Alzheimer's Disease: Exploratory Analysis of Genetic and Demographic Associations in a Synthetic Dataset

Pranavi Raina[1]

Received May 12, 2026
Accepted July 21, 2026
Electronic access September 15, 2026

Alzheimer's disease (AD) is a neurodegenerative disease characterized by cognitive decline, memory loss, and later, loss of basic physical functioning. It is the leading cause of dementia and represents a growing global health burden, projected to rise in the future. This study carried out an exploratory analysis of genetic and demographic risk factors associated with Alzheimer's disease diagnostic status using the Alzheimer's Disease Data Initiative (ADDI) Dataset, which included 1,500 individuals classified as cognitively healthy, having mild cognitive impairment (MCI), or having Alzheimer's disease. The variables examined included age, sex, APOE genotype, Mini Mental State Examination (MMSE) scores, and geographic region. There were statistical analyses conducted in R using descriptive statistics and contingency table methods, including odds ratios with 95% confidence intervals, Fisher's exact test, and chi-square tests. Age above 65 years was significantly associated with MCI relative to cognitively healthy status (OR = 3.18; 95% CI: 1.24-8.58; p = 0.017), whereas APOE $\varepsilon 4$ carriage did not have any statistical significance in any disease group comparison. There was no statistically significant relationship between APOE $\varepsilon 4$ carriage and diagnostic classification in the exploratory pairwise analyses (OR = 1.01; 95% CI: 0.68-1.51; p = 0.95). Such results may be due to some limitations in the dataset rather than biological processes. Because the synthetic group lacked environmental exposures, future research should associate genetic testing with lifestyle measurements to evaluate genetic and demographic relationships.

Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by the loss of neurons and their connections in the cerebral cortex, within the hippocampus[1]. These changes lead to cognitive decline, memory impairment, and eventual loss of independent functioning[1,2]. Globally, neurodegenerative diseases represent a rising public health crisis. Around 55 million people worldwide are currently living with dementia, and this number is projected to rise to 139 million by 2050, according to the World Health Organization (WHO)[3]. Within the United States, current clinical records indicate that over 6.9 million individuals aged 65 and older suffer from an active AD diagnosis, generating healthcare infrastructure expenditures exceeding $360 billion annually, with long-term cost projections tracking past $1 trillion by mid-century[4]. In addition to the financial strain on healthcare companies, the toll on society includes an estimated 18.4 billion hours of informal care provided once a year by family members and other support systems[4]. So, improving models that analyze early diagnosis and patient factors is essential for better screening and treatment[5].

The biological basis of late-onset Alzheimer's disease (AD) is complex and involves the interaction of genetic risk factors, demographic factors, and environmental exposures[6]. The main hallmarks of the disease, deposits of toxic amyloid-beta outside cells and tangles of tau proteins inside cells, can take decades to develop before symptoms appear[2]. Large-scale genetic studies have significantly expanded our understanding of the many genes involved in late-onset AD, identifying numerous risk areas related to immune system responses[7–10]. The most important of these genes is apolipoprotein E (APOE), which is found on chromosome 19[11]. Variations of the APOE gene produce three main types of proteins, $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$, which have different shapes and interact uniquely with lipid molecules[12].

Individuals with one copy of the APOE $\varepsilon 4$ allele have 3 to 4 times the risk of developing AD, and individuals with two copies of the APOE $\varepsilon 4$ allele have 8 to 12 times the risk of developing AD compared to the most common $\varepsilon 3/\varepsilon 3$ type[13]. The $\varepsilon 4$ protein structurally changes its interaction with fats and hinders the brain's ability to remove amyloid plaque[12,14]. However, the less common $\varepsilon 2$ gene is seen as a protective factor that might delay the age at which the disease begins[15].

Overall, genetic risk tests should be considered with fixed factors such as biological sex and age[16,17]. Long-term clinical systems show that advanced age is the main predictor for the progression from mild cognitive impairment (MCI) to an official dementia diagnosis[18]. On the other hand, deep-phenotyping research has found that the female sex is linked with increased tau pathology build-up and symptoms, and the

[1] *Kent Place School, USA*

biological mechanisms of this vulnerability are still under investigation [19].

