

Distinct medical and substance use histories associate with cognitive decline in Alzheimer's Disease

Clayton Mansel¹⁺, B.A.; Diego R. Mazzotti², Ph.D.; Ryan Townley³, M.D.; Mihaela E Sardu⁴, Ph.D.; Russell H. Swerdlow³, M.D.; Robyn A Honea³, Ph.D.; and Olivia J Veatch¹, Ph.D.

1. Department of Cell Biology and Physiology, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS, 66160
2. Department of Internal Medicine, Division of Medical Informatics, Division of Pulmonary Critical Care and Sleep Medicine, University of Kansas Medical Center, Kansas City, KS 66160
3. Alzheimer's Disease Research Center, University of Kansas Medical Center, 4350 Shawnee Mission Pkwy, Mail Stop 6002, Fairway, KS, 66205
4. Department of Biostatistics and Data Science, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS, 66160

+ Corresponding Author

Abstract

INTRODUCTION: Phenotype clustering reduces patient heterogeneity and could be useful when designing precision clinical trials. We hypothesized that the onset of early cognitive decline in patients would exhibit variance predicated on the clinical history documented prior to an Alzheimer's Disease (AD) diagnosis

METHODS: Self-reported medical and substance use history (i.e., problem history) was used to cluster participants from the National Alzheimer's Coordinating Centers (NACC) into distinct subtypes. Linear mixed effects modeling was used to determine the effect of problem history subtype on cognitive decline over two years.

RESULTS: 2754 individuals were partitioned into three subtypes: minimal (n = 1380), substance use (n = 1038), and cardiovascular (n = 336) subtypes. The cardiovascular problem history subtype had significantly worse cognitive decline over a two-year follow-up period (p = 0.013).

DISCUSSION: Our study highlights the need to account for problem history to reduce heterogeneity of outcomes in AD clinical trials.

1. Background

Alzheimer's disease (AD) is the most common form of dementia, with an estimated prevalence in the U.S. of 6 million as of 2020 [1]. AD presentation is clinically and biologically heterogeneous with many factors affecting the progression of the disease including socioeconomic status [2], nutrition [3], apolipoprotein E (*APO3*) genotype [4], sex [4], and co-morbidities including diabetes, depression, and hypertension [5–7]. Decades of clinical trials have resulted in only three Food and Drug Administration approved therapies for AD in the last 20 years which showed only a modest slowing of cognitive decline [8,9]. One reason for these underwhelming results may be that AD trials generally enroll participants as one homogenous group, despite evidence that AD progression and response to therapy is highly heterogeneous [10]. When not accounted for, this heterogeneity could inflate the type II error rate in clinical trials and minimize the average treatment effect of interventions when relevant subgroups are not accounted for.

The advent of large biomedical databases and machine learning algorithms, such as clustering, enables precision medicine approaches focused on identifying subgroups of individuals according to complex patterns in a hypothesis-independent manner and reducing heterogeneity in patient populations [11–13]. These data-driven clusters can then be used to augment future clinical trial enrollment and predict treatment response [14,15]. For example, Seymour and colleagues [16] used 29 variables to retrospectively cluster participants in the ProCESS clinical trial which aimed to improve outcomes in patients with sepsis [17]. Even though the original trial showed nearly zero percent chance of benefit, when informed by phenotype clusters derived in this study, the chance of benefit rose to 35% [16].

Most previous clustering studies in AD utilize either cognitive tests [18–21] or biological data such as neuroimaging or fluid biomarkers in patients after a diagnosis has been made [22–24]. However, when using these data, it can be difficult to distinguish between true subgroups or different stages of the disease after onset [25]. Moreover, clinical trials generally try to target individuals early in the disease progression [8,9] when cognitive decline and biomarkers are less pronounced. One approach to address this dilemma is to use medical and substance use history (i.e., problem history) prior to disease onset which is commonly collected at every routine primary care visit. We hypothesized that problem history subtypes could be especially relevant to AD heterogeneity given that many previous studies have shown that co-morbidities and substance use, such as smoking and alcohol intake, affect AD progression [6,26,27]. A few previous studies have performed clustering in AD using problem history data; however, they utilize data from the electronic health record (EHR) [28–31]. The accuracy of AD diagnoses using International Classification of Diseases (ICD) billing codes is mixed at best [32–35] with one study identifying a lack of cognitive testing and time as major barriers in a primary care setting [36]. To ensure accurate AD diagnosis, we utilized the Uniform Data Set (UDS) [37], with consistent testing and diagnosis procedures, from the National Alzheimer's Coordinating Centers (NACC) which is comprised of specialist Alzheimer's Disease Research Centers

(ADRCs) across the U.S. and is one of the largest clinical databases of individuals with AD in the world. In a cohort of individuals with AD, we performed multivariate clustering of medical and substance use survey items (problem history items) collected prior to AD diagnosis. To infer whether problem history subtypes could be used to inform future AD clinical trials, we compared each cluster's mean trajectory of cognitive decline over the next two years.

2. Methods

2.1 Data source

This cross-sectional study analyzed data from UDS annual visits between September 2005 and November 2020 across 40 ADRCs in the NACC database. Details regarding data collection are well documented.[38] For individual problem history, self-reported data (available 2005-present) were chosen instead of clinician-assessed medical conditions (available 2015-present) to minimize missingness and increase statistical power.

29,818 individuals in the NACC database had at least 1 follow-up visit. Individuals were included if he or she had: 1) normal cognitive status, impaired but not mild cognitive impairment (MCI), or MCI at the initial visit, 2) dementia at any follow-up visit, and 3) dementia was determined to have a “primary etiology” of AD. The primary etiology of dementia was determined using the individual's most recent visit to maximize diagnosis accuracy. Individuals <50 years old were removed to exclude autosomal-dominant forms of AD. Additionally, 298 individuals with missing problem history data were excluded. Ultimately, data from 2,754 individuals were included in the final cluster analysis (Figure 1; Table 1).

