alzheimer's disease and dementia with lewy bodies mistaken for creutzfeldt-jakob disease, *Journal of Neurology, Neurosurgery & Psychiatry* **71**, 33 (2001).
34. F. Emamian, H. Khazaie, M. Tahmasian, G. D. Leschziner, M. J. Morrell, G.-Y. R. Hsiung, I. Rosenzweig and A. A. Sepehry, The association between obstructive sleep apnea and alzheimer's disease: a meta-analysis perspective, *Frontiers in aging neuroscience* **8**, p. 78 (2016).
35. A. G. Andrade, O. M. Bubu, A. W. Varga and R. S. Osorio, The relationship between obstructive

sleep apnea and alzheimer's disease, *Journal of Alzheimer's Disease* **64**, S255 (2018).
36. C. Goehring, A. Perrier and A. Morabia, Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance, *Statistics in medicine* **23**, 125 (2004).
37. S. A. Mulherin and W. C. Miller, Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation, *Annals of internal medicine* **137**, 598 (2002).