It can be noted that currently up to around 45% of the global burden of dementia can potentially be prevented or delayed through the proper management of 14 modifiable risk factors [6,20]. Randomized controlled trials such as FINGER have demonstrated that a number of interventions are effective in maintaining cognitive performance and reducing the rate of cognitive decline in at-risk older patients [21], including dietary modification, physical activity, cognitive training and metabolic monitoring. These changing factors interact within a complex system of socio-economic and structural factors such as childhood education, air quality, cardiovascular health, sleep hygiene and access to healthcare [6,22].

While much has been done in terms of health factor studies, it is very complicated to analyze multiple variables due to ethical restrictions, costs, and confidential information problems [23]. To overcome these challenges, researchers are using synthetic data, computer-generated information, to simulate large groups of patients [24]. Some wonder whether these groups actually reflect real-world patterns observed in real human trials, although this method does make research more accessible [25].

Figures 1-3 represent background images provided from outside resources. The original analysis results used in this research are shown in Figures 4-6, and details on the data sources are given in figure captions and Methods.

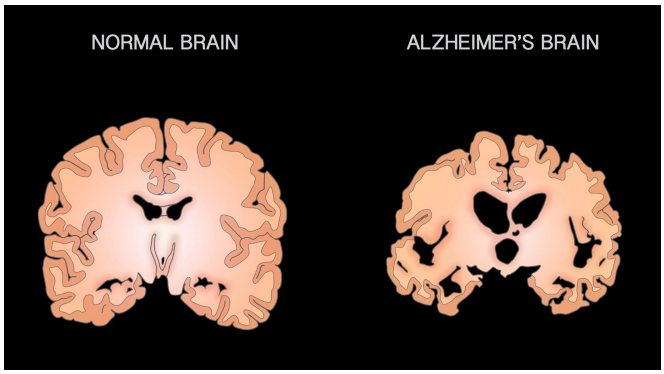

Figure. 1 Structural differences between a normal brain and an Alzheimer's disease brain. Image used under academic publishing guidelines courtesy of BrightFocus Foundation [26].

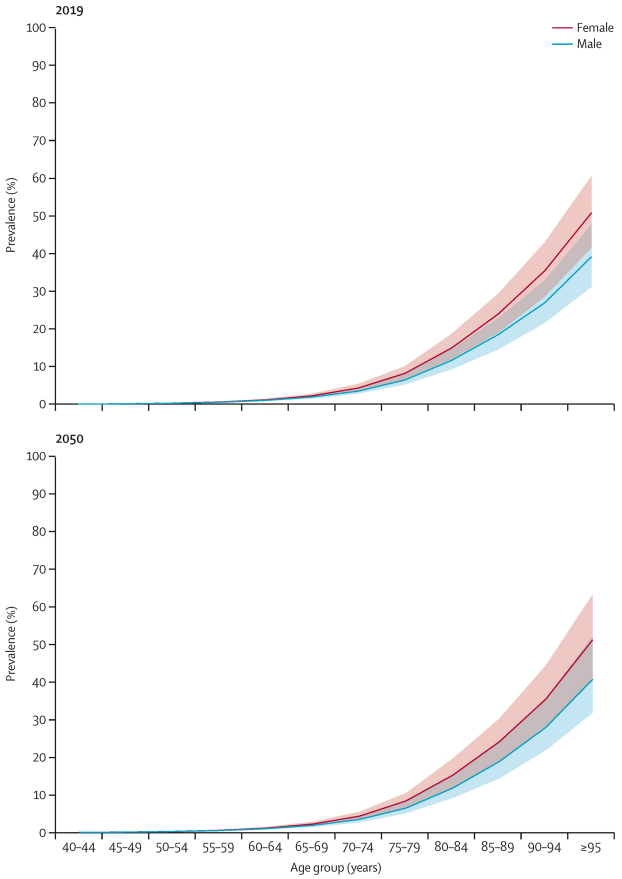

Figure. 2 Estimated prevalence of dementia (%), by age group and sex in 2019 and projections to 2050. Risk curves stratified by biological sex across global age parameters. Reproduced from GBD 2019 Dementia Forecasting Collaborators [4].