2.2 Cluster Analysis

In total, 26 self-reported variables reflecting problem history were used as input for clustering (Table 2). Variables ranged from cardiovascular disease history to neurological conditions and substance use. For each variable, individuals could indicate “Absent”, “Recent/Active”, or “Remote/Inactive.” Both the “Recent/Active” and “Remote/Inactive” answers were collapsed into “Present” because 1) the difference was often not clinically relevant and 2) the individuals' judgment of what is “remote” may introduce bias into the study.

Data preprocessing and clustering was conducted in R (version 4.3.1) and RStudio (version 2023.06.1+524). Initially, multiple correspondence analysis (MCA) [39] was performed to: 1) reduce the overall dimensionality of the data, and 2) transform categorical data into continuous component scores for clustering. To determine the number of components necessary to explain the majority of the variability observed in the dataset, the mean-squared error of prediction (MSEP) was plotted after performing k-fold cross-validation with 5% missing values added to the dataset across 100 simulations using the missMDA (v1.18) package in R. Five components substantially decreased the MSEP and were thus retained as input for clustering. The FactoMineR (v2.8) [40] and factoextra (v1.07) packages were used to perform MCA and visualize results.

Agglomerative hierarchical clustering was performed on individual component scores using the FactoMineR (v2.8) package. Individual similarity was determined using Euclidean distance and Ward's method to build the tree [41]. Inertia gain estimates were calculated when dividing the dataset between 2 and 10 clusters. The final cluster solution was determined by the largest relative drop in inertia gain which resulted in three clusters [39]. See Supplementary Figure 1 for the dendrogram as well as the top five most closely associated categories for each problem history cluster. Overall, clusters characterized and named according to which problem history items were present in higher proportions than the other clusters (Table 2).

2.3 Supplemental Variable Definitions

Clusters were further characterized by other features not used as input in the clustering algorithm. Age, sex, race, and ethnicity were self-reported and provided by the NACC. *APOE* genotypes were supplied by the NACC when available [42–44]. The NACC received genotypes from the participating ADRCs, ADGC, and the National Centralized Repository for Alzheimer's Disease (<https://naccddata.org>).

2.4 Statistics and Longitudinal Analysis

First, chi-square tests of independence (or Fisher's exact test when the expected count of a category was $n < 5$) were performed to determine differences in categorical non-transformed input variables and supplemental variables between clusters. One-way analyses of variance (ANOVA) were used to compare continuous supplemental measures among clusters. Unadjusted p-values less than 0.05 were considered significant.

The progression of cognitive decline for individuals in each cluster was characterized using a linear mixed effects model. The outcome variable was the Clinical Dementia Rating Sum of Boxes (CDRSUM) which ranges from 0 to 18, with higher numbers indicating worsening cognitive decline. CDRSUM was chosen because of its clinical relevance and use as an outcome in recent AD clinical trials [9]. Three annual study visits were included in the longitudinal analysis to: 1) minimize the effect of study dropout between subsequent visits and 2) focus on a follow-up period that is typical for a phase III AD clinical trial [8,9]. Age, sex, *APOE* genotype, baseline CDR, visit number, and cluster membership were included as fixed effects and subject ID was included as a random effect to account for repeated measures. Prior to conducting analyses, 316 (11.4%) individuals were dropped because *APOE* genotype was missing, leaving 2,438 individuals. Of these, 2,407 individuals (98.7%) completed at least three NACC study visits. To account for non-random censoring between clusters, the model was weighted by the inverse probability of censoring estimated using a binomial general linear model with the same fixed effects as the unweighted model. Sex by cluster and age by cluster interaction terms were tested in the weighted model and not included because they were not significant at $p = 0.05$.

A likelihood ratio test was used to evaluate the contribution of the problem history cluster on CDRSUM trajectory. To investigate differences between clusters at visit three, the pairwise

contrasts between the marginal means of each cluster was compared using a Wald test. Unadjusted p values less than 0.05 were considered significant. Longitudinal analyses were conducted using R version 4.3.1 and the *lme4*, *ggeffects*, and *ipw* packages.

3. Results

Overall, there were 2,754 participants in the NACC database who did not have AD at their initial visit but were diagnosed at a later visit and met cohort criteria (see Methods and Figure 1). The overall dataset had a mean (SD) age of 76.2 (8.3) years, was 55.1% female, 84.9% white, 94.4% non-Hispanic, and had a mean (SD) of 15.7 (6.0) years of education (Table 1). We then performed hierarchical clustering on principal components from MCA, which resulted in three problem history subtypes summarized in Table 2. The three subtypes can be described as 1) minimal problem history (n = 1,380), 2) substance use history (n = 1,038), and 3) cardiovascular problem history (n = 336). Age at initial visit, sex, and ethnicity were significantly different across the three subtypes (Table 1). The cardiovascular problem history subtype had a numerically higher mean age of 78.6, lower proportion of females (28.0%), and higher proportion of non-Hispanic individuals (97.0%) (Table 1). The demographics of race and years of education were not significantly different across the subtypes. Every problem history item was significantly different between the subtypes except vitamin B12 deficiency, thyroid disease, traumatic brain injury, seizures, other Parkinsonian disorders, urinary incontinence, and depression (Table 2). The minimal problem history subtype was noted to have the most problem history variables reported as ‘Absent’ across nearly all categories, including fewer reports of smoking >100 cigarettes lifetime (1.5%), angioplasty/endarterectomy/stent placement (0.4%), and cardiac arrest (0.6%) (Table 2). The substance use history subtype had the most individuals indicating a smoking (6.9%) and alcohol abuse (6.5%) history compared to other subtypes (Table 2). The cardiovascular problem history subtype had the highest proportion of individuals indicating a significant cardiovascular disease history including a heart attack/cardiac arrest (42.9%), atrial fibrillation (20.5%), and hypertension (74.4%), among others (Table 2).

