

Integrative clinical–genomic modeling evaluates exploratory Non-APOE candidate signals in Alzheimer’s disease

Eun Kyung Paik

Hanyang University

Jaechul Lee

Kyung Hee University

Eek-Sung Lee

Soonchunhyang University Bucheon Hospital

Jong Kuk Park

Korea Institute of Radiological and Medical Sciences

Jie-Young Song

Korea Institute of Radiological and Medical Sciences

SiYeon Kim

Hanyang University

Jong-Young Lee

Oneomics

Insong Koh

Hanyang University

Woojin Kim

Kyung Hee University

Youngchan Park

`ycpark1228@kiram.s.re.kr`

Korea Institute of Radiological and Medical Sciences

Research Article

Keywords: Alzheimer’s disease, amnesic mild cognitive impairment, APOE, whole-exome sequencing, whole-genome sequencing, XGBoost, clinical–genomic modeling, out-of-fold validation

Posted Date: August 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9617891/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Alzheimer's Research & Therapy on October 16th, 2026. See the published version at <https://doi.org/10.1186/s13195-026-02231-0>.

Integrative clinical–genomic modeling evaluates experimental Non-APOE candidate signals in Alzheimer’s disease

Eun Kyung Paik^{1,2,†}, Jaechul-Lee^{3,†}, Eek-Sung Lee⁶, Jong Kuk Park², Jie-Young Song², SiYeon Kim¹, Jong-Young Lee⁴, Insong Koh^{1,*}, Woojin Kim^{3,*}, Youngchan Park^{2,5,*}

1 Department of Biomedical Informatics, Graduate School of Biomedical Science and Engineering, Hanyang University, Seoul 04763, Republic of Korea.

2 Division of Radiation Biomedical Research, Korea Institute of Radiological and Medical Sciences, Seoul 01812, Republic of Korea.

3 Department of Physiology, College of Korean Medicine, Kyung Hee University, Seoul, Republic of Korea.

4 OneOmics Co., Ltd., Gyeonggi-do 14585, Republic of Korea

5 Division of Bio Bigdata, Department of Precision Medicine, Korea National Institute of Health, Korea Disease Control and Prevention Agency, Cheongju, Republic of Korea

6 Soonchunhyang University Bucheon Hospital, Bucheon, Republic of Korea

Present address of Youngchan Park : Division of Radiation Biomedical Research, Korea Institute of Radiological and Medical Sciences, Seoul 01812, Republic of Korea.

† These authors contributed equally to this work.

*Corresponding authors: Insong Koh, insong@hanyang.ac.kr; Woojin Kim,
wojin@khu.ac.kr; Youngchan Park, ycpark1228@kiram.s.re.kr

Abstract

Background

Alzheimer's disease (AD) is characterized by a genetic architecture dominated by *APOE*, while predictive modeling in clinically enriched cohorts is often driven by neuropsychological and clinical variables. This combination limits the interpretability of secondary genetic signals in multimodal frameworks.

Methods

We applied a two-stage Korean WES–WGS design across the Alzheimer's continuum. Candidate loci were nominated in a whole-exome sequencing (WES) amnesic mild cognitive impairment (aMCI) cohort (n=346) and evaluated in an independent whole-genome sequencing (WGS) cohort (n=995) with epidemiologic and neuropsychological variables. Gradient-boosted decision trees (XGBoost) were evaluated using a common stratified five-fold cross-validation scheme with pooled out-of-fold (OOF) predictions. Preprocessing and one-hot encoding were fitted within each training fold only. Identifier/linkage fields were excluded before modeling, and a redundant categorical age-band variable was removed. Feature attribution and APOE sensitivity analyses were used for descriptive model interpretation.

Results

The full clinical–genomic model showed near-perfect internal discrimination (AUROC 0.998; 95% CI 0.996–0.999) but included diagnostic-proximal clinical and neuropsychological variables and was therefore interpreted as an upper-bound clinical separability model rather than evidence of prospective predictive utility. In the strict leakage-reduced sensitivity analysis, AUROC decreased to 0.713 (95% CI 0.682–0.745) and decreased further to 0.625 (95% CI 0.589–0.658) after removal of rs429358 and APOE proxy information. The five-SNP model had modest discrimination (AUROC 0.637; 95% CI 0.605–0.670), whereas the Non-APOE SNP-only model was weak (AUROC 0.533; 95% CI 0.497–0.568). In fold-wise validation permutation analysis of the strict model, Test_Age ranked first overall and rs429358 was the highest-ranking genetic feature; the Non-APOE candidates had small or near-zero mean AUROC decreases after permutation.

Conclusions

This study presents a Korean WES–WGS exploratory clinical–genomic candidate-evaluation framework. The findings indicate that high multimodal discrimination in clinically enriched AD datasets can reflect diagnostic-proximal clinical separability, while Non-APOE candidate variants should be interpreted as hypothesis-generating model features rather than confirmed susceptibility loci, independent predictors, or causal variants. A second independent Korean cohort with harmonized candidate genotypes and comparable clinical and neuropsychological variables was not available within the present study, precluding direct external replication, and the WES and WGS stages used differed

nt clinical endpoints.

Keywords: Alzheimer’s disease; amnesic mild cognitive impairment; APOE; whole-exome sequencing; whole-genome sequencing; XGBoost; clinical-genomic modeling; out-of-fold validation.

Background

Alzheimer’s disease (AD) is a complex neurodegenerative disorder with a heterogeneous genetic architecture. Genome-wide association studies have identified multiple susceptibility loci; however, most variants have modest effect sizes, and their contributions are often overshadowed by the dominant influence of *APOE* [1, 2]. *APOE* ϵ 4 remains the strongest common genetic risk factor for late-onset AD and can obscure secondary genetic effects in both association and predictive modeling contexts [3].

In clinically characterized cohorts, the challenge is further compounded by the presence of neuropsychological and biomarker variables that are proximal to diagnostic labeling. These features frequently dominate predictive performance, making it difficult to interpret weaker genomic contributions. As a result, high classification accuracy in multimodal models does not necessarily reflect independent genetic signal strength.

Recent advances in machine learning provide tools for integrating heterogeneous data and for interpreting feature contributions within complex models [4, 5]. However, genome-wide modeling in moderately sized cohorts remains underpowered and prone to overfitting, particularly when interpretability is required. Under these conditions, candidate-based approaches combined with interpretable modeling offer a practical alternative.

In addition, emerging evidence indicates that genetic effects related to cognition may extend across neurodegenerative contexts rather than mapping exclusively to a single diagnostic category [6, 7]. This supports an exploratory framework in which candidate variants are evaluated as potential contributors to cognitive vulnerability, while avoiding disease-specific causal claims without external replication. However, most existing studies do not explicitly integrate functional genomic annotations when interpreting genetic contributions within clinically dominant predictive models.

In this study, we used a two-stage Korean WES–WGS design across the Alzheimer’s continuum. Candidate loci were nominated in a WES-based aMCI cohort and evaluated in an independent Korean WGS cohort from the Korea National Institute of Health (KNIH), which included detailed epidemiologic and neuropsychological variables [8]. We applied interpretable machine learning and *APOE* sensitivity analyses to assess whether exploratory Non-APOE candidate features retained model reliance after reducing APOE-driven attribution. Our objective was not to establish novel AD susceptibility loci, but to evaluate how WES-nominated exploratory candidate variants behave within clinically integrated WGS modeling frameworks and to separate model reliance from genetic associati

on or causal interpretation.

Methods

Study design and cohorts

We used a two-stage design across the Alzheimer's continuum (Figure 1). Candidate loci were nominated in a Korean WES case-control cohort comparing aMCI and cognitively normal controls. Candidate loci were then evaluated in an independent Korean WGS cohort, which includes Alzheimer's diagnosis labels and detailed epidemiologic and neuropsychological variables. All genomic coordinates and variant extraction were performed using GRCh38.

WES candidate-nomination cohort

The WES candidate-nomination cohort included 346 Korean participants, comprising 263 individuals with aMCI and 83 cognitively normal controls, recruited at Soonchunhyang University Bucheon Hospital (January 2020–July 2021). Participants with documented family relationships were excluded based on internal recruitment and clinical records. Therefore, WES candidate-variant analyses were conducted under an unrelated-participant framework and interpreted as exploratory screening rather than family-based association testing. Demographic and cognitive characteristics are summarized

d (Table 1). Written informed consent was obtained, and the study was approved by the institutional review board in accordance with the Declaration of Helsinki.

Independent WGS evaluation cohort

The WGS Alzheimer cohort was obtained from the Korea National Institute of Health (KNIH) Clinical & Omics Data Archive (CODA). In the GARD clinical metadata, Diagnose=1 corresponded to clinically diagnosed Alzheimer's disease and Diagnose=0 to control status. It included 995 participants with epidemiologic and neuropsychological variables and a binary phenotype variable, Diagnose (0=control, 1=Alzheimer's disease). For the WGS clinical-genomic evaluation cohort, participants with documented family relationships were excluded based on the clinical epidemiologic registry. Accordingly, WGS analyses were conducted under an unrelated-participant framework. Because this cohort was used as an independent clinical-genomic evaluation cohort rather than a family-based association cohort, no additional family-based modeling was applied.

Candidate-based strategy and claim boundaries

Given the modest sample size of the WES candidate-nomination cohort relative to the exome-wide variant space, this study did not train a genome-wide machine-learning model for de novo variant discovery. Instead, WES screening was used to nominate a small set of exploratory candidate va

riants, which were then evaluated in the independent WGS cohort within predefined clinical-genomic modeling frameworks. This design was intended for hypothesis generation and model-based candidate evaluation rather than genome-wide association discovery, causal inference, or confirmation of AD susceptibility loci. The rationale and claim boundaries for each candidate variant are summarized in Supplementary Table S12. Sample sizes were determined by the available cohorts rather than by an a priori power calculation. Given the modest WES cohort size relative to the exome-wide variant space, WES findings were therefore treated as exploratory candidate nominations and not as definitive genetic association evidence.

WES processing and association testing

We performed sequencing and variant calling on GRCh38 (BWA-MEM (BWA RRID:SCR_010910) and GATK 4.1.0.0 (RRID:SCR_001876)) [9, 10]. Candidate-level PLINK analyses were conducted using PLINK (RRID:SCR_001757) and the available PLINK-formatted genotype dataset, in which all five variants carried forward for WGS evaluation were present. The unfiltered PLINK BIM file contained 142,740 variant entries. After genotype-level quality control using --mind 0.10, --geno 0.10, --maf 0.01, and --hwe 1e-6, the QC-filtered candidate-level analytic dataset included 332 samples and 53,858 variants. All five carried-forward candidate variants passed these QC thresholds. Across these five variants, minor allele frequencies ranged from 0.0661 to 0.4189, Hardy-Weinberg equilibrium P values ranged from 0.0644 to 0.8937, and genotype call rates ranged from 97.08% to 97.37%. A summ

ary of candidate screening and PLINK genotype quality-control metrics is provided in Supplementary Table S13.