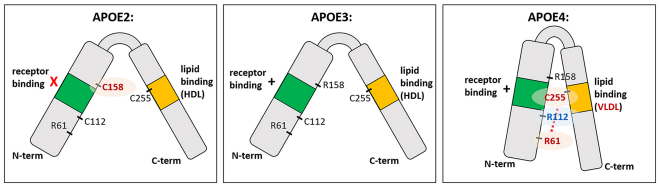

Figure. 3 Structural differences between APOE ε2, APOE ε3, and APOE ε4 isoforms and their connection with the risk of Alzheimer's disease development. Reproduced from Fernandez et al. (2019), Frontiers in Aging Neuroscience.

Methodology

Data Source and Study Population

Analysis is based on the ADDI (Alzheimer's Disease Data Initiative) Synthetic Dataset. This dataset contains clinical and demographic information about 1,500 subjects diagnosed with cognitive health, mild cognitive impairment (MCI), and Alzheimer's disease (AD). Variables analyzed were age, sex, APOE genotype, MMSE score, geographic region, and diagnosis.

Data Processing

The dataset was imported into R (version 4.3.2) for analysis. Completeness of the data was tested prior to the analysis. In the case of all primary variables used in this study using the ADDI Synthetic Dataset, there were no missing values in case of age, sex, APOE genotype, MMSE score, and diagnosis. Therefore, no imputation procedures were needed.

Age was grouped into clinically studied age groups, including $<$65 years and $\geq$65 years. Participants were measured for APOE genotype and carrier status (both carrier and non-carrier). The MMSE test score was grouped into clinically important categories: 25-30, 20-24, 15-19, and $<$10. These bins correspond to the bands of dementia severity that are mapped onto the Mini-Mental State Examination: questionable/normal, mild, moderate and severe. This is similar to what was characterized by Perneczky et al.[27], who associated MMSE scores with Clinical Dementia Rating staging (questionable 26-29, mild 21-25, moderate 11-20, and severe 0-10). Each group consisted of five per level. The cut-off points are approximate and have no correlation with anything mentioned in the literature.

All analyses were performed in R version 4.3.2[28]. Contingency table comparisons were performed using the base R functions chisq.test() and fisher.test(), with Fisher's exact test used if any expected cell count was less than five. The odds ratios and 95% confidence intervals were determined by applying the continuity correction. Cumulative link mixed models were fitted using the function clmm() from the ordinal package[29]. The synthetic ADDI dataset[30] was obtained from the AD Workbench. Subjects used in the analysis came from 9 geographical sites (countries). All key variables such as age, gender, APOE status, MMSE score, and diagnosis were present without any need for imputation.

After introducing age, sex, APOE genotypes and MMSE scores into the analysis, a further analysis of the relationship between these factors was performed, controlling for any differences between sites. This confirmed the results to be correct.

Statistical Analysis

Demographic and clinical characteristics were summarized by the diagnostic group using descriptive statistics. Contingency tables were used to compare disease groups within each APOE genotype group and within each age group.

Odds ratios (ORs) with 95% confidence intervals (CIs) were calculated to assess the strength of associations between predictor variables and disease status. Chi-square tests were used when expected cell counts were large enough, and Fisher's exact tests were used when expected frequencies were less than five. Statistical significance was defined as $p < 0.05$.

As the primary aim of the present study was exploratory, the analyses were focused on pairwise associations rather than prediction. Thus, the results are interpreted as descriptive and hypothesis-generating.

To further investigate the relationships between variables, cumulative link mixed models were fitted with disease status treated as an ordinal outcome (Healthy $<$ MCI $<$ AD). Age, sex, APOE genotype, and MMSE score were included as fixed effects, and country was included as a random effect to account for geographical variation across the dataset.