Next, we further characterized these problem history subtypes according to the age of AD diagnosis, family history, *APOE* genotype, and co-occurrence of other types of dementia. The age of AD diagnosis was significantly different across subtypes with the cardiovascular problem history cluster having the numerically highest mean age (80.9) (Table 3). It should be noted that the overall mean (SD) age at the NACC initial visit was 76.2 (8.3) and the overall mean (SD) age of AD diagnosis was 78.5 (8.8) which results in a mean (SD) difference of 2.4 (2.5) between evaluation and later diagnosis. The subtypes were also significantly different with respect to early-onset (age range 50 to 64) versus late-onset (>65 years) AD. The minimal problem history subtype had the numerically highest proportion of early-onset AD (6.6%) and the cardiovascular problem history subtype had numerically highest proportion of late-onset AD (97.0%) (Table 3). The proportion of a self-reported family history of AD was also significantly different on both the maternal and paternal side (Table 3). The cardiovascular problem history subtype had a markedly lower proportion of individuals reporting a maternal family history (28.9%) than the

overall dataset (36.3%). The cardiovascular problem history subtype also had a numerically lower proportion of individuals reporting a paternal family history, although the comparison to the overall dataset was less (14.9% vs 17.1% for the cardiovascular problem history subtype compared to the overall dataset respectively). The difference in the frequency of *APOE* genotypes across subtypes was nearly significant ($p = 0.08$). Compared to the overall dataset, the largest differences with respect to *APOE* genotype were the $\epsilon 3/\epsilon 3$ genotype in the cardiovascular problem history subtype (36.0% vs 43.8%) and the $\epsilon 4/\epsilon 4$ genotype in the cardiovascular problem history subtype (9.8% vs 6.2%) in the overall dataset versus within-subtype respectively. No differences between subtypes were observed with respect to the co-occurrence of vascular, Lewy body, or frontotemporal dementia (Table 3).

To determine the significance of problem history on the subsequent cognitive decline of the NACC participants we modeled the longitudinal Clinical Dementia Rating Sum of Boxes (CDRSUM) scale over the next two annual follow-up visits (Figure 2). Problem history subtypes had significantly different changes in CDRSUM while controlling for age, sex, *APOE* genotype, baseline CDRSUM, and visit number ($p = 0.013$). At visit three, the substance use subtype had the lowest adjusted mean (\pm SE) change in CDRSUM compared to baseline of 1.51 ± 0.07 . The cardiovascular problem history subtype had the highest adjusted mean (\pm SE) change in CDRSUM of 1.96 ± 0.16 . Post-hoc analyses indicated that these two subtypes were significantly different at visit 3 ($p = 0.005$) (Figure 2; Supplementary Table 1).

4. Discussion

In this study, we leveraged the NACC database to perform clustering of medical and substance use history (i.e., problem history) in a cohort of participants who developed incident Alzheimer's Disease at a follow-up visit. We identified three unique problem history subtypes: minimal problem history, substance use history, and cardiovascular problem history. Notably, cognitive decline among problem history subtypes varied with the level of effect being clinically significant over a subsequent follow-up of two years. The difference between the adjusted mean change in CDRSUM between the substance use history and cardiovascular problem history subtype at visit three was 0.45. This is the same difference that was seen between the treatment and placebo groups at 18 months of follow-up in the phase III lecanemab trial [9]. Thus, heterogeneity in problem history among clinic trial participants is likely a significant factor in clinical trials. Future clinical trials should therefore consider problem history in their inclusion/exclusion criteria and their analysis.

Furthermore, our AD clustering study expands on findings from other studies that have clustered on EHR-derived problem history variables in patients already diagnosed with AD. Notably, all three previous clustering studies found a group with higher prevalence of cardiovascular disease—a finding corroborated by our study [28,30,31]. Moreover, we showed a significantly faster decline in cognition over a two-year follow-up period in the cardiovascular problem history subtype. Cardiorespiratory fitness has been observed as a mediating factor in AD

treatment [45] and AD progression [46]. Future clinical trials should therefore either specifically target these individuals or account for a cardiovascular disease history to minimize the chance of a type II error. Alexander et al., found a subtype of AD patients with higher rates of smoking that was also associated with more depression and anxiety diagnoses and a faster rate of AD progression [31]. Our study, however, detected a substance use history cluster with no significant difference in rates of depression and a slower cognitive decline over two years. Although depression and substance use disorders are often comorbid [47,48], other reports have found conflicting results in different sexes [49] and ages [50] which could account for our different results. It is also possible that our study was not sensitive enough to characterize depression and anxiety histories given that the data we analyzed included only two questions pertaining to mental health (Table 2). Finally, our study identified a novel minimal problem history cluster with a higher proportion of female and Hispanic individuals. This group of individuals is historically understudied in healthcare [51] and the prevalence of AD in this population is expected to grow by an estimated 460% by 2060 [1]—highlighting the need for future studies to target these individuals.

Our study has a few limitations. First, the NACC cohort analyzed in this study was comprised primarily of highly educated white individuals and thus may not as generalizable as studies using the EHR. However, one advantage of our NACC cohort is that all participants are seen at specialist memory centers and thus likely have more accurate diagnoses. By contrast, EHR-derived data are limited by the issue that AD is often misdiagnosed outside of specialist memory centers [52,53] and that co-pathology can make it difficult to distinguish between different types of dementia [54]. Second, our inclusion criteria targeted individuals who did not have a diagnosis of AD but would receive one in the future—thus the estimated cognitive decline over a two-year period is not generalizable to a wide clinic population of older adults with early cognitive decline. However, our cohort is similar to individuals who would be targeted for a late-phase clinical trial and thus our results remain very relevant to future AD intervention strategies [55]. Third, we were limited to the 26 problem history variables in the NACC questionnaire. Future studies should include additional problem history variables such as generalized anxiety disorder, sleep disorders, kidney disease, autoimmune conditions, etc.

In conclusion, we found three problem history subtypes in a clinical cohort of AD patients: minimal problem history, substance use history, and cardiovascular problem history. The minimal problem history cluster had a higher proportion of Hispanic females and earlier onset AD. The substance use history subtype was comprised primarily of individuals who had smoked more than one hundred cigarettes over their lifetime and had the slowest decline in cognition over a two-year follow-up period. The cardiovascular problem history subtype was older than the other clusters and had a higher proportion of white male individuals. This subtype also had a significantly higher decline in cognition over a two-year follow-up period—indicating results from our study may be informative to future work aimed at designing precision clinical trials.