Separately, the annotated WES screening table used for candidate review contained 5,215 unique variant records, including 4,753 variants annotated as PASS. The annotated screening table and the PLINK genotype matrix represented different downstream data objects. The screening table was used for exploratory candidate nomination and was not treated as a formal exome-wide association scan. Candidate selection was therefore interpreted as exploratory nomination rather than genome-wide discovery, and no candidate variant was interpreted as exome-wide significant, independently associated, causal, or confirmed as an Alzheimer's disease susceptibility locus.

Additive models were prespecified as the primary candidate-level comparison, whereas dominant, recessive, and genotypic models were reported as sensitivity analyses. Because the WES candidate-nomination cohort showed an age difference between aMCI cases and cognitively normal controls, candidate-variant associations were additionally evaluated using covariate-adjusted logistic regression. For Non-APOE candidate variants, the primary adjusted model included age, sex, and *APOE* ϵ 4 carrier status. APOE-region variants were analyzed separately using age and sex adjustment, because *APOE* ϵ 4 carrier status is partly determined by *APOE* genotype and may introduce collinearity. For the candidate-panel analysis, Benjamini–Hochberg FDR- and Bonferroni-adjusted P values were calculated across the five primary additive candidate-variant tests.

The five carried-forward SNPs were *PEMT* rs7946, *XRCC1* rs25489, *APOE* rs440446, *APOE* rs429358,

and *NKPD1* rs28469095. External reference-projected ancestry principal components were not available for the WES cohort; therefore, residual population structure could not be fully excluded and was acknowledged as a limitation. As an additional internal sensitivity analysis, genotype-derived principal components were calculated from QC-filtered, LD-pruned autosomal WES variants after excluding the APOE-region and candidate loci [11, 12]. PCA was performed using 26,479 LD-pruned variants with a total genotyping rate of 0.9957. The first five internal WES principal components were then added to candidate-level logistic regression models. For Non-APOE candidate variants, the PC-adjusted models included SNP dosage, age, sex, *APOE* ϵ 4 carrier status, and PC1–PC5. APOE-region variants were analyzed using SNP dosage, age, sex, and PC1–PC5 to avoid collinearity with *APOE* ϵ 4 carrier status. These analyses were interpreted as internal population-structure sensitivity analyses rather than external ancestry validation.

Functional annotation of candidate variants

To provide biological context for exploratory variants, we performed post hoc in silico annotation using public functional genomics and clinical genetics resources. Variant consequences and genomic coordinates were obtained using GRCh38-based Ensembl Variant Effect Predictor annotation (RRID:SCR_007931) [13]. Gene expression and single-tissue eQTL evidence were queried using GTEx resources (GTEx project RRID:SCR_013042; GTEx eQTL Browser RRID:SCR_001618), focusing on brain tissues and whole blood [14]. Clinical interpretation was assessed using ClinVar (RRID:SCR_006169)

[15], and external AD genetic context was assessed by comparison with large-scale AD GWAS findings and public variant-to-gene resources [16, 17]. These annotations were not used in model training, association testing, or candidate selection and were incorporated only as post hoc biological context.

Genotype extraction and feature construction

We extracted genotypes at the five target loci from WGS-derived site VCFs and encoded genotypes as alternate-allele dosage: 0/0→0, 0/1 or 1/0→1, 1/1→2. Dosage features were merged with the epidemiologic/neuropsychological table by subject ID. ID and FAMILY_ID were linkage identifiers in the analytic export and were excluded before model fitting; PNTTM was also excluded as a non-predictive procedural field. Age denotes age at Alzheimer’s disease diagnosis, whereas Test_Age denotes age at neuropsychological testing. Ages is a categorical five-year age-band variable derived from Test_Age and was excluded from all model configurations to avoid redundant age encoding. Clinical variables were processed using the codebook mapping, and sentinel values (e.g., 66666/77777/99999) were treated as missing before modeling [11, 12, 18]. Missing values were not imputed and were handled natively by XGBoost, which allows sparse-aware split finding.

Machine learning and evaluation

We modeled the binary Diagnose phenotype in the Alzheimer WGS cohort (0 = control, 1 = Alzheimer's disease) using gradient-boosted decision trees (XGBoost). XGBoost fits an ensemble of shallow trees sequentially, where each tree reduces the residual error of the current ensemble [4, 19].

$$F(x) = \sum_{k=1}^K f_k(x), \hat{p} = \sigma(F(x))$$

(where f_k is the k -th tree and σ is the logistic function)

The framework can capture non-linear relationships and higher-order interactions between epidemiologic/neuropsychological variables and genotype dosage features without specifying parametric interaction terms. Clinical variables were taken from codebook-guided preprocessing, and genotype features were encoded as alternate-allele dosage (0/1/2) from WGS site-level VCFs for the candidate panel (rs7946, rs25489, rs440446, rs429358, rs28469095). Primary WGS performance evaluation used stratified five-fold cross-validation across the full WGS evaluation cohort to generate one OOF prediction for each participant. The same fold assignments were used for all prespecified model configurations. Within every fold, preprocessing and one-hot encoding were fitted only on the four training folds and then applied to the corresponding validation fold. The XGBoost configuration was fixed before OOF evaluation and was not optimized using validation-fold outcomes. Pooled OOF predictions were used to calculate AUROC, AUPRC, accuracy, balanced accuracy, sensitivity, specificity, Brier score, and confusion matrices. Classification metrics used a prespecified probability threshold of 0.50. AUROC and AUPRC 95% confidence intervals were estimated from the pooled OOF

F predictions using 2,000 class-stratified nonparametric bootstrap resamples and percentile 2.5th and 97.5th limits. These estimates represent internal cross-validated performance and are not external or prospective validation.

To address potential diagnostic reconstruction, we additionally built a strict leakage-reduced model. This model excluded amyloid PET, CDR and CDR-SB-related variables, MMSE total score, MMSE item-level variables q01–q19, SVLT, RCFT, SNSB domain scores, memory-domain variables, ADL or functional assessment variables, and broader neuropsychological test-derived variables including Stroop, TMT, COWAT, DSC, CDT, digit span, naming, praxis, calculation, and related cognitive test scores. The retained feature set included candidate SNP dosages, demographic variables, education-related variables, sex, job, handgrip, and *APOE* proxy information when applicable. The categorical age-band variable Ages was excluded as redundant. We also evaluated a stricter *APOE* sensitivity model removing rs429358 and *APOE* proxy information, as well as a Non-*APOE* SNP-only model.

***APOE* sensitivity and interpretation**

Because *APOE* rs429358 is a major *APOE*-region marker indexing *APOE* ϵ 4-related Alzheimer's disease risk [3, 20], we refit sensitivity models excluding rs429358 and reassessed feature attribution among the remaining clinical and candidate genetic variables. These sensitivity analyses were designed

ned to evaluate whether Non-APOE candidate variants retained descriptive model reliance after reducing APOE-driven attribution [21]. No variant-specific interaction terms were included in the primary models. Feature attribution results from these analyses were interpreted as descriptive model-reliance measures and not as evidence of independent genetic association, causality, or biological mechanism [5, 21].

Linkage disequilibrium assessment

To evaluate whether candidate signals near the *APOE* locus reflect independent effects or linkage disequilibrium (LD) tagging, we estimated pairwise LD among the candidate variants and, when applicable, an expanded AD GWAS SNP panel within the WGS Alzheimer cohort. Genotypes were encoded as alternate-allele dosage (0, 1, 2) per variant and sample [11, 18]. We computed r^2 as the primary LD metric because it directly reflects correlation between alleles and is most interpretable for tagging in association/prediction contexts [22, 23]. Pairwise r^2 was calculated using cohort-derived dosages after excluding samples with missing genotypes for the corresponding SNP pair (pairwise complete observations).

$$r^2 = \left(\text{corr}(g_a, g_b) \right)^2$$

Here, g_a and g_b denote genotype-dosage vectors for SNPs a and b , respectively. For reporting near APOE, we focused on key pairs involving rs429358 and rs7412, as well as nearby candidates su

ch as *NKPD1* rs28469095. LD estimates from external reference resources, including LDlink/LDpair, were used only as supportive context, whereas the main interpretation relied on cohort-derived LD estimates [24].

Feature attribution

To quantify feature contributions within multivariable prediction models, we used three complementary attribution measures. Gain, defined as XGBoost “gain,” summarizes the improvement in the training objective attributable to splits on a feature across all trees; relative gain was calculated as feature-level gain normalized by total model gain [4].

$$\text{rel_gain}_j = \frac{\text{gain}_j}{\sum_m \text{gain}_m}$$

Permutation importance is computed within each outer validation fold as the decrease in AUROC after random permutation of one raw input feature while all other features remain fixed [21, 25]. The preprocessing pipeline and XGBoost model used for each permutation analysis are fitted only on the corresponding training folds. For the strict leakage-reduced model, validation-fold permutation estimates are aggregated across the five folds and repeated permutations to summarize mean and standard deviation of AUROC decrease. This analysis is interpreted as out-of-sample model reliability under the observed cohort distribution, not as a genetic association effect estimate [21].

$$\Delta AUC_j = AUC_{\text{orig}} - AUC_{\text{perm}(j)}$$

SHAP values were estimated using TreeExplainer from SHAP (RRID:SCR_021362) and summarized as mean absolute SHAP values per feature, with relative SHAP values reported only for comparison across model configurations [5]. Feature attribution analyses were interpreted as descriptive model-reliance measures rather than evidence of genetic association, causality, or biological mechanism. Genetic association evidence was evaluated separately using WES logistic regression, whereas biological relevance was assessed only as post hoc functional context. Because relative gain and relative SHAP values can be affected by the total attribution denominator, especially after exclusion of dominant features such as *APOE* rs429358, these measures were not interpreted as effect sizes or as evidence of independent predictive utility [5, 21].

$$I_j = \frac{\mathbb{E}(|\phi_j|)}{\sum_m \mathbb{E}(|\phi_m|)}$$

Model configurations are summarized in Supplementary Table S4. The raw feature inventory and explicit exclusions are provided in Supplementary Table S5. Interaction-containing exploratory model configurations were not part of the prespecified primary analyses and were not used for candidate nomination, association inference, biological interpretation, or the primary reported results. They are presented only as supplementary descriptive analyses of feature-ranking stability (Table S9). Analyses were performed using Python 3.13.9 (RRID:SCR_008394), XGBoost 3.2.0 (RRID:SCR_021361), scikit-learn 1.8.0 (RRID:SCR_002577), NumPy 2.2.6 (RRID:SCR_008633), and pandas 2.3.3 (RRID:S

CR_018214), with a fixed random seed 42. The same fixed XGBoost configuration was applied across OOF folds and prespecified model configurations. As a separate sensitivity analysis distinguishing candidate-level association from model attribution in the WGS evaluation cohort, we performed covariate-adjusted logistic regression for the five carried-forward variants. For Non-APOE candidate variants, models included SNP dosage, Age (age at diagnosis), sex, and APOE rs429358 dosage. APOE-region variants were analyzed using SNP dosage, Age (age at diagnosis), and sex. Benjamini–Hochberg FDR and Bonferroni-adjusted P values were calculated across the five WGS candidate-variant tests [26]. These analyses were interpreted as exploratory candidate-level sensitivity analyses and not as validation of genetic association, causality, or clinical utility.