Methodological Considerations

As this is not a predictive clinical model with alpha control, but rather a hypothesis generating and exploratory investigation, no formal statistical corrections for multiple comparisons (e.g., Bonferroni adjustments) were used for the pairwise computations. All reported p-values are to be considered as purely exploratory markers. Moreover, these pairwise contingency designs do not statistically control for confounding factors such as age or sex, and test raw cross-tabulations. Finally, the precise procedure for obtaining the synthetic diagnostic labels is not fully described in the materials available in the dataset. Therefore, it is impossible to verify whether MMSE scores were used directly for the assignment of the labels. Thus, MMSE was used descriptively for cognitive severity and not as an independent predictor. Since clinical diagnoses of AD and MCI are based on cognitive testing, MMSE would be expected to be strongly associated with diagnostic status by construction, and this association should be interpreted as a confirmation of internal consistency rather than as a demonstration of predictive value.

Results

Sample Composition

The dataset included 1,500 total participants across three disease groups:

APOE ε4 Carriage and Disease Status

There was no statistically significant association between diagnostic group and APOE ε4 carriage in any pairwise comparison (Table 2). For the Healthy vs. AD comparison, the odds ratio was 1.01 (95% *CI*: 0.68-1.51; $p = 0.95$). For the Healthy vs. MCI comparison, the *OR* was 1.16 (95% *CI*: 0.73-1.84; $p = 0.53$), and for the MCI vs. AD comparison, the *OR* = 0.87 (95% *CI*: 0.66-1.16; $p = 0.35$).

Table 1 Baseline demographics and clinical characteristics of cognitively healthy, mild cognitive impairment (MCI), and Alzheimer's disease (AD) cohorts (N=1,500).

Diagnostic Group	Sample Size (n)	Mean Age (SD)	Sex (% Female)	APOE ε4 Carriers (n)	Mean MMSE Score
Healthy	591	68.2 (5.1)	52.4%	145	28.4
MCI	528	71.4 (6.2)	50.1%	132	22.1
AD	381	75.6 (7.1)	55.3%	94	14.2

Table 2 Pairwise comparison of odds ratios for general APOE ε4 carrier status (carriers vs. non-carriers) in clinical diagnostic groups.

Comparison	Odds Ratio	95% CI	*p*-value	Significant?
Healthy vs AD	1.01	0.68-1.51	0.95	No
Healthy vs MCI	1.16	0.73-1.84	0.53	No
MCI vs AD	0.87	0.66-1.16	0.35	No

APOE ε4/ε4 and ε2 Carriage

The *OR* for homozygous APOE ε4/ε4 carriage in the comparison Healthy vs. AD was 0.80 (95% *CI*: 0.49-1.37; $p = 0.40$), not significant. In the same comparison, APOE ε2 carriage had an *OR* of 1.45 (95% *CI*: 0.97-2.17, $p = 0.07$). Although the odds ratio was > 1.0, this association was not statistically significant and is considered to be an exploratory result only (Table 3).

Table 3 Odds ratios for exploratory analysis and comparison of homozygous APOE ε4/ε4 and general ε2 alleles in healthy subjects versus patients with Alzheimer's disease.

Comparison	Odds Ratio	95% CI	p-value	Significant?
APOE ε4/ε4 (Healthy vs AD)	0.80	0.49-1.37	0.40	No
APOE ε2 carriage (Healthy vs AD)	1.45	0.97-2.17	0.07	No

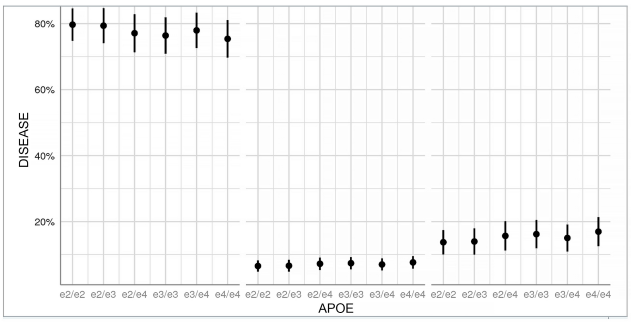

Figure. 4 Probability of Alzheimer's disease according to APOE genotype groups. Circles represent the predicted probability for each category with a 95% confidence interval using the cumulative link mixed model (age, sex, APOE genotype, MMSE; country).