References

1. Rajan KB, Weuve J, Barnes LL, McAninch EA, Wilson RS, Evans DA. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimers Dement J Alzheimers Assoc.* 2021;17(12):1966-1975. doi:10.1002/alz.12362
2. Hu D, Liu C, Xia K, Abramowitz A, Wu G, Alzheimer's Disease Neuroimaging Initiative (ADNI). Characterizing the Resilience Effect of Neurodegeneration for the Mechanistic Pathway of Alzheimer's Disease. *J Alzheimers Dis JAD.* 2021;84(3):1351-1362. doi:10.3233/JAD-215160
3. Loeffler DA. Modifiable, Non-Modifiable, and Clinical Factors Associated with Progression of Alzheimer's Disease. *J Alzheimers Dis JAD.* 2021;80(1):1-27. doi:10.3233/JAD-201182
4. Neu SC, Pa J, Kukull W, et al. Apolipoprotein E Genotype and Sex Risk Factors for Alzheimer Disease: A Meta-analysis. *JAMA Neurol.* 2017;74(10):1178-1189. doi:10.1001/jamaneurol.2017.2188
5. Modrego PJ, de Cerio LD, Lobo A. The Interface between Depression and Alzheimer's Disease. A Comprehensive Approach. *Ann Indian Acad Neurol.* 2023;26(4):315-325. doi:10.4103/aian.aian_326_23
6. Santiago JA, Potashkin JA. The Impact of Disease Comorbidities in Alzheimer's Disease. *Front Aging Neurosci.* 2021;13:631770. doi:10.3389/fnagi.2021.631770
7. Ravona-Springer R, Luo X, Schmeidler J, et al. Diabetes is associated with increased rate of cognitive decline in questionably demented elderly. *Dement Geriatr Cogn Disord.* 2010;29(1):68-74. doi:10.1159/000265552
8. Sims JR, Zimmer JA, Evans CD, et al. Donanemab in Early Symptomatic Alzheimer Disease: The TRAILBLAZER-ALZ 2 Randomized Clinical Trial. *JAMA.* 2023;330(6):512-527. doi:10.1001/jama.2023.13239
9. Van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in Early Alzheimer's Disease. *N Engl J Med.* 2023;388(1):9-21. doi:10.1056/NEJMoa2212948
10. Jutten RJ, Sikkes SAM, Van Der Flier WM, et al. Finding Treatment Effects in Alzheimer Trials in the Face of Disease Progression Heterogeneity. *Neurology.* 2021;96(22). doi:10.1212/WNL.0000000000012022
11. Pan Q, Hu T, Malley JD, Andrew AS, Karagas MR, Moore JH. A system-level pathway-phenotype association analysis using synthetic feature random forest. *Genet Epidemiol.* 2014;38(3):209-219. doi:10.1002/gepi.21794
12. Veatch OJ, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes Brain Behav.* 2014;13(3):276-285. doi:10.1111/gbb.12117
13. Tai AMY, Albuquerque A, Carmona NE, et al. Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artif Intell Med.* 2019;99:101704. doi:10.1016/j.artmed.2019.101704

14. Loftus TJ, Shickel B, Balch JA, et al. Phenotype clustering in health care: A narrative review for clinicians. *Front Artif Intell.* 2022;5:842306. doi:10.3389/frai.2022.842306
15. Sharma A, Zheng Y, Ezekowitz JA, et al. Cluster Analysis of Cardiovascular Phenotypes in Patients With Type 2 Diabetes and Established Atherosclerotic Cardiovascular Disease: A Potential Approach to Precision Medicine. *Diabetes Care.* 2022;45(1):204-212. doi:10.2337/dc20-2806
16. Seymour CW, Kennedy JN, Wang S, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *JAMA.* 2019;321(20):2003-2017. doi:10.1001/jama.2019.5791
17. ProCESS Investigators, Yealy DM, Kellum JA, et al. A randomized trial of protocol-based care for early septic shock. *N Engl J Med.* 2014;370(18):1683-1693. doi:10.1056/NEJMoa1401602
18. Scheltens NME, Tijms BM, Koene T, et al. Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. *Alzheimers Dement.* 2017;13(11):1226-1236. doi:10.1016/j.jalz.2017.03.002
19. Scheltens NME, Galindo-Garre F, Pijnenburg YAL, et al. The identification of cognitive subtypes in Alzheimer's disease dementia using latent class analysis. *J Neurol Neurosurg Psychiatry.* 2016;87(3):235-243. doi:10.1136/jnnp-2014-309582
20. Davidson JE, Irizarry MC, Bray BC, et al. An exploration of cognitive subgroups in Alzheimer's disease. *J Int Neuropsychol Soc JINS.* 2010;16(2):233-243. doi:10.1017/S1355617709991160
21. Libon DJ, Drabick DAG, Giovannetti T, et al. Neuropsychological syndromes associated with Alzheimer's/vascular dementia: a latent class analysis. *J Alzheimers Dis JAD.* 2014;42(3):999-1014. doi:10.3233/JAD-132147
22. Förstl H, Levy R, Burns A, Luthert P, Cairns N. Pathways and patterns of cell loss in verified Alzheimer's disease: a factor and cluster analysis of clinico-pathological subgroups. *Behav Neurol.* 1994;7(3):175-180. doi:10.3233/BEN-1994-73-411
23. Poulakis K, Pereira JB, Mecocci P, et al. Heterogeneous patterns of brain atrophy in Alzheimer's disease. *Neurobiol Aging.* 2018;65:98-108. doi:10.1016/j.neurobiolaging.2018.01.009
24. Varol E, Sotiras A, Davatzikos C. HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *NeuroImage.* 2017;145:346-364. doi:10.1016/j.neuroimage.2016.02.041
25. Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun.* 2018;9(1):4273. doi:10.1038/s41467-018-05892-0
26. Anstey KJ, Von Sanden C, Salim A, O'Kearney R. Smoking as a Risk Factor for Dementia and Cognitive Decline: A Meta-Analysis of Prospective Studies. *Am J Epidemiol.* 2007;166(4):367-378. doi:10.1093/aje/kwm116
27. Lo AHY, Pachana NA, Byrne GJ, Sachdev PS. A Review of Tobacco, Alcohol, Adiposity, and Activity as Predictors of Cognitive Change. *Clin Gerontol.* 2012;35(2):148-194. doi:10.1080/07317115.2011.639856