Results

WES screening nominates exploratory candidate variants for downstream evaluation

The WES cohort showed clinical differences between aMCI cases and cognitively normal controls, including older age, lower cognitive performance, and higher *APOE* ϵ 4 carrier frequency in the aMCI group. Several loci showed nominal evidence of association in the initial WES screening and were carried forward for exploratory evaluation in the independent WGS clinical–genomic evaluation cohort (Table 2). Secondary inheritance-model sensitivity analyses are provided in Supplementary Tables S1–S3; primary candidate-level inference was based on the age-adjusted additive models with

h candidate-panel multiple-testing correction reported in Supplementary Table S10. In covariate-adjusted sensitivity analyses, *PEMT* rs7946 and *XRCC1* rs25489 were attenuated after adjustment for age, sex, and *APOE* ϵ 4 carrier status. *NKPD1* rs28469095 retained nominal evidence after covariate adjustment but did not survive candidate-panel multiple-testing correction (Supplementary Table S10) [26]. *APOE* rs429358 showed the expected APOE-region positive-control signal [3, 20]. Therefore, the WES candidate variants were interpreted as exploratory signals rather than confirmed AD susceptibility loci [16]. The PLINK genotype dataset contained 142,740 BIM variant entries before filtering and 53,858 variants after standard PLINK quality control. However, the WES stage was not treated as a formal exome-wide discovery analysis; candidate findings were interpreted conservatively and evaluated using candidate-panel correction in Supplementary Table S10. Additional sensitivity analyses including the first five internal WES genotype-derived principal components did not alter the conservative interpretation of the candidate variants. *PEMT* rs7946 and *NKPD1* rs28469095 showed borderline nominal evidence after PC adjustment, whereas *XRCC1* rs25489 was attenuated; however, no candidate variant survived candidate-panel FDR or Bonferroni correction (Supplementary Table S14). These findings further supported interpretation of the WES signals as exploratory candidate nominations rather than corrected susceptibility loci.

WGS cohort enables integrated clinical-genomic modeling

The independent WGS cohort (n=995; 509 AD, 486 controls) enabled integration of genotype, epi

demographic, and neuropsychological variables. Genotypes at the nominated loci were encoded as dosage and merged with clinical features to construct a combined dataset. This stage serves as an integration layer, allowing evaluation of candidate variants within a clinically informed predictive framework rather than as isolated associations. The model configurations and pooled five-fold OOF performance metrics are summarized in Table 3. The feature composition of the combined model is summarized in Table 5.

Clinical variables dominate predictive performance

The full clinical-genomic model showed near-perfect pooled OOF discrimination (AUROC 0.998; 95% CI 0.996–0.999; AUPRC 0.998; 95% CI 0.996–0.999), but this model included diagnostic-proximal variables and was therefore interpreted as an upper-bound clinical separability model rather than evidence of prospective predictive utility. Accuracy was 0.981 at the prespecified 0.50 threshold. The common OOF pipeline excluded identifier fields and the redundant age-band variable before feature construction, and all preprocessing was fitted within training folds only.

To address potential diagnostic reconstruction, we performed a strict leakage-reduced sensitivity analysis excluding amyloid PET, CDR-related variables, MMSE total and item-level variables q01–q19, SVLT, RCFT, broader neuropsychological test scores, memory-domain variables, and functional assessment variables. Under the strict leakage-reduced feature set, AUROC decreased to 0.713 (95%

CI 0.682–0.745; AUPRC 0.723, 95% CI 0.686–0.758). After additional removal of rs429358 and APOE proxy information, AUROC decreased to 0.625 (95% CI 0.589–0.658; AUPRC 0.623, 95% CI 0.589–0.659). The five-SNP model showed modest discrimination (AUROC 0.637; 95% CI 0.605–0.670), whereas the Non-APOE SNP-only model showed weak discrimination (AUROC 0.533; 95% CI 0.497–0.568). These comparisons used identical OOF fold assignments across model configurations.

Feature attribution describes model reliance within clinical-genomic models

In the strict leakage-reduced model, primary attribution was summarized using permutation importance calculated only on outer validation folds (Figure 5; Table 4; full feature-level results in Supplementary Table S11). Test_Age ranked first overall (mean AUROC decrease 0.0863), followed by rs429358 (0.0759), YearsOfEducation (0.0353), APOE (0.0346), and Age (0.0267). Among Non-APOE candidates, rs25489 and rs7946 had small mean decreases (0.00230 and 0.00207, respectively), whereas rs28469095 was near zero (–0.00032). Attribution was interpreted as descriptive model reliance rather than evidence of independent genetic association, biological mechanism, or causality [5, 21]. APOE-region signals were interpreted in the context of established APOE risk, whereas PEMT rs7946, XRCC1 rs25489, and NKPD1 rs28469095 remained exploratory Non-APOE model features only. Covariate-adjusted WGS logistic regression was reported separately to distinguish candidate-level association from model attribution.

367

368 **Functional annotation provides post hoc biological context**

369 To assess whether the exploratory candidate variants had post hoc biological context, we annotate
370 d the five candidate variants using GRCh38-based variant consequences, GTEx single-tissue eQTL r
371 esources, ClinVar annotations, in silico variant-effect predictions, and external Alzheimer's disease g
372 enetic evidence (Tables S8A-S8B). Four candidate variants were consistently represented by missen
373 se consequences in the functional annotation summary (*PEMT* rs7946, *XRCC1* rs25489, *APOE* rs429
374 358, and *NKPD1* rs28469095). *APOE*-region rs440446 showed transcript-dependent VEP consequen
375 ces. Its overall most severe VEP consequence was missense_variant, arising from an alternative AP
376 OE transcript (ENST00000434152; C>G, p.Asn14Lys), whereas the *APOE* MANE Select transcript (EN
377 ST00000252486/NM_000041.4) was annotated as intron_variant; 5' UTR consequences were presen
378 t for other alternative transcripts (Table S8A).

379 *PEMT* rs7946 was located outside the *APOE* region and showed minimal LD with rs429358, suppor
380 ting its interpretation as a Non-*APOE* candidate in this dataset, although this does not establish bi
381 ological or causal independence from *APOE*-related pathways (Figure 4; Table S6; Table S8A). No s
382 ignificant brain-tissue eQTL was detected for rs7946 in the GTEx significant single-tissue eQTL look
383 up performed in this study; therefore, rs7946 was not interpreted as a confirmed regulatory varian
384 t (Table S8B). *PEMT* participates in phosphatidylcholine biosynthesis and phospholipid homeostasi
385 s, which are relevant to membrane lipid metabolism and are biologically adjacent to *APOE*-mediat

ed lipid transport pathways in the central nervous system [27, 28]. Variant-level annotation provide
d post hoc functional context for *PEMT* rs7946, but in silico prediction scores were not interpreted
as evidence of pathogenicity, causality, regulatory function, *APOE* independence, or independent A
lzheimer's disease association. These annotations were used only as post hoc biological context fo
r cautious follow-up consideration within the exploratory candidate framework and did not establi
sh rs7946 as a confirmed regulatory, causal, independently associated, or independently predictive
Alzheimer's disease variant (Table S8B).

XRCC1 rs25489 was annotated as a missense variant in a DNA repair gene (Table S8A). Although D
NA repair and oxidative-stress pathways are broadly relevant to neuronal aging and neurodegener
ation, rs25489 was interpreted as an exploratory candidate because its WES association was attenu
ated after covariate adjustment and its WGS model attribution was limited (Figure 5; Table 4; Tabl
es S8A-S8B). *APOE* rs429358 served as the expected dominant Alzheimer's disease genetic signal a
nd positive-control locus, consistent with the established role of *APOE* ϵ 4 in late-onset Alzheimer's
disease risk [3]. rs440446 was interpreted as an APOE-region marker requiring cautious interpretati
on because of local LD and potential tagging effects, whereas NKPD1 rs28469095 was interpreted
cautiously despite its missense annotation because its WES evidence did not survive multiple-testi
ng correction and its WGS model attribution was near zero (Figure 5; Table 4; Tables S8A-S8B). Co
mparison with large-scale Alzheimer's disease GWAS findings showed that *APOE* is repeatedly esta
blished as the dominant common genetic risk locus, whereas *PEMT* rs7946 has not been consisten

tly reported as a primary Alzheimer's disease GWAS locus [29]. This supports interpretation of rs7946 as a secondary, context-dependent exploratory candidate finding rather than a confirmed susceptibility locus. Additional variant-to-gene resources were used as external context, but did not alter the conservative interpretation that rs7946 should be treated as a non-primary Alzheimer's disease candidate requiring replication and functional validation (Table S8B).

No candidate variant had a ClinVar annotation indicating established Alzheimer's disease pathogenicity, further supporting conservative interpretation (**Table S8B**). Overall, the integrated framework combining model attribution, LD structure, allele-frequency comparison, and functional annotation supports considering *PEMT* rs7946 as one of several hypothesis-generating Non-APOE candidate variants for follow-up study, but not as a confirmed causal, independently associated, or independently predictive Alzheimer's disease variant (Figure S1; Table S8A-S8B).

LD analysis provides structural context for APOE-region and Non-APOE candidates

LD analysis showed minimal correlation between *PEMT* rs7946 and APOE-region variants, providing structural context for interpreting *PEMT* rs7946 separately from APOE-region markers, although LD patterns alone cannot establish statistical or biological independence from APOE-related effects (Figure 4; Table S6) [11]. In contrast, variants located near the *APOE* region were interpreted cautiously because local LD and tagging effects may complicate variant-specific interpretation (Figure 4;

Table S6) [24]. Allele-frequency comparisons showed an rs429358 C allele frequency of 0.252 in the WGS cohort, consistent with the AD-enriched cohort composition rather than a general Korean population frequency (Figure 6; Table S7) [3, 30]. Together, these analyses provided structural context for interpreting candidate variant contributions within the integrated modeling framework. Cross-stage integration of WES association results and WGS allele-frequency patterns identified APOE rs429358 as the dominant APOE-region signal. In contrast, the Non-APOE candidate variants, including PEMT rs7946, XRCC1 rs25489, and NKPD1 rs28469095, showed limited and variable model reliability across exploratory analyses. These findings were interpreted as exploratory candidate-evaluation context rather than confirmation of an independent susceptibility locus.

Discussion

This study evaluated how WES-nominated candidate variants behave within clinical-genomic Alzheimer models in which diagnostic-proximal variables can dominate discrimination. The full model showed near-perfect internal OOF discrimination (AUROC 0.998), but AUROC decreased by 0.285 to 0.713 after diagnostic-proximal variables were removed. Further removal of rs429358 and APOE proxy information reduced AUROC to 0.625. These estimates indicate that the high full-model performance largely reflected clinical and neuropsychological features closely related to diagnostic labeling rather than strong standalone genetic discrimination. APOE rs429358 remained the highest

t-ranking genetic feature when included, whereas the Non-APOE candidates contributed little to model performance. Feature attribution is therefore treated as model reliance under the observed data distribution, separate from genetic association or causal interpretation [3, 5, 20, 21].