Age and Cognitive Impairment

Individuals aged 65 years and older had a significantly increased risk for MCI compared to cognitively healthy status (*OR* = 3.18; 95% *CI*: 1.24-8.58; $p = 0.017$), indicating that those over 65 years had about 3 times the odds of MCI relative to healthy controls. The AD vs. Healthy comparisons between those over 65 years of age showed an odds ratio greater than 1.0 (*OR* = 1.54; 95% *CI*: 0.75-2.89; $p = 0.22$), but this was not statistically significant. For subjects younger than 65 years, neither comparison showed a statistically significant association (Table 4).

Table 4 Associations between age group and disease status.

Comparison	Age Group	Odds Ratio	95% CI	*p*-value	Significant?
AD vs Healthy	Over 65	1.54	0.75-2.89	0.22	No
AD vs Healthy	Under 65	0.72	0.41-1.40	0.32	No
MCI vs Healthy	Over 65	3.18	1.24-8.58	0.017	Yes
MCI vs Healthy	Under 65	0.57	0.27-1.25	0.16	No

Sex Comparisons

Difference in APOE genotypes in relation to sex among male and female subjects was analyzed. There was no significant association between sex and APOE ε4 carriage (*OR* = 1.16, 95% *CI*: 0.95-1.42, $p = 0.14$), APOE ε4/ε4 genotype (*OR* = 1.28, 95% *CI*: 0.97-1.69, $p = 0.08$), APOE ε3/ε3 genotype (*OR* = 0.82, 95% *CI*: 0.62-1.08, $p = 0.17$), or APOE ε2/ε2 genotype (*OR* = 1.08, 95% *CI*: 0.84-1.40, $p = 0.54$). These results demonstrate that the APOE genotypes are not significantly different with regard to sex.

Table 5 Sex-based comparisons of APOE genotype frequencies.

Comparison	Odds Ratio	95% CI	*p*-value
Female vs Male (APOE ε4)	1.16	0.95-1.42	0.14
Female vs Male (APOE ε4/ε4)	1.28	0.97-1.69	0.08
Female vs Male (APOE ε3/ε3)	0.82	0.62-1.08	0.17
Female vs Male (APOE ε2/ε2)	1.08	0.84-1.40	0.54

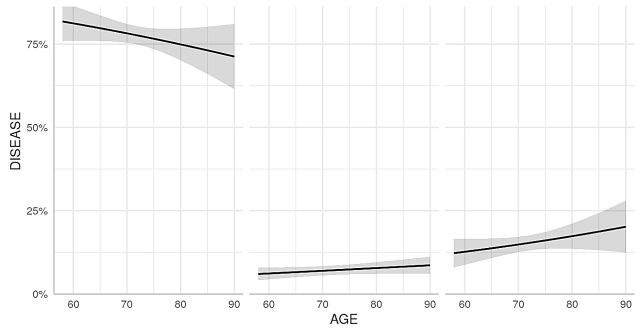

Figure. 5 Participant age and estimated prevalence of Alzheimer's disease and MCI. The x-axis is the age of participants, and the y-axis is the proportion of people with that diagnosis. Shaded bands represent the 95% confidence interval of the smoother applied to the cross-sectional synthetic sample with locally estimated scatterplot smoothing (LOESS) in ggplot2. The curves are descriptive, not fitted disease-progression trajectories in time.

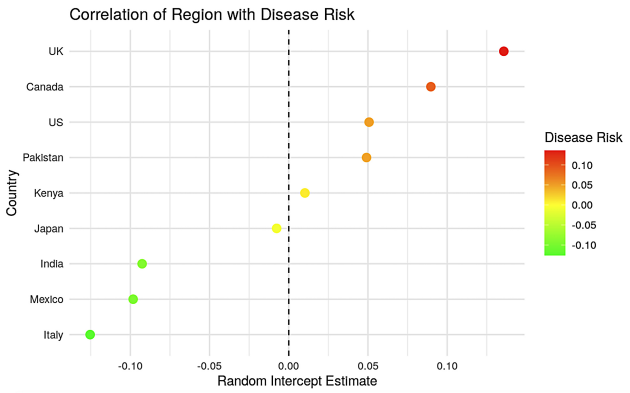

Figure. 6 Cumulative link mixed model country-level random intercept estimates. The x-axis indicates the country-specific random intercept relative to the overall mean (0). Higher values indicate a greater model-estimated propensity for more severe categories of disease after accounting for fixed effects. These estimates reflect between-country variation in the synthetic dataset and should not be interpreted as true geographic prevalence at the population level. The country-specific random intercepts were obtained from the cumulative link mixed model after adjustment for age, sex, APOE genotype, and MMSE score.