28. 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). doi:10.1002/lrh2.10246
29. Peter J, Abdulkadir A, Kaller C, et al. Subgroups of Alzheimer's Disease: Stability of Empirical Clusters Over Time. *J Alzheimers Dis.* 2014;42(2):651-661. doi:10.3233/JAD-140261
30. He Z, Tian S, Erdengasileng A, Charness N. Temporal Subtyping of Alzheimer's Disease Using Medical Conditions Preceding Alzheimer's Disease Onset in Electronic Health Records. Published online May 23, 2022.
31. Alexander N, Alexander DC, Barkhof F, Denaxas S. Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Med Inform Decis Mak.* 2021;21(1):343. doi:10.1186/s12911-021-01693-6
32. Sommerlad A, Perera G, Singh-Manoux A, Lewis G, Stewart R, Livingston G. Accuracy of general hospital dementia diagnoses in England: Sensitivity, specificity, and predictors of diagnostic accuracy 2008–2016. *Alzheimers Dement.* 2018;14(7):933-943. doi:10.1016/j.jalz.2018.02.012
33. Ponjoan A, Garre-Olmo J, Blanch J, et al. How well can electronic health records from primary care identify Alzheimer's disease cases? *Clin Epidemiol.* 2019;Volume 11:509-518. doi:10.2147/CLEP.S206770
34. Schliep K, Shepelak Z, Bitter N, et al. Detecting early signs of Alzheimer's disease and related dementia onset from the EHR. *Innov Aging.* 2021;5(Supplement_1):643-644. doi:10.1093/geroni/igab046.2441
35. Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis Assoc Disord.* 2009;23(4):306-314. doi:10.1097/WAD.0b013e3181a6bebc
36. Maserejian N, Krzywy H, Eaton S, Galvin JE. Cognitive measures lacking in EHR prior to dementia or Alzheimer's disease diagnosis. *Alzheimers Dement.* 2021;17(7):1231-1243. doi:10.1002/alz.12280
37. Besser L, Kukull W, Knopman DS, et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Dis Assoc Disord.* 2018;32(4):351-358. doi:10.1097/WAD.0000000000000279
38. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) Database: The Uniform Data Set. *Alzheimer Dis Assoc Disord.* 2007;21(3):249-258. doi:10.1097/WAD.0b013e318142774e
39. Husson F. *Exploratory Multivariate Analysis by Example Using R*. 1st ed. CRC Press; 2011.
40. Lê S, Josse J, Husson F. FactoMineR : An R Package for Multivariate Analysis. *J Stat Softw.* 2008;25(1). doi:10.18637/jss.v025.i01
41. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc.* 1963;58(301):236-244. doi:10.1080/01621459.1963.10500845

42. Naj AC, Jun G, Beecham GW, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet.* 2011;43(5):436-441. doi:10.1038/ng.801
43. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat Genet.* 2019;51(3):414-430. doi:10.1038/s41588-019-0358-2
44. Wang LS, Naj AC, Graham RR, et al. Rarity of the Alzheimer disease-protective APP A673T variant in the United States. *JAMA Neurol.* 2015;72(2):209-216. doi:10.1001/jamaneurol.2014.2157
45. Yogev-Seligmann G, Eisenstein T, Ash E, et al. Neurocognitive Plasticity Is Associated with Cardiorespiratory Fitness Following Physical Exercise in Older Adults with Amnesic Mild Cognitive Impairment. Hackney M, ed. *J Alzheimers Dis.* 2021;81(1):91-112. doi:10.3233/JAD-201429
46. Dougherty RJ, Jonaitis EM, Johnson SC, Okonkwo OC, Cook DB. Cardiorespiratory fitness mitigates brain atrophy and cognitive decline in preclinical Alzheimer's disease. *Alzheimers Dement.* 2021;17(S10):e053971. doi:10.1002/alz.053971
47. Calarco CA, Lobo MK. Depression and substance use disorders: Clinical comorbidity and shared neurobiology. *Int Rev Neurobiol.* 2021;157:245-309. doi:10.1016/bs.irm.2020.09.004
48. Swendsen JD, Merikangas KR. The comorbidity of depression and substance use disorders. *Clin Psychol Rev.* 2000;20(2):173-189. doi:10.1016/s0272-7358(99)00026-4
49. Massak A, Graham K. Is the smoking-depression relationship confounded by alcohol consumption? An analysis by gender. *Nicotine Tob Res.* 2008;10(7):1231-1243. doi:10.1080/14622200802163449
50. Quittschalle J, Pabst A, Löbner M, et al. Association of Alcohol and Tobacco Consumption with Depression Severity in the Oldest Old. Results from the Age Different Old Age Cohort Platform. *Int J Environ Res Public Health.* 2021;18(15):7959. doi:10.3390/ijerph18157959
51. Lange JW. Methodological concerns for non-Hispanic investigators conducting research with Hispanic Americans. *Res Nurs Health.* 2002;25(5):411-419. doi:10.1002/nur.10049
52. Larner AJ. Getting it wrong: the clinical misdiagnosis of Alzheimer's disease: The Clinical Misdiagnosis of Alzheimer's Disease. *Int J Clin Pract.* 2004;58(11):1092-1094. doi:10.1111/j.1368-5031.2004.00314.x
53. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology.* 1984;34(7):939-939. doi:10.1212/WNL.34.7.939
54. Karantzoulis S, Galvin JE. Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Rev Neurother.* 2011;11(11):1579-1591. doi:10.1586/ern.11.155
55. Grill JD, Flournoy C, Dhadda S, et al. Eligibility Rates among Racially and Ethnically Diverse US Participants in Phase 2 and Phase 3 Placebo-Controlled, Double-Blind, Randomized Trials of

Lecanemab and Elenbecestat in Early Alzheimer Disease. *Ann Neurol.* 2024;95(2):288-298.
doi:10.1002/ana.26819