PEMT rs7946 remained a hypothesis-generating candidate only. Its initial WES signal was attenuated after adjustment for age, sex, and APOE ϵ 4 carrier status, and the WGS association sensitivity analysis did not support corrected independent association. In the fold-wise validation permutation analysis, rs7946 produced a small mean AUROC decrease (0.00207), similar in scale to rs25489 (0.00230), while rs28469095 was near zero. The Non-APOE SNP-only model had AUROC 0.533 (95% CI 0.497–0.568), with the interval including chance-level discrimination. These results argue against interpreting the Non-APOE candidates as independent predictors, genetic effect-size estimates, or confirmed susceptibility loci.

Post hoc annotation provides biological context but does not strengthen the statistical evidence for these variants. *PEMT* participates in phosphatidylcholine biosynthesis and phospholipid homeostasis [2, 27], but rs7946 did not show a significant brain-tissue eQTL in the GTEx lookup used here, had no ClinVar annotation establishing Alzheimer's disease pathogenicity, and lacks independent replication. These observations do not establish regulatory function, pathogenicity, or mechanism. The cross-neurodegenerative examples cited above provide context for exploratory candidate evaluation but do not establish *PEMT* rs7946 as a cross-disease locus [6, 7].

Limitations

Several limitations remain. Candidate screening was restricted to five carried-forward variants and was not a formal exome-wide discovery analysis. The WES and WGS stages used different endpoints, so the WGS stage is not a direct replication of the aMCI association findings. A second independent Korean cohort with harmonized candidate genotypes and comparable clinical and neuropsychological variables was not available within the present study, precluding direct external replication. Accordingly, the machine-learning estimates represent internal pooled OOF performance rather than prospective or external validation. The bootstrap confidence intervals resample the fixed pooled OOF predictions and therefore do not incorporate the full uncertainty that would arise from repeating model fitting in new samples. Age and Test_Age represent age at diagnosis and age at neuropsychological testing, respectively; the categorical variable Ages was an exact five-year banding of Test_Age in this cohort and was excluded to avoid redundant age encoding. External reference-projected ancestry principal components were unavailable for WES, so residual population structure cannot be fully excluded despite internal genotype-derived PC sensitivity analyses. The full clinical-genomic model includes diagnostic-proximal variables and should be interpreted as an upper-bound clinical separability model rather than evidence of prospective clinical utility.

Conclusions

In this Korean WES–WGS exploratory framework, the full clinical–genomic model showed near-perfect internal OOF discrimination, but AUROC decreased to 0.713 after diagnostic-proximal cognitive and functional variables were removed and to 0.625 after additional removal of rs429358 and APOE proxy information. Non-APOE SNPs alone showed weak discrimination (AUROC 0.533; 95% CI 0.497–0.568). Age-related, education-related, demographic, and APOE-region variables accounted for most strict-model reliance, while PEMT rs7946, XRCC1 rs25489, and NKPD1 rs28469095 remained hypothesis-generating candidates only. These findings support cautious model-based candidate evaluation rather than confirmation of novel Alzheimer’s disease susceptibility loci, independent predictors, or causal variants.

List of abbreviations

AD, Alzheimer’s disease; aMCI, amnestic mild cognitive impairment; APOE, apolipoprotein E; AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; CDR, Clinical Dementia Rating; FDR, false discovery rate; GTEx, Genotype-Tissue Expression; KNIH, Korea National Institute of Health; LD, linkage disequilibrium; MMSE, Mini-Mental State Examination; OOF, out-of-fold; PCA, principal component analysis; RCFT, Rey Complex Figure Test; SHAP, SHapley Additive exPlanations; SNP, single-nucleotide polymorphism; SVLT, Seoul Verbal Learning Test; VEP, Variant Effect Predictor; WES, whole-exome sequencing; WGS, whole-genome sequencing

g.

Declarations

Ethics approval and consent to participate

This study protocol was reviewed and approved by the Institutional Review Board of Soonchunhyang University Bucheon Hospital (SCHBC-2020-03-016-002) and Korea Institute of Radiological & Medical Sciences (KIRAMS 2026-02-004). All participants provided written informed consent using IRB-approved documents prior to participation.

Consent for publication

Not applicable.

Availability of data and materials

The WES datasets are not publicly available but are available from the corresponding author upon reasonable request. The WGS dataset was accessed through the Korea National Institute of Health Clinical & Omics Data Archive (CODA), Republic of Korea, subject to the data-access policies of th

e Korea Disease Control and Prevention Agency (KDCA). This study used biomedical research resources, including genetic and health information, provided through CODA under access approval No. CODA_S2600052R-02. Reproducibility code supporting the reported WES candidate analyses, WGS pooled out-of-fold modeling, candidate-level sensitivity analyses, linkage-disequilibrium calculations, and manuscript figure/table generation is publicly available at GitHub: https://github.com/robertpark1228/Alzheimer_Amci/tree/main/reproducibility.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2020-NR049559), the Soonchunhyang University Research Fund, and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (No. HI19C1132). This work was also supported by grants from the Korea Institute of Radiological and Medical Sciences (KIRAMS), funded by the Ministry of Science and ICT (MSIT), Republic of Korea.

529 a (Grant Nos. 50531-2026, 50558-2026, and 50559-2026).

530

531 **Authors' Contributions**

532 Conceptualization: YP, EP; Data curation: EP, ESL; Formal analysis: YP, SYK; Funding acquisition: JYL,

533 WK, JYS; Investigation: JYL, EP, ESL; Methodology: YP, JKP; Project administration: JYL; Resources: JY

534 L, WK, ESL, JKP; Software: YP; Supervision: JL, IK; Validation: YP, SYK, JYL, IK; Visualization: YP; Writi

535 ng–original draft: EP; Writing–review & editing: all authors. All authors read and approved the final

536 manuscript.

537

538 **AI-assisted tools**

539 AI was used during manuscript revision to assist with code drafting and language editing. All anal

540 yses and numerical outputs were executed and verified by the authors.

541

542 **Acknowledgements**

543 Not applicable.

544

References

1. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, Jun G, DeStefano AL, Bis JC, Beecham GW *et al.* **Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.** *Nature Genetics* 2013, **45**(12):1452-1458.
2. Serrano-Pozo A, Das S, Hyman BT: **APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches.** *Lancet Neurol* 2021, **20**(1):68-80.
3. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA: **Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families.** *Science* 1993, **261**(5123):921-923.
4. Chen T, Guestrin C: **XGBoost: A Scalable Tree Boosting System.** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA.* Association for Computing Machinery 2016: 785–794.
5. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Banti N, Lee SI: **From Local Explanations to Global Understanding with Explainable AI for Trees.** *Nat Mach Intell* 2020, **2**(1):56-67.
6. Dugan AJ, Nelson PT, Katsumata Y, Shade LMP, Teylan MA, Boehme KL, Mukherjee S, Kaur JSK, Hohman TJ, Schneider JA, Fardo DW: **Association between WWOX/MAF variants and dementia-related neuropathologic endophenotypes.** *Neurobiol Aging* 2022, **111**:95-106.

- 564 7. Yuan YT, Hong WP, Tan CH, Yu RL: **Influence of WWOX/MAF genes on cognitive perfor**
565 **mance in patients with Parkinson's disease.** *Neurobiol Dis* 2025, **208**:106887.
- 566 8. Park H-Y: **Genome-based medicine in Korea: the Korea National Institute of Health infr**
567 **astructure for precision medicine.** *Child Kidney Dis.* 2026, **30**(1):15-23.
- 568 9. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.**
569 *Bioinformatics* 2010, **26**(5):589-595.
- 570 10. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jorda
571 n T, Shakir K, Roazen D, Thibault J *et al.* **From FastQ Data to High-Confidence Variant Ca**
572 **lls: The Genome Analysis Toolkit Best Practices Pipeline.** *Current Protocols in Bioinforma*
573 *tics* 2013, **43**(1):11.10.11-11.10.33.
- 574 11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de B
575 akker PIW, Daly MJ, Sham PC: **PLINK: A Tool Set for Whole-Genome Association and Po**
576 **pulation-Based Linkage Analyses.** *The American Journal of Human Genetics* 2007, **81**(3):5
577 59-575.
- 578 12. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: **Data qualit**
579 **y control in genetic case-control association studies.** *Nature Protocols* 2010, **5**(9):1564-1
580 573.
- 581 13. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F: **The**
582 **Ensembl Variant Effect Predictor.** *Genome Biology* 2016, **17**(1):122.

- 583 14. **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue g**
584 **ene regulation in humans.** *Science* 2015, **348**(6235):648-660.
- 585 15. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Li
586 u C: **ClinVar: improvements to accessing data.** *Nucleic acids research* 2020, **48**(D1):D835-
587 D844.
- 588 16. Bellenguez C, Küçükali F, Jansen IE, Kleindan L, Moreno-Grau S, Amin N, Naj AC, Campo
589 s-Martin R, Grenier-Boley B, Andrade V *et al.* **New insights into the genetic etiology of A**
590 **lzheimer's disease and related dementias.** *Nature Genetics* 2022, **54**(4):412-436.
- 591 17. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, Fumis L, Hay
592 hurst J, Buniello A, Karim MA *et al.* **An open approach to systematically prioritize causal**
593 **variants and genes at all published human GWAS trait-associated loci.** *Nature Genetics*
594 2021, **53**(11):1527-1533.
- 595 18. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G,
596 Marth GT, Sherry ST *et al.* **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**(1
597 5):2156-2158.
- 598 19. Friedman JH: **Greedy function approximation: a gradient boosting machine.** *Annals of s*
599 *tistics* 2001, **29**(5):1189-1232.
- 600 20. Serrano-Pozo A, Escott-Price V, Grinberg LT, Pascoal T, Suárez-Calvet M, Dubois B, Sperling
601 RA: **Alzheimer's disease.** *The Lancet* 2026, **407**(10544):2241-2262.

- 602 21. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J Mach*
603 *Learn Res.* 2019;20(177):1–81.
- 605 22. Hill WG, Robertson A: **The effect of linkage on limits to artificial selection.** *Genetical Res*
606 *earch* 1966, **8**(3):269-294.
- 607 23. Slatkin M: **Linkage disequilibrium — understanding the evolutionary past and mapping**
608 **the medical future.** *Nature Reviews Genetics* 2008, **9**(6):477-485.
- 609 24. Machiela MJ, Chanock SJ: **LDlink: a web-based application for exploring population-spe**
610 **cific haplotype structure and linking correlated alleles of possible functional variants.** *B*
611 *ioinformatics* 2015, **31**(21):3555-3557.
- 612 25. Breiman L: **Random Forests.** *Machine Learning* 2001, **45**(1):5-32.
- 613 26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful app
614 roach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995;57(1):289–300.
- 615 27. Li J, Xin Y, Li J, Chen H, Li H. Phosphatidylethanolamine N-methyltransferase: from function
616 s to diseases. *Aging Dis.* 2023;14(3):879–891.
- 617 28. Mahley RW: **Central Nervous System Lipoproteins: ApoE and Regulation of Cholesterol**
618 **Metabolism.** *Arterioscler Thromb Vasc Biol* 2016, **36**(7):1305-1315.
- 619 29. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, va
620 n der Lee SJ, Amlie-Wolf A *et al.* **Genetic meta-analysis of diagnosed Alzheimer's diseas**

621 **e identifies new risk loci and implicates A β , tau, immunity and lipid processing.** *Nat Ge*
622 *net* 2019, **51**(3):414-430.