MMSE Group Comparisons

Distributions of APOE genotypes were also compared among five groups according to the levels of MMSE score (25-30, 20-24, 15-19, 10-14 and <10). No subgroups showed any significant relationship (Table 6). These results are considered to be exploratory.

Regional Variation in Disease Risk

Exploratory visual analysis and cumulative link mixed modeling (CLMM) suggested modest variation in estimated disease risk across geographic regions (Figure 6). In the CLMM, country was included as a random effect while age, sex, APOE genotype, and MMSE score were included as fixed effects. The random intercept estimates were positive for the United Kingdom, Canada, and the United States and negative for Italy, Mexico, and India. Nevertheless, considering the small number of subgroups in each region and also the fact that the ADDI Synthetic Dataset is not created for simulating the actual geographic prevalence of the disease, the trends observed in these graphs must be cautiously analyzed. Moreover, the observed regional variation should not be interpreted as evidence for environmental risk factors, as the dataset does not contain direct environmental or lifestyle variables but may be a consequence of the synthetic data generation process.

Discussion

The analysis of the generated data resulted in a very unexpected outcome; namely, the carriage of the APOE $\varepsilon 4$ allele had no statistically significant correlation with Alzheimer's disease diagnosis (OR = 1.01; 95% CI: 0.68-1.51; p = 0.95). Compared to years of research proving that the $\varepsilon 4$ allele is one of the major risk factors for developing late onset Alzheimer's disease [13]. Research has shown that the $\varepsilon 4$ allele hampers the brain's ability to clear lipids, accelerates the formation of amyloid-beta plaques, and encourages neuroinflammation [12,16]. Consequently, these results should not be interpreted as a reflection of real-world biology. On the contrary, the null hypothesis highlights the shortcomings of the process of generating the synthetic data and the fact that the model does not represent the existing association between genetics and disease outcomes [25].

The $\varepsilon 4$ frequency was similar across the healthy, MCI, and AD cohorts, showing that the synthetic dataset did not reproduce the $\varepsilon 4$ enrichment observed in real AD populations, where the $\varepsilon 4$ frequency is significantly increased in AD patients. Lack of association $\varepsilon 4$ in the simulated data is an apparent mistake as far as actual medical records are concerned.

Such limitations are also seen with the APOE $\varepsilon 2$ allele. The $\varepsilon 2$ variant has been demonstrated in real-world cases to have a protective effect against the development of cortical atrophy and tau pathology [15]. Pairwise analysis on this synthetic dataset shows a non-significant association suggesting increased disease risk (OR = 1.45; 95% CI: 0.97-2.17; p = 0.07). This difference may be because of statistical noise or the synthetic data generating data for low frequency alleles. This is why synthetic models must be confirmed in the real

Table 6 MMSE genotype subgroup analyses: Pairwise statistical distribution and exploratory p-values.

MMSE Category	Genotypes Tested	Sample Size (n)	Chi-Square / Fisher p-value	Overall Pattern
25-30	$\varepsilon4$, $\varepsilon4/\varepsilon4$, $\varepsilon3/\varepsilon3$, $\varepsilon2/\varepsilon2$	106	$p = 0.48$	Non-Significant
20-24	$\varepsilon4$, $\varepsilon4/\varepsilon4$, $\varepsilon3/\varepsilon3$, $\varepsilon2/\varepsilon2$	282	$p = 0.62$	Non-Significant
15-19	$\varepsilon4$, $\varepsilon4/\varepsilon4$, $\varepsilon3/\varepsilon3$, $\varepsilon2/\varepsilon2$	261	$p = 0.31$	Non-Significant
10-14	$\varepsilon4$, $\varepsilon4/\varepsilon4$, $\varepsilon3/\varepsilon3$, $\varepsilon2/\varepsilon2$	294	$p = 0.75$	Non-Significant
<10	$\varepsilon4$, $\varepsilon4/\varepsilon4$, $\varepsilon3/\varepsilon3$, $\varepsilon2/\varepsilon2$	557	$p = 0.89$	Non-Significant

world before being used to guide clinical decisions[24].