Figures

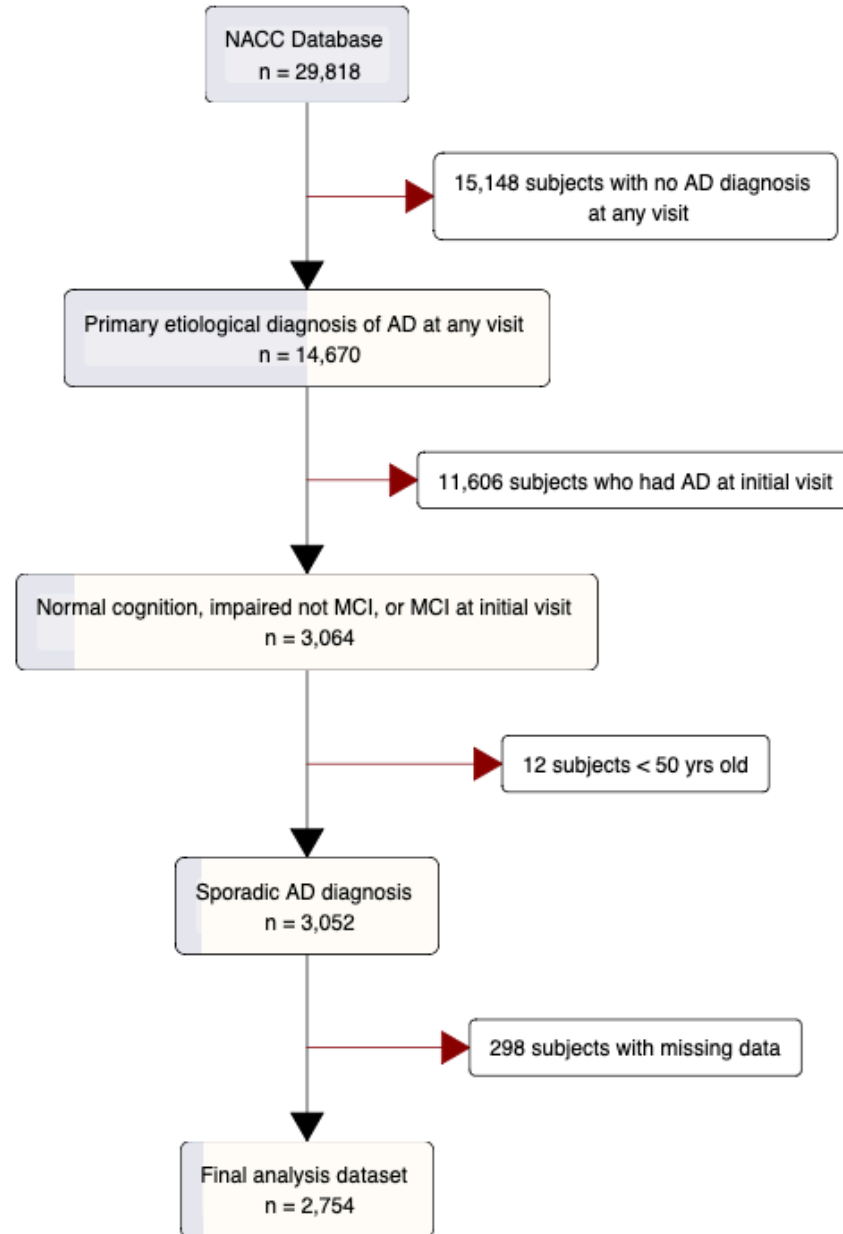


Figure 1: Flowchart of study cohort selection. NACC: National Alzheimer’s Coordinating Centers, AD: Alzheimer’s Disease, MCI: Mild Cognitive Impairment

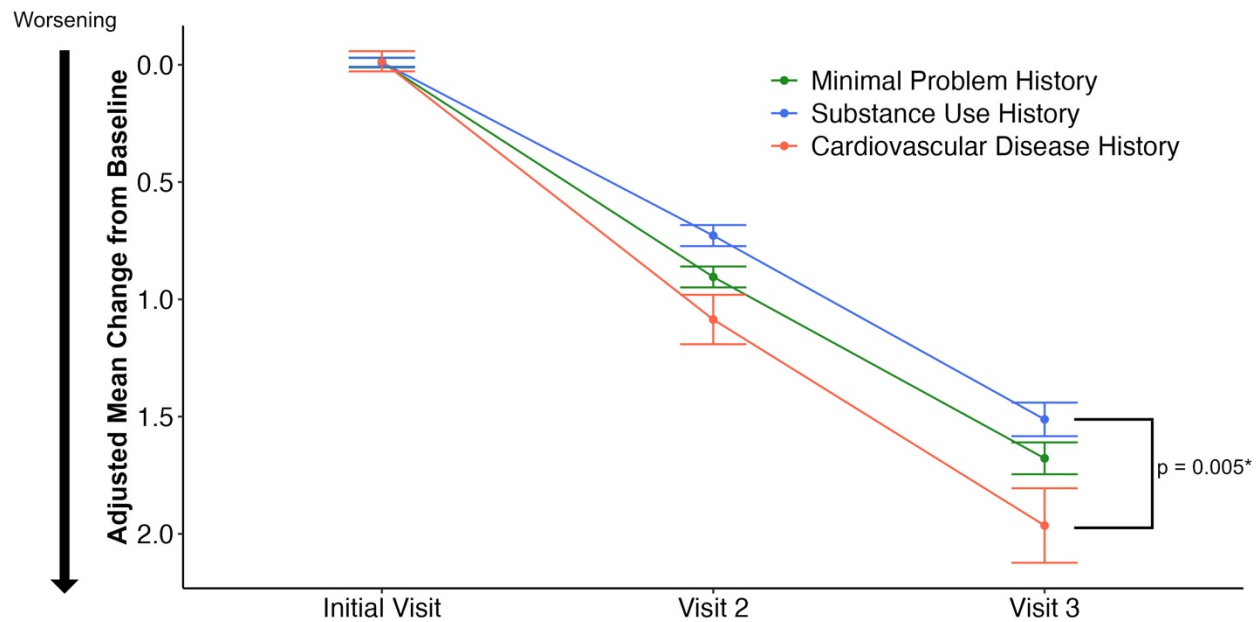


Figure 2: Change in CDR Sum of Boxes for each Problem History Cluster. Change in the Clinical Dementia Rating Sum of Boxes (CDRSUM) in each problem history cluster. Total scores range from 0-18 with higher numbers indicating worsening cognitive decline. Data are plotted as the adjusted mean CDRSUM (\pm standard error) change from baseline. Data are adjusted using a linear mixed effects model with age, sex, *APOE* genotype, and baseline CDRSUM included as fixed effects and subject ID included as a random effect. *p value corresponds to pairwise comparisons between the marginal mean of each cluster at visit 3.