623 30. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S,
624 McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015,
625 **526**(7571):68-74.

626

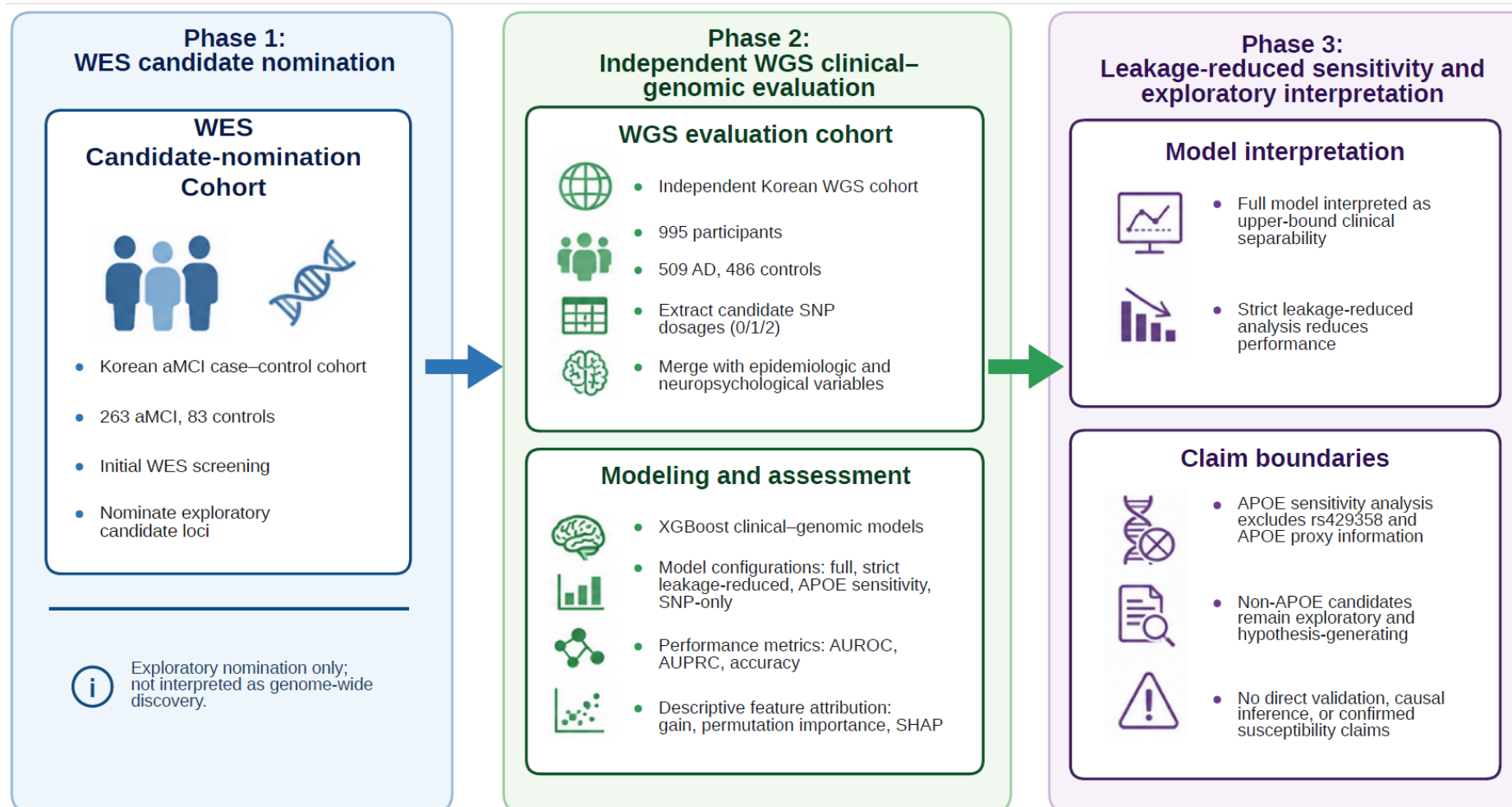


Figure 1 Study design and analysis workflow across the Alzheimer's continuum.

Candidate loci were nominated in a Korean whole-exome sequencing (WES) aMCI case–control cohort (263 aMCI, 83 controls). The nominated panel was evaluated in an independent WGS cohort (n=995; 509 Alzheimer’s disease cases, 486 controls). WGS genotypes were extracted and encoded as allele dosage (0/1/2), merged with epidemiologic and neuropsychological variables using the subject ID, and modeled using interpretable machine learning (XGBoost). Model configurations included SNP-only, full clinical–genomic, strict leakage-reduced, APOE-sensitivity, and Non-APOE SNP-only analyses. Variant contributions were assessed as descriptive model-reliance measures using gain, permutation importance, and SHAP summaries. The workflow was designed for exploratory candidate evaluation rather than genome-wide discovery, causal inference, or clinical validation.

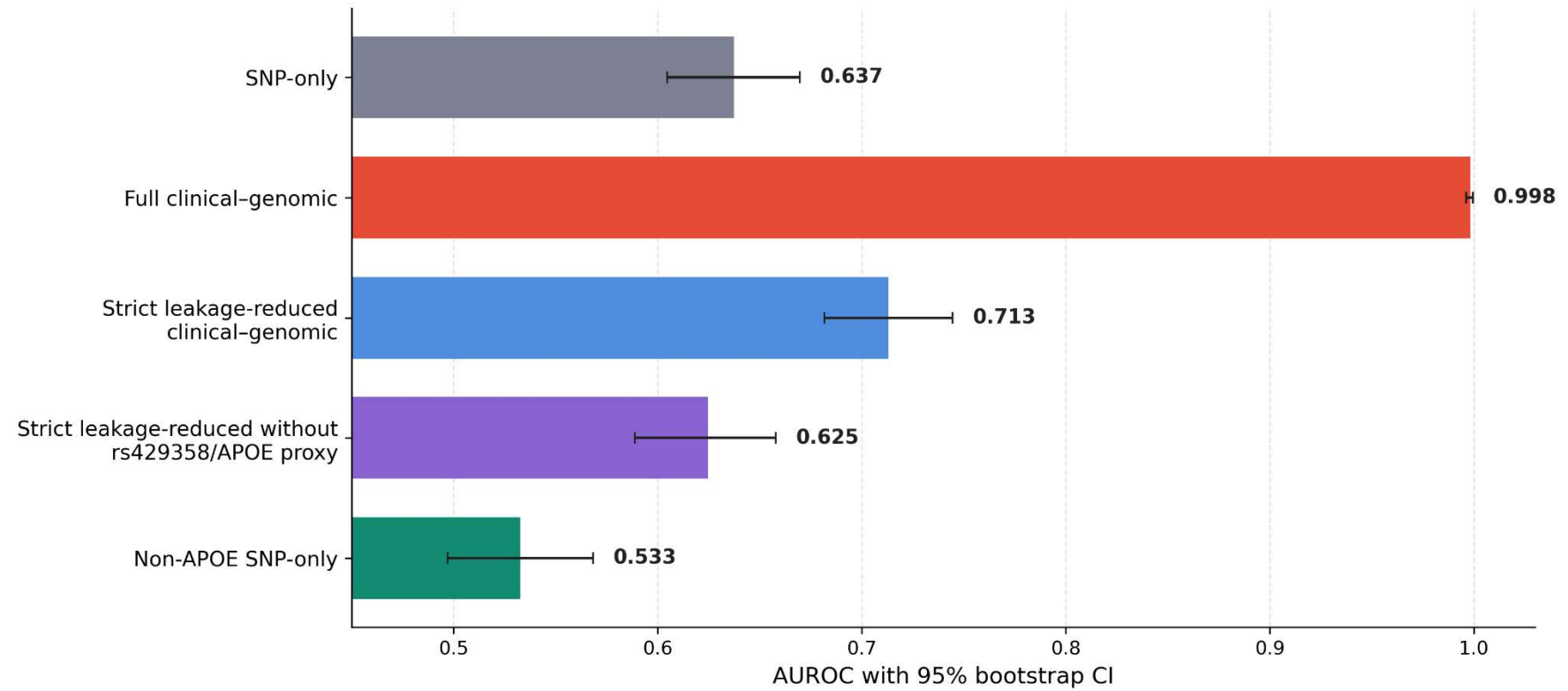
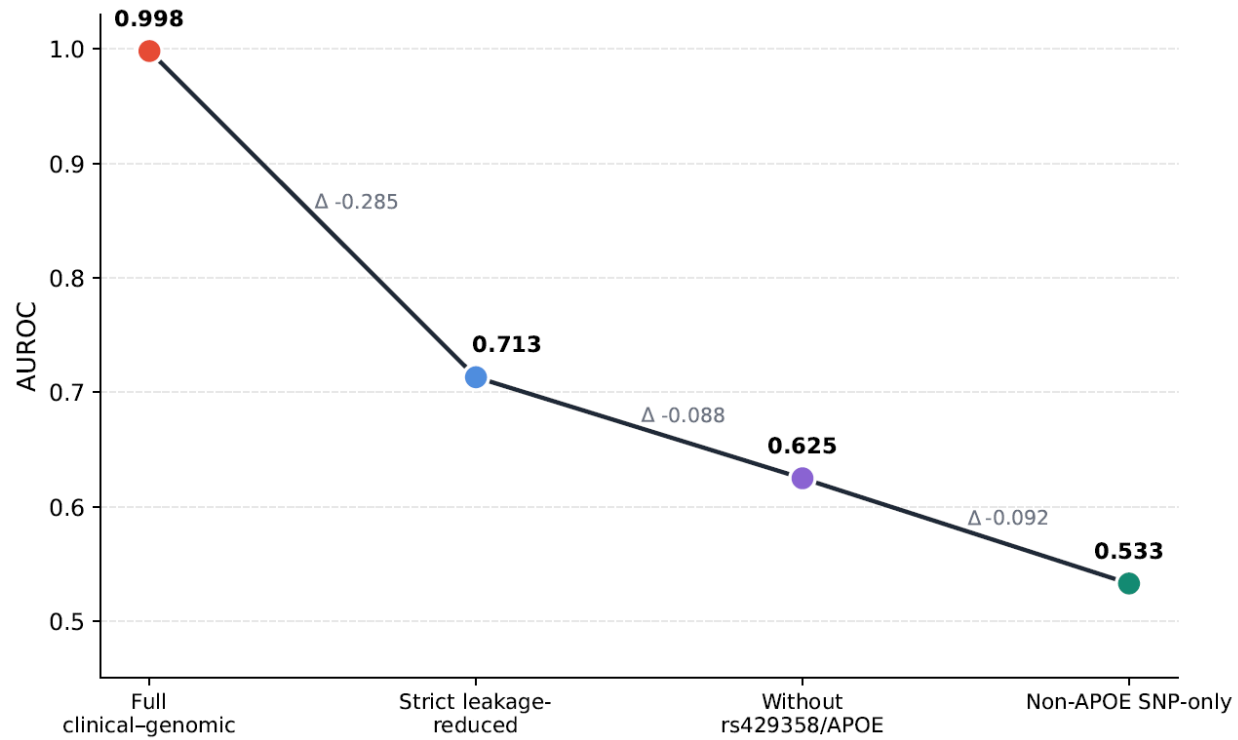


Figure 2 Model performance ladder after diagnostic-leakage reduction using pooled out-of-fold evaluation.