On the other hand, age assessment was verified by the existing data. Strong association with mild cognitive impairment compared to the healthy control group in those >65 years (OR = 3.18; 95% CI: 1.24-8.58; p = 0.017). This is in line with clinical data that show that after the age of 65, metabolic changes increase cognitive vulnerability[18].

The strong association of the MMSE with diagnosis status supports the hypothesis about consistency between cognition score and severity of the condition. Comparing these scores only verifies how the dataset was constructed. Because MMSE scores were used to build the synthetic diagnoses, this doesn't actually prove real-world accuracy.

The results of the study show how complicated neurodegenerative diseases are. In the modern medicine field, genetics alone cannot be relied upon. The use of the polygenic risk score combined with tracking of the patient's lifestyle is needed[31]. This study was confined to basic demographics since the ADDI dataset lacks other important factors such as diet, physical activity, and pollution.

Moving forward, only a joint investigation of the genetic data and actual environment will be applicable. The use of statistical models and analysis in large-scale real-life patient populations is necessary for proper identification of genetic and demographic risks and formulation of preventive measures.

Limitations

These limitations are important to keep in mind when interpreting our results. The first limitation of the study is the use of a synthetic dataset, which results in uncertainty as to whether the relationship observed is characteristic of the actual population of patients with AD. Synthetic data may not replicate the distributional characteristics of clinical cohorts. Null findings should not be interpreted as evidence against well-established associations reported in the literature. The second limitation is that the primary analyses were confined to pairwise contingency tables, while an exploratory cumulative link mixed model was applied as well. The third one is that only genetic and demographic factors could be studied, as environmental/lifestyle factors were not measured. Fourth, the sample size within individual subgroups, and in particular for MMSE-stratified analyses, was too small to detect small effect sizes. Null findings in those analyses should be viewed as inconclusive and not negative.

Conclusion

The relationships among APOE genotype, age, MMSE score, and geographical region in relation to AD diagnostic status were analyzed using the ADDI Synthetic Dataset. Being older than 65 was found to be significantly related to MCI as compared to cognitively healthy status; on the other hand, no significant relationship emerged from any pairwise analysis between APOE $\varepsilon4$ carrier and other variables. This indicates that Alzheimer's disease is a multifactorial disorder; moreover, analyzing synthetic data in terms of complex gene relationships carries inherent weaknesses. The results presented in this paper can be used as the basis for future studies that will incorporate actual environmental and lifestyle factors together with genetic data.

Future Research Directions

Follow-up studies that would be most beneficial include adding real-world data regarding environmental exposures and lifestyle choices such as dietary quality, physical activity, pollution, cardiovascular history, educational attainment, and healthcare access. When paired with APOE genotype data in a larger study, this would allow testing of hypotheses as opposed to generating them. Another follow-up method that could prove useful in early prediction of disease progression would be biomarker data, brain imaging, and possibly machine learning algorithms.

References

1 Alzheimer's Association, *2024 Alzheimer's disease facts and figures*, 2024, 10.1002/alz.13809.

2 C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein and N. Silverberg, *NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease*, 2018, 10.1016/j.jalz.2018.02.018.

3 World Health Organization, *Global status report on the public health response to dementia*, 2021, https://www.who.int/publications/i/item/9789240033245.

4 GBD 2019 Dementia Forecasting Collaborators, *Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050*, 2022, 10.1016/S2468-2667(21)00249-8.

5 B. Winblad *et al.*, *Defeating Alzheimer's disease and other dementias: a priority for European science and society*, 2016, 10.1016/S1474-4422(16)00062-4.

6 G. Livingston *et al.*, *Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission*, 2024, 10.1016/S0140-6736(24)01296-0.

7 C. Bellenguez, F. Küçükali, I. E. Jansen, V. Andrade, S. Moreno-Grau, N. Amin and P. Amouyel, *New insights into the genetic etiology of Alzheimer's disease and related dementias*, 2022, 10.1038/s41588-022-01024-z.