Tables

Table 1. Demographics of NACC Cohort by Problem History Cluster

Factor	Unclustered (n = 2754)	Problem History Cluster			χ^2 ** (df) or Eta2	p value
		Minimal Problem History (n = 1380)	Substance Use History (n = 1038)	Cardiovascular Problem History (n = 336)		
Age at Initial Visit	76.18 (8.29)	76.09 (8.7)	75.55 (7.80)	78.55 (7.64)	0.012	<0.01*
Sex					152.9 (2)	<0.01*
Male	1237 (44.9)	492 (35.7)	503 (48.5)	242 (72.0)		
Female	1517 (55.1)	888 (64.3)	535 (51.5)	94 (28.0)		
Race					-	0.17
White	2337 (84.9)	1146 (83.0)	889 (85.6)	302 (89.9)		
Black or African American	295 (10.7)	160 (11.6)	112 (10.8)	23 (6.8)		
American Indian or Alaska Native	7 (0.3)	5 (0.4)	2 (0.2)	0 (0.0)		
Native Hawaiian or Pacific Islander	1 (0.0)	1 (0.1)	0 (0.0)	0 (0.0)		
Asian	67 (2.4)	41 (3.0)	18 (1.7)	8 (2.4)		
Other	38 (1.4)	22 (1.6)	14 (1.3)	2 (0.6)		
Unknown	9 (0.3)	5 (0.4)	3 (0.3)	1 (0.3)		
Ethnicity					-	0.01*
Non-Hispanic	2601 (94.4)	1285 (93.1)	990 (95.4)	326 (97.0)		
Hispanic	146 (5.3)	90 (6.5)	47 (4.5)	9 (2.7)		
Unknown	7 (0.3)	5 (0.4)	1 (0.1)	1 (0.3)		
Years of Education	15.7 (6.0)	15.7 (6.0)	15.7 (5.5)	16.0 (7.3)	<0.001	0.67

Continuous factors reported as mean (SD) and categorical factors reported as n (%). * denotes a p value of <0.05. ** Differences between clusters evaluated using Chi-square or Fisher's exact test for categorical variables and one-way ANOVA for quantitative variables. Chi-square (χ^2) or correlation ratio (Eta2) reported when applicable.

Table 2. Problem History Items by Cluster

	Item	Unclustered n = 2754	Minimal Problem History n = 1380	Substance Use History n = 1038	Cardiovascul ar Problem History n = 336	χ^2 (df)	p value
Cardiovascular Disease	Heart Attack/Cardiac Arrest	169 (6.13)	8 (0.58)	17 (1.64)	144 (42.86)	897.03 (2)	<0.01 *
	Atrial Fibrillation	200 (7.26)	59 (4.28)	72 (6.94)	69 (20.54)	106.34 (2)	<0.01 *
	Angioplasty/Endarterectomy/ Stent	203 (7.37)	5 (0.36)	16 (1.54)	182 (54.17)	1228.58 (2)	<0.01 *
	Cardiac Bypass Procedure	137 (4.97)	1 (0.07)	2 (0.19)	134 (39.88)	-	<0.01 *
	Pacemaker	95 (3.45)	23 (1.67)	16 (1.54)	56 (16.67)	200.76 (2)	<0.01 *
	Congestive Heart Failure	56 (2.03)	5 (0.36)	8 (0.77)	43 (12.80)	223.09 (2)	<0.01 *
	Other Cardiovascular Disease	289 (10.49)	119 (8.62)	120 (11.56)	50 (14.88)	13.28 (2)	<0.01 *
	Diabetes	334 (12.13)	159 (11.52)	112 (10.79)	63 (18.75)	16.05 (2)	<0.01 *
	Hypertension	1465 (53.20)	690 (50)	524 (50.48)	251 (74.70)	71.15 (2)	<0.01 *
	Hypercholesterolemia	1474 (53.52)	651 (47.17)	557 (53.66)	266 (79.17)	111.19 (2)	<0.01 *
Metabolic Conditions	Vitamin B12 Deficiency	155 (5.62)	79 (5.72)	59 (5.68)	17 (5.06)	0.23 (2)	0.89

	Thyroid Disease	539 (19.57)	280 (20.29)	194 (18.69)	65 (19.35)	0.98 (2)	0.61
Neurological Conditions	Stroke	111 (4.03)	57 (4.13)	25 (2.41)	29 (8.63)	25.48 (2)	<0.01 *
	Traumatic Brain Injury (TBI)	273 (9.91)	126 (9.13)	116 (11.18)	31 (9.23)	2.98 (2)	0.23
	Seizures	51 (1.85)	23 (1.67)	20 (1.93)	8 (2.38)	0.81 (2)	0.67
	Parkinson's Disease	8 (0.29)	8 (0.58)	0 (0)	0 (0)	-	0.02*
	Other Parkinsonian Disorder	16 (0.58)	8 (0.58)	6 (0.58)	2 (0.60)	-	0.99
	Urinary Incontinence	353 (12.82)	176 (12.75)	126 (12.14)	51 (15.18)	2.11 (2)	0.35
	Fecal Incontinence	72 (2.61)	53 (3.84)	13 (1.25)	6 (1.79)	16.62 (2)	<0.01 *
Psychiatric Disorders	Active Depression in the last 2 years	821 (29.81)	417 (30.22)	315 (30.35)	89 (26.49)	2.02 (2)	0.36
	Other Psychiatric Disorder	128 (4.65)	57 (4.13)	62 (5.97)	9 (2.68)	7.89 (2)	0.019 *
Tobacco Use	Smoked Cigarettes in the Last 30 Days	85 (3.09)	1 (0.07)	72 (6.94)	12 (3.57)	-	<0.01 *
	Smoked More than 100 Cigarettes in Life	1199 (43.54)	20 (1.45)	1027 (98.94)	152 (45.24)	2290.94 (2)	<0.01 *
	Average Number of Packs Smoked Per Day:					-	<0.01 *
	0	1552 (56.35)	1360 (98.55)	9 (0.87)	183 (54.46)		
	1 cigarette to < 1/2 pk	417 (15.14)	10 (0.72)	367 (35.36)	40 (11.90)		
	1/2 pk to < 1 pk	406 (14.74)	9 (0.65)	358 (34.49)	39 (11.61)		
	1 pk to 1 & 1/2 pks	196 (7.12)	0 (0)	163 (15.70)	33 (9.82)		