AUROC values with 95% bootstrap confidence intervals are shown for the five-SNP model, full clinical-genomic model, strict leakag

e-reduced clinical-genomic model, strict leakage-reduced model without rs429358 and APOE proxy information, and Non-APOE SNP-only model. All values were calculated from the same stratified five-fold pooled OOF predictions. AUROC decreased from 0.998 in the full model to 0.713 after removal of diagnostic-proximal variables and to 0.625 after additional APOE sensitivity removal. The Non-APOE SNP-only model had AUROC 0.533.



Interpretation: near-perfect full-model AUROC reflects upper-bound clinical separability.

Figure 3 Reduction in pooled OOF model discrimination after removing diagnostic-proximal variables.

The full clinical-genomic model (AUROC 0.998) is compared with the strict leakage-reduced model (0.713), the strict model without

rs429358/APOE proxy information (0.625), and the Non-APOE SNP-only model (0.533), using identical stratified five-fold OOF assignments. The 0.285 AUROC decrease from the full to strict model supports interpretation of the full model as upper-bound clinical separability rather than prospective predictive utility.

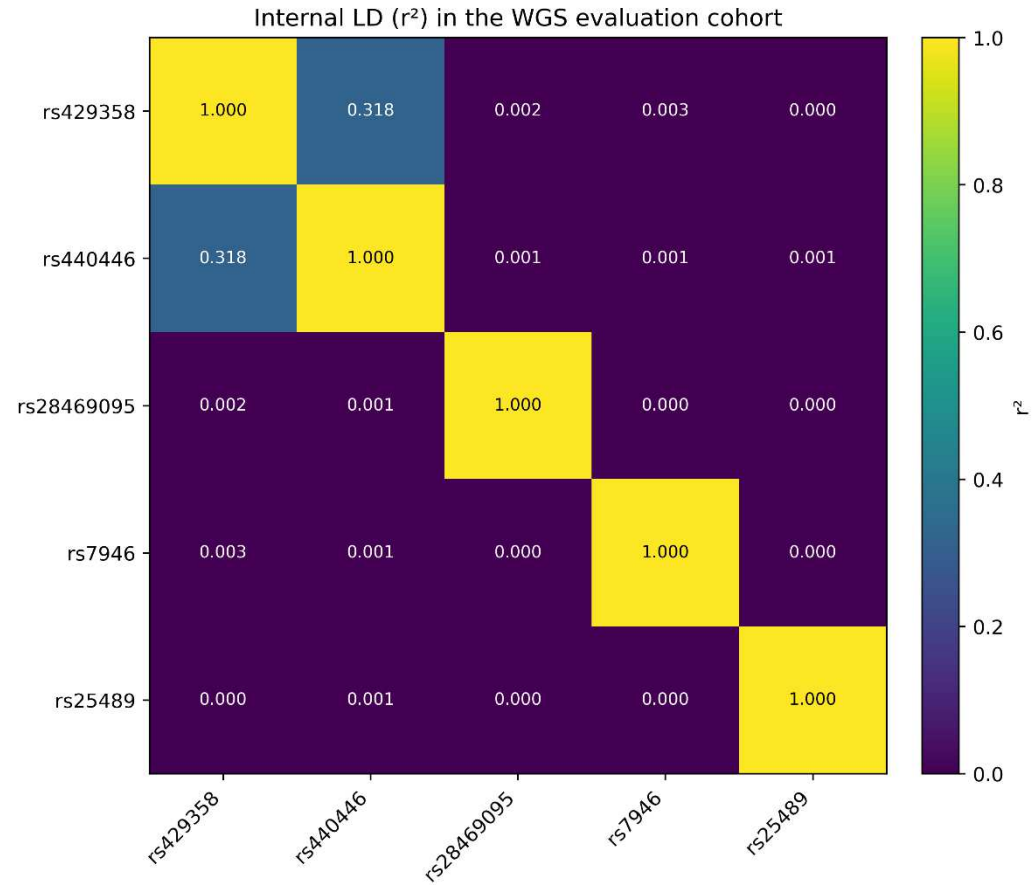


Figure 4. Internal linkage disequilibrium among the five candidate loci in the WGS evaluation cohort.

Pairwise LD was computed as r^2 from dosage-encoded genotypes using pairwise complete observations. rs429358 and rs440446 showed moderate LD, whereas *PEMT* rs7946 showed minimal LD with APOE-region variants. These LD results provide structural context for candidate interpretation but do not establish biological or statistical independence from APOE-related mechanisms.

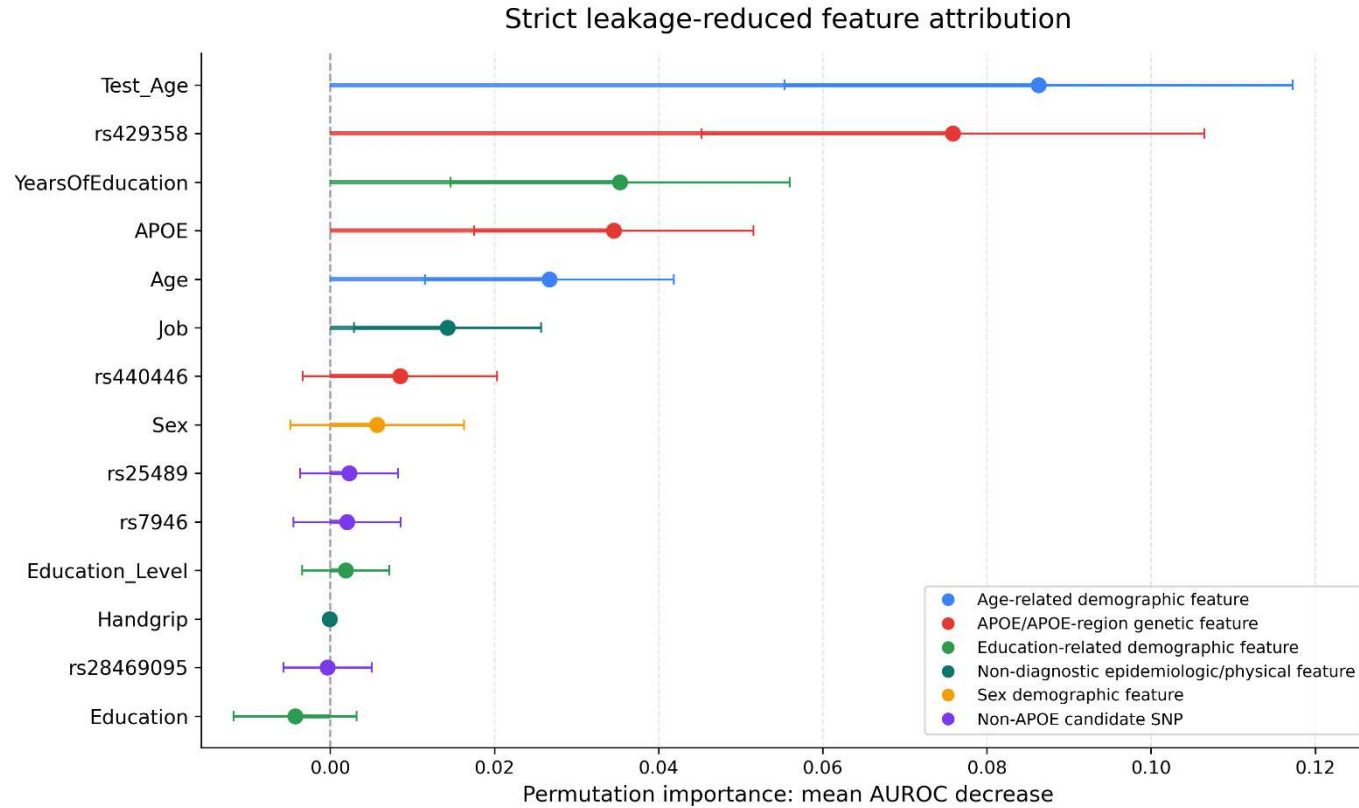


Figure 5. Fold-wise validation feature attribution in the strict leakage-reduced model.

Permutation importance was calculated as the AUROC decrease after feature shuffling within each outer validation fold and aggreg

ated across five folds and 50 repeats per fold. Test_Age ranked first overall (mean decrease 0.0863), while rs429358 was the highest-ranking genetic feature (0.0759). The Non-APOE candidates showed small or near-zero mean changes: rs25489 0.00230, rs7946 0.00207, and rs28469095 -0.00032 . Age was removed as a redundant age-band variable. Attribution values are interpreted as model reliance only.

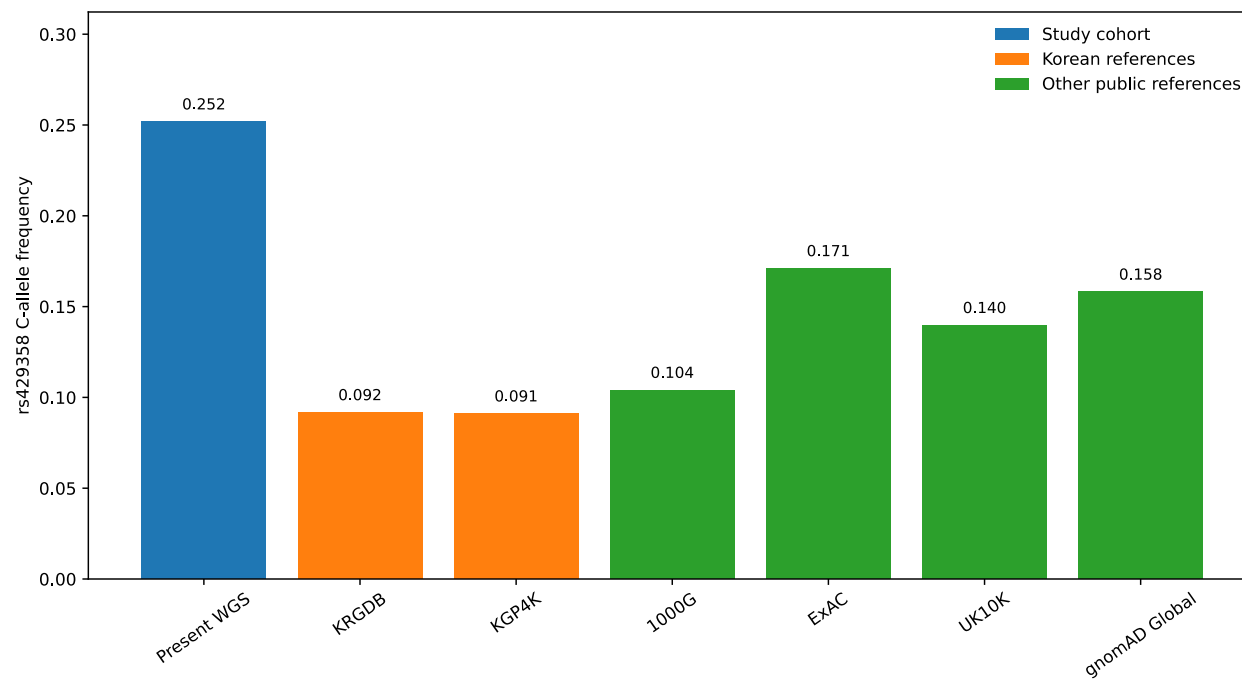


Figure 6. Comparison of rs429358 C-allele frequency between the present WGS cohort and public reference datasets.