8 J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez and P. Amouyel, *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease*, 2013, 10.1038/ng.2802.

9 B. W. Kunkle, B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj and M. A. Pericak-Vance, *Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing*, 2019, 10.1038/s41588-019-0358-2.

10 I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg and D. Posthuma, *Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk*, 2019, 10.1038/s41588-018-0311-9.

11 E. H. Corder *et al.*, *Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families*, 1993, 10.1126/science.8346443.

12 P. B. Verghese, J. M. Castellano and D. M. Holtzman, *Apolipoprotein E in Alzheimer's disease and other neurological disorders*, 2011, 10.1016/S1474-4422(10)70325-2.

13 W. J. Strittmatter *et al.*, *Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease*, 1993, 10.1073/pnas.90.5.1977.

14 C. G. Fernandez, M. E. Hamby, M. L. McReynolds and W. J. Ray, *The role of APOE4 in disrupting the homeostatic functions of astrocytes and microglia in aging and Alzheimer's disease*, 2019, 10.3389/fnagi.2019.00014.

15 E. M. Reiman *et al.*, *Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5,000-person neuropathological study*, 2020, 10.1038/s41467-019-14279-8.

16 C. C. Liu, T. Kanekiyo, H. Xu and G. Bu, *Apolipoprotein E and Alzheimer's disease: risk, mechanisms and therapy*, 2013, 10.1038/nrneurol.2012.263.

17 L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux and R. Duara, *Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis from the APOE e4 Joint Study*, 1997, https://jamanetwork.com/journals/jamapsychiatry/fullarticle/209307.

18 A. J. Mitchell and M. Shiri-Feshki, *Rate of progression of mild cognitive impairment to dementia — meta-analysis*, 2009, 10.1111/j.1600-0447.2008.01326.x.

19 A. Serrano-Pozo, S. Das and B. T. Hyman, *APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapies*, 2021, 10.1016/S1474-4422(20)30412-9.

20 D. E. Barnes and K. Yaffe, *The projected effect of risk factor reduction on Alzheimer's disease prevalence*, 2011, 10.1016/S1474-4422(11)70072-2.

21 T. Ngandu, J. Lehtisalo, A. Solomon, E. Levälahti, S. Ahtiluoto, R. Antikainen and M. Kivipelto, *A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial*, 2015, https://pubmed.ncbi.nlm.nih.gov/25771249/.

22 S. Licher, S. Ahmad, H. Karamujić-Čomić, T. Voortman, J. J. G. Leening, M. A. Ikram and M. K. Ikram, *Genetic predisposition, modifiable risk factors, and long-term risk of dementia*, 2019, https://journals.sagepub.com/url/10.3233/JAD-180631.

23 K. B. Rajan *et al.*, *Population-attributable fractions of risk factors for Alzheimer disease in the United States*, 2021, 10.1002/alz.12362.

24 L. Gootjes-Dreesbach *et al.*, *Variational autoencoder modelling of in silico (virtual) patient cohorts in dementia*, 2020, 10.3389/frai.2020.613956.

25 R. J. Chen *et al.*, *Synthetic data in machine learning for medicine and healthcare*, 2021, 10.1038/s41551-021-00751-8.

26 BrightFocus Foundation, *Alzheimer's disease shrinks the brain [Image]*, 2026, https://www.brightfocus.org/alzheimers/.

27 R. Perneczky, S. Wagenpfeil, K. Komossa, T. Grimmer, J. Diehl and A. Kurz, *Mapping scores onto stages: Mini-Mental State Examination and Clinical Dementia Rating*, 2006, 10.1097/01.JGP.0000192478.82189.a8.

28 R Core Team, *R: A language and environment for statistical computing*, 2023, https://www.R-project.org/.

29 R. H. B. Christensen, *ordinal—Regression Models for Ordinal Data*, 2023, R package.

30 Alzheimer's Disease Data Initiative (ADDI), *AD Workbench Synthetic Dataset*, 2024, https://www.alzheimersdata.org.

31 D. P. Wightman, I. E. Jansen, J. E. Savage, A. A. Shadrin, S. Bahrami, D. Holland and D. Posthuma, *A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease*, 2021, 10.1038/s41588-021-00921-z.