	1 & 1/2 pks to 2 pks	99 (3.59)	0 (0)	79 (7.61)	20 (5.95)		
	> 2 pks	84 (3.05)	1 (0.07)	62 (5.97)	21 (6.25)		
Substance Use	Alcohol Abuse - Clinically Significant	98 (3.56)	20 (1.45)	67 (6.45)	11 (3.27)	43.34 (2)	<0.01 *
	Other Substance Abuse	13 (0.47)	0 (0)	13 (1.25)	0 (0)	-	<0.01 *

Variable categories (n (%)) shown for every variable used to construct the problem history clusters. Differences between the clusters were tested using a Chi-square test of independence. If the expected frequency was less than five, a Fisher's exact test was utilized. * denotes a p value of <0.05.

Table 3. Alzheimer's Disease Age of Onset, Family History, APOE Genotype, and Other Dementia Co-Occurrence for Each Problem History Cluster

	Problem History Cluster				χ^2 ** (df) or Eta2	p value
	Unclustered (n = 2754)	Minimal Problem History (n = 1380)	Substance Use History (n = 1038)	Cardiovascular Problem History (n = 336)		
Age of AD Diagnosis	78.54 (8.79)	78.47 (9.2)	77.90 (8.33)	80.85 (8.0)	0.01 0	<0.01 *
AD Onset					7.2 (2)	0.03*
Early-Onset AD (>49 yrs & <65 yrs)	155 (5.60)	91 (6.6)	54 (5.2)	10 (3.0)		
Late-Onset AD (\geq 65 yrs)	2599 (94.4)	1289 (93.4)	984 (94.8)	326 (97.0)		
Maternal Family History of AD					15.1 (4)	<0.01 *
Present	1001 (36.3)	507 (36.7)	397 (38.2)	97 (28.9)		
Absent	1661 (60.3)	817 (59.2)	613 (59.1)	231 (68.8)		
Unknown	92 (3.3)	56 (4.1)	28 (2.7)	8 (2.4)		
Paternal Family History of AD						0.03*

Present	471 (17.1)	232 (16.8)	189 (18.2)	50 (14.9)	10.8 (4)	
Absent	2152 (78.1)	1068 (77.4)	806 (77.6)	278 (82.7)		
Unknown	131 (4.8)	80 (5.8)	43 (4.1)	8 (2.4)		
APOE Genotype					-	0.08
E3/E3	992 (36.0)	500 (36.2)	345 (33.2)	147 (43.8)		
E3/E4	952 (34.6)	461 (33.4)	380 (36.6)	111 (33.0)		
E3/E2	154 (5.6)	79 (5.7)	57 (5.5)	18 (5.4)		
E4/E4	270 (9.8)	143 (10.4)	106 (10.2)	21 (6.2)		
E4/E2	67 (2.4)	32 (2.3)	25 (2.4)	10 (3.0)		
E2/E2	3 (0.1)	2 (0.1)	1 (0.1)	0 (0.0)		
Missing/Unknown	316 (11.5)	163 (11.8)	124 (11.9)	29 (8.6)		
Co-Occurrence of Vascular Dementia						
Present	320 (11.6)	162 (11.7)	109 (10.5)	49 (14.6)		
Absent	2434 (88.40)	1213 (88.3)	929 (89.5)	287 (85.4)		
Co-Occurrence of Lewy Body Dementia					0.44 (2)	0.80
Present	90 (3.3)	44 (3.2)	33 (3.2)	13 (3.9)		
Absent	2664 (96.70)	1336 (96.8)	1005 (96.8)	323 (96.1)		
Co-Occurrence of Frontotemporal Dementia					-	0.45
Present	23 (0.80)	9 (0.7)	10 (1.0)	4 (1.2)		
Absent	2731 (99.20)	1371 (99.3)	1028 (99.0)	332 (98.8)		

Continuous factors reported as mean (SD) and categorical factors reported as n (%). * denotes a p value of <0.05. ** Differences between clusters evaluated using Chi-square or Fisher's exact test for categorical variables and one-way ANOVA for quantitative variables. Chi-square (χ^2) or correlation ratio (Eta²) reported when applicable.

Acknowledgements/Conflicts/Funding Sources/Consent Statement

Acknowledgements: RAH and OJV conceived the presented idea with input from MES and RT. CM and OJV designed the analysis plan with critical feedback from MES. CM and OJV performed data quality control procedures, conducted clustering analysis, and drafted the initial manuscript. RAH developed the approach for defining maternal family history. RS, RT, RAH, and DRM provided expertise regarding the analysis and interpretation of findings. All authors revised, discussed, and approved the final manuscript.

Conflicts: The authors have no conflicts of interest

Funding Sources: The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

The Alzheimer's Disease Genetic Consortium (ADGC) data is funded by NIA/NIH Grant U01 AG032984.

This work was also performed with funding from NIH P30 AG072973 and T32 AG078114 (CM); NIH P20 GM130423 (OJV) and P30 AG035982 (OJV, RAH).

Consent Statement: Informed consent was not necessary for this study because all data were de-identified.

Keywords: Alzheimer's Disease, Phenotype Clustering, Cardiovascular Disease, National Alzheimer's Coordinating Centers, Clinical Dementia Rating, Substance Use, Cognitive Decline