The rs429358 C-allele frequency in the present WGS cohort was 0.2518 (501 C alleles among 1,990 chromosomes; n=995). The same compared allele and the same reference-source set shown in Supplementary Table S7 were used in the figure. External frequencies are presented for descriptive population context only and should be interpreted cautiously because reference datasets differ in population composition, sampling frame, sequencing platform, release, and allele-frequency estimation procedures.

Table 1 Baseline characteristics of the WES candidate-nomination cohort.

Group	N	Age, mean ± SD	MMSE, mean ± SD	Male, n (%)	<i>APOE</i> ε4 carrier, n (%)	P value for age	P value for MMSE	P value for sex	P value for <i>APOE</i> ε4
aMCI	263	74.42 ± 8.33	22.61 ± 4.42	96 (36.5)	77 (29.3)	0.001	0.001	0.176	0.006

Control	83	65.27 ± 9.4 0	27.34 ± 1.87	24 (28.9)	12 (14.5)	0.001	0.001	0.176	0.006
---------	----	------------------	--------------	-----------	-----------	-------	-------	-------	-------

Demographic and key clinical characteristics of the Korean WES candidate-nomination cohort, comprising 263 participants with amnesic mild cognitive impairment (aMCI) and 83 cognitively normal controls.

Table 2. Initial nominal WES screening results for nominated candidate variants.

Gene	SNP	Alternate allele count in aMCI	Alternate allele count in controls	Chi-square statistic	P value	Odds ratio (OR)	Standard error (SE)	95% confidence interval (CI)
<i>PEMT</i>	rs7946	155	36	4.038	0.044	1.53	0.213	1.009–2.320
<i>XRCC1</i>	rs25489	69	11	5.324	0.021	2.149	0.338	1.107–4.169
<i>APOE</i>	rs440446	223	55	4.397	0.036	1.486	0.190	1.025–2.155
<i>APOE</i>	rs429358	88	13	7.93	0.004	2.359	0.312	1.279–4.352
<i>NKPD1</i>	rs28469095	39	5	4.177	0.040	2.6	0.484	1.007–6.713

Allele-based screening statistics are shown for the five variants carried forward for exploratory WGS evaluation. These results were used for initial candidate nomination only. Covariate-adjusted and multiple-testing-adjusted analyses are reported in Supplementar

y Table S10. Candidate status does not imply corrected statistical significance, causal inference, biological validation, or confirmation of Alzheimer’s disease susceptibility loci.

Table 3. Performance of prespecified model configurations in the independent WGS evaluation cohort.

Model	N	Feat	AUROC (95%	AUPRC (95%	Accur	Balanced	Sensiti	Specifi	Brier	Confusion matrix (T
		ures	CI)	CI)	acy	accuracy	vity	city		N, FP, FN, TP)
SNP-only	995	5	0.637 (0.605–0.670)	0.656 (0.624–0.690)	0.600	0.599	0.637	0.562	0.235	273, 213, 185, 324

Model	N	Feat	AUROC (95%	AUPRC (95%	Accur	Balanced	Sensiti	Specifi	Brier	Confusion matrix (T
		ures	CI)	CI)	acy	accuracy	vity	city		N, FP, FN, TP)
Full clinical-genomic	995	142	0.998 (0.996–0.999)	0.998 (0.996–0.999)	0.981	0.981	0.990	0.971	0.014	472, 14, 5, 504
Strict leakage-reduced clinical-genomic	995	14	0.713 (0.682–0.745)	0.723 (0.686–0.758)	0.655	0.656	0.644	0.667	0.217	324, 162, 181, 328
Strict leakage-reduced without rs429358/APOE proxy	995	12	0.625 (0.589–0.658)	0.623 (0.589–0.659)	0.590	0.589	0.611	0.568	0.243	276, 210, 198, 311
Non-APOE SNP-only	995	3	0.533 (0.497–0.568)	0.537 (0.508–0.569)	0.533	0.530	0.654	0.405	0.250	197, 289, 176, 333

(Note. All rows use one common stratified five-fold pooled OOF evaluation with identical fold assignments. AUROC and AUPRC are

reported with 95% confidence intervals from 2,000 paired class-stratified bootstrap resamples of the pooled OOF predictions. Classification metrics use a prespecified 0.50 threshold. Confusion matrices are ordered as TN, FP, FN, TP.)

Table 4. Fold-wise validation permutation importance in the strict leakage-reduced model.

Rank	Feature	Feature group	Permutation importance mean AUROC decrease	Permutation importance SD	Interpretation
1	Test_Age	Age-related demographic feature	0.086317	0.030968	Descriptive demographic model reliance
2	rs429358	APOE/APOE-region genetic feature	0.075857	0.030636	Known APOE-region model reliance; not novel

Ran k	Feature	Feature group	Permutation importance mean AUROC decrease	Permutation importance SD	Interpretation
3	YearsOf Educati on	Education-related demographic fe ature	0.035303	0.020655	Descriptive demographic model reliance
4	APOE	APOE/APOE-region genetic featur e	0.034571	0.017022	APOE proxy model reliance
5	Age	Age-related demographic feature	0.026704	0.015162	Descriptive demographic model reliance
6	Job	Non-diagnostic epidemiologic/ph ysical feature	0.014297	0.011404	Model reliance only

Rank	Feature	Feature group	Permutation importance mean AUROC decrease	Permutation importance SD	Interpretation
7	rs440446	APOE/APOE-region genetic feature	0.008493	0.011849	APOE-region marker; not interpreted as APOE-independent
8	Sex	Sex demographic feature	0.005724	0.010562	Descriptive demographic model reliance
9	rs25489	Non-APOE candidate SNP	0.002302	0.005959	Exploratory model-reliance signal only
10	rs7946	Non-APOE candidate SNP	0.002066	0.006520	Exploratory model-reliance signal only
11	Education_Level	Education-related demographic feature	0.001893	0.005285	Descriptive demographic model reliance

Rank	Feature	Feature group	Permutation importance mean AUROC decrease	Permutation importance SD	Interpretation
12	Handgrip	Non-diagnostic epidemiologic/physical feature	-0.000063	0.000563	Model reliance only
13	rs28469095	Non-APOE candidate SNP	-0.000321	0.005368	Near-zero exploratory model reliance
14	Education	Education-related demographic feature	-0.004250	0.007463	Descriptive demographic model reliance

Permutation importance was summarized as mean AUROC decrease from feature shuffling performed only in outer validation folds and aggregated across five folds and 50 repeats per fold (250 estimates per feature). Ages was excluded as a redundant categorical age-band variable. Attri

bution values indicate model reliance under the observed data distribution and were not interpreted as genetic association, causality, biological validation, or independent predictive utility.

Table 5. Feature inclusion and exclusion audit across model configurations.

Model	Feature category	Number of variables	Status	Examples / notes
Full clinical–genomic model	Diagnostic-proximal / neuropsychological / functional variables	128	Included	Amyloid PET, CDR, CDR-SB, MMSE total and items, S VLT, RCFT, SNSB, ADL, Stroop, TMT, COWAT, DSC, CDT, and related cognitive/functional variables
Full clinical–genomic model	Retained demographic / non-diagnostic variables	8	Included	Age, Test_Age, education-related variables, sex, job, and handgrip; Ages excluded as redundant
Full clinical–genomic model	Candidate SNP dosages	5	Included	rs7946, rs25489, rs440446, rs429358, rs28469095
Full clinical–genomic model	APOE proxy/genotype information	1	Included	APOE genotype/proxy variable
Strict leakage-reduced model	Diagnostic-proximal / neuropsychological / functional variables	128	Excluded	Removed to reduce diagnostic reconstruction

Model	Feature category	Number of variables	Status	Examples / notes
Strict leakage-reduced model I	Retained demographic / non-diagnostic variables	8	Included	Age, Test_Age, education-related variables, sex, job, and handgrip; Ages excluded as redundant
Strict leakage-reduced model II	Candidate SNP dosages + APOE proxy	6	Included	Five candidate SNP dosages plus APOE genotype/proxy information
Strict leakage-reduced without rs429358/APOE proxy	APOE sensitivity removal	2	Excluded	rs429358 and APOE genotype/proxy information removed from the 14-feature strict model, leaving 12 features
Non-APOE SNP-only	Non-APOE candidate SNPs	3	Included	rs7946, rs25489, rs28469095 only

The table summarizes which feature classes are included or excluded in the full clinical-genomic, strict leakage-reduced, APOE-sensitivity, and Non-APOE SNP-only models. Diagnostic-proximal, neuropsychological, cognitive, memory-domain, and functional variables

bles are retained in the full model but excluded from the strict leakage-reduced model. ID, FAMILY_ID, and PNTTM are excluded from all models. Ages is removed as a redundant categorical age-band variable.

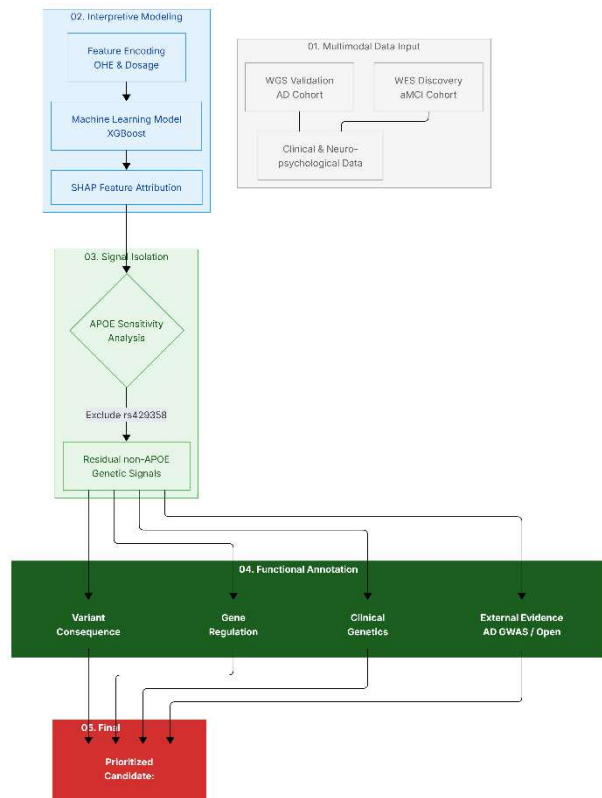


Figure S1. Functional annotation framework for candidate variant prioritization.

Candidate variants nominated from the WES cohort were evaluated in an independent Korean WGS cohort using machine learning–based attribution, *APOE* sensitivity analysis, and population-genetic context. Post hoc functional annotation incorporated variant consequence, GTEx/eQTL lookup, ClinVar interpretation, and comparison with external AD genetic resources (GWAS and Open Targets). This framework supports conservative prioritization rather than causal inference.

Table S1. Additive logistic-regression sensitivity analysis of WES candidate variants.

SNP	A1	NMISS	TEST	OR	SE	L95	U95	STAT	P
rs7946	T	332	ADD	1.672	0.2335	1.058	2.643	2.203	0.02763
			Gender	0.592	0.3016	0.3278	1.069	-1.738	0.08221

			<i>APOE</i>	2.417	0.3593	1.195	4.887	2.456	0.01405
rs25489	T	333	ADD	1.909	0.3708	0.923	3.949	1.744	0.08118
			Gender	0.6641	0.3025	0.3671	1.202	-1.353	0.1761
			<i>APOE</i>	2.178	0.3569	1.082	4.383	2.18	0.02923
rs440446	G	333	ADD	1.207	0.2294	0.7699	1.892	0.8202	0.4121
			Gender	0.6247	0.2986	0.3479	1.122	-1.576	0.1151
			<i>APOE</i>	1.959	0.3948	0.9034	4.246	1.703	0.08864
rs429358	C	332	ADD	1.785	1.037	0.2339	13.62	0.5589	0.5763
			Gender	0.6319	0.2988	0.3518	1.135	-1.536	0.1244
			<i>APOE</i>	1.215	1.159	0.1253	11.78	0.1677	0.8668
rs28469095	C	333	ADD	2.478	0.4905	0.9476	6.481	1.85	0.06428

	Gender	0.6242	0.2999	0.3468	1.124	-1.571	0.1161
	<i>APOE</i>	2.301	0.3565	1.144	4.628	2.338	0.0194

Supplementary additive models are presented to characterize inheritance-model sensitivity and were not used for primary candidate-level inference. Because age differed between aMCI cases and cognitively normal controls, the primary candidate-level analysis was as the age-adjusted additive model with candidate-panel multiple-testing correction reported in Supplementary Table S10. APOE-region variants in the primary analysis were modeled separately from APOE ϵ 4 carrier status to avoid genotype-related collinearity.

Table S2. Dominant logistic-regression sensitivity analysis of WES candidate variants.

SNP	A1	NMISS	TEST	OR	SE	L95	U95	STAT	P
rs7946	T	332	DOM	1.884	0.2816	1.085	3.271	2.249	0.02451
			Gender	0.5871	0.3017	0.325	1.06	-1.765	0.0775
			<i>APOE</i>	2.448	0.3599	1.209	4.957	2.488	0.01286
rs25489	T	333	DOM	1.903	0.3738	0.9148	3.96	1.722	0.08513
			Gender	0.6618	0.3024	0.3659	1.197	-1.365	0.1722
			<i>APOE</i>	2.175	0.3569	1.081	4.379	2.177	0.02945

rs440446	G	333	DOM	1.277	0.3036	0.7045	2.316	0.8063	0.42
			Gender	0.6279	0.2988	0.3496	1.128	-1.558	0.1193
			<i>APOE</i>	2.017	0.3827	0.9526	4.271	1.833	0.0668
rs429358	C	332	DOM	NA	NA	NA	NA	NA	NA
			Gender	NA	NA	NA	NA	NA	NA
			<i>APOE</i>	NA	NA	NA	NA	NA	NA
rs28469095	C	333	DOM	2.478	0.5159	0.9015	6.812	1.759	0.07859
			Gender	0.62	0.2998	0.3445	1.116	-1.595	0.1107
			<i>APOE</i>	2.289	0.3563	1.139	4.602	2.324	0.0201

Dominant-model results are provided as a supplementary inheritance-model sensitivity analysis and were not used for primary candidate-level inference. The primary candidate-level analysis used age-adjusted additive models with candidate-panel multiple-testin

g correction, as reported in Supplementary Table S10.

Table S3. Recessive logistic-regression sensitivity analysis of WES candidate variants.

SNP	A1	NMISS	TEST	OR	SE	L95	U95	STAT	P
rs7946	T	332	REC	1.817	0.5822	0.5805	5.688	1.026	0.305
			Gender	0.6149	0.299	0.3422	1.105	-1.626	0.1039
			<i>APOE</i>	2.293	0.3555	1.142	4.602	2.334	0.01958
rs25489	T	333	REC	NA	NA	NA	NA	NA	NA
			Gender	NA	NA	NA	NA	NA	NA
			<i>APOE</i>	NA	NA	NA	NA	NA	NA

rs440446	G	333	REC	1.198	0.434	0.5118	2.806	0.4169	0.6767
			Gender	0.618	0.2981	0.3445	1.109	-1.614	0.1065
			<i>APOE</i>	2.141	0.3754	1.026	4.468	2.028	0.04261
rs429358	C	332	REC	1.526	1.128	0.1671	13.93	0.3744	0.7081
			Gender	0.6272	0.2984	0.3495	1.126	-1.563	0.118
			<i>APOE</i>	2.187	0.3645	1.071	4.467	2.147	0.0318
rs28469095	C	333	REC	NA	NA	NA	NA	NA	NA
			Gender	NA	NA	NA	NA	NA	NA
			<i>APOE</i>	NA	NA	NA	NA	NA	NA

Recessive-model results are provided as a supplementary inheritance-model sensitivity analysis. These models were not used for primary candidate-level inference. Because age differed between aMCI cases and cognitively normal controls, the primary candidate-l

level analysis was based on age-adjusted additive models with candidate-panel multiple-testing correction, as reported in Supplementary Table S10.

Table S4. Prespecified WGS model configurations used for pooled out-of-fold evaluation.

			Candidate			Analytical	Primary
Model	N features	Core feature set	genetic features included	Diagnostic- proximal variables	APOE-related information	purpose / interpretatio n	result locatio n
SNP-only	5	Candidate SNP dosages only	rs7946, rs25489, rs28469095, rs429358, rs440446	Excluded	Includes rs429358 and APOE-region rs440446	Genetic	Table 3; Figure 2
						baseline	
						model; not	
						intended to	
						establish	
						standalone	
						clinical utility	

Full clinical– genomic	142	Candidate SNP			Upper-bound	
			rs7946,			
		dosages plus full			clinical	
			rs25489,			
		available			separability	
			rs28469095,			
		epidemiologic,			model; high	
			rs429358,			
		clinical, biomarker,			performance	Table 3;
			rs440446;	Included	Included	
		cognitive,			interpreted	Figure 2
			<i>APOE</i>			
		neuropsychological,			cautiously	
			proxy/genotyp			
		and functional			because	
			e information			
		variables after			diagnostic-	
			when available			
		preprocessing			proximal	

Model	N features	Core feature set	Candidate			Analytical	Primary
			genetic	Diagnostic-	APOE-related	purpose /	result
			features	proximal variables	information	interpretatio	locatio
			included			n	n
						variables are	
						included	

Strict leakage-reduced clinical-genomic 14	Candidate SNP dosage s plus retained demog raphic, education-relat ed, and non-diagnostic epidemiologic/physical variables; redundant a ge-band variable Ages removed when available	rs7946, rs25489, rs28469095, rs429358, rs440446; <i>APOE</i> proxy/genotyp e information 	Excluded: amyloid PET, CDR-related variables, MMSE total and q01–q19 item-level variables, SVLT, RCFT, broader neuropsychological scores, memory- domain variables, and functional	Primary diagnostic- leakage sensitivity model; used to evaluate model performance after removing diagnostic	Table 3; Figure 2; Figure 3; Table 4; Table S11
			Included		

Model	N features	Core feature set	Candidate			Analytical	Primary
			genetic	Diagnostic-	APOE-related	purpose /	result
			features	proximal variables	information	interpretatio	locatio
			included			n	n
			assessment		reconstruction		
			variables		variables		

Strict leakage- reduced without rs429358/<i>APO</i> <i>E</i> proxy	12	Strict leakage- reduced feature set with additional <i>APOE</i> sensitivity removal	rs7946,			<i>APOE</i>	
			rs25489,			sensitivity	
			rs28469095,			model;	
			rs440446		rs429358 and	evaluates how	Table 3;
			retained;	Excluded as in	<i>APOE</i>	performance	Figure
			rs429358 and	strict leakage-	proxy/genotyp	changes after	2;
			<i>APOE</i>	reduced model	e information	removing the	Figure 3
			proxy/genotyp		removed	dominant	
			e information			<i>APOE</i> -region	
			removed			marker and	

Model	N features	Core feature set	Candidate			Analytical	Primary
			genetic	Diagnostic-	APOE-related	purpose /	result
			features	proximal variables	information	interpretatio	locatio
			included			n	n
						<i>APOE</i> proxy	
						information	

					Standalone					
					Non-APOE					
Non-APOE SNP-only	3	Non-APOE	rs7946,	Excluded	Excluded	genetic signal	Table 3;			
		candidate SNP	rs25489,			check; weak	Figure			
		dosages only	rs28469095			discrimination	2;			
						supports	Figure 3			
							exploratory			
							interpretation			

The table defines the five prespecified feature sets used in the common stratified five-fold pooled OOF evaluation. ID, FAMILY_ID, PNTTM, and Ages were excluded before fitting. Performance estimates are reported in Table 3.

Table S5. WGS feature inventory, variable definitions, and explicit exclusions used for modeling.

Feature	Data_type	Group	Description
PNTTM	Procedural	Identifier/procedural variable	Procedural/constant field; excluded before model fitting and never used as a predictor
Amyloid_PET	Numeric	Biomarker/Imaging	Amyloid PET variable
Age	Numeric	Demographic/Socioeconomic	Age at Alzheimer diagnosis (diagnosis age)
Sex	Numeric	Demographic/Socioeconomic	Sex
q01	Numeric	MMSE item-level / diagnostic-proximal	MMSE item-level variable; excluded from strict leakage-reduced model
q02	Numeric		
q03	Numeric		
q04	Numeric		

Feature	Data_type	Group	Description
q05	Numeric		
q06	Numeric		
q07	Numeric		
q08	Numeric		
q09	Numeric		
q10	Numeric		
q11_1	Numeric		
q11_2	Numeric		
q11_3	Numeric		
q12_1	Numeric		

Feature	Data_type	Group	Description
q12_2	Numeric		
q12_3	Numeric		
q12_4	Numeric		
q12_5	Numeric		
q13_1	Numeric		
q13_2	Numeric		
q13_3	Numeric		
q14_1	Numeric		
q14_2	Numeric		
q15	Numeric		

Feature	Data_type	Group	Description
q16_1	Numeric		
q16_2	Numeric		
q16_3	Numeric		
q17	Numeric		
q18	Numeric		
q19	Numeric		
total	Numeric	-	MMSE total score
Test_Age	Numeric	Demographic/Socioeconomic	Age at neuropsychological testing; distinct from diagnosis age

Feature	Data_type	Group	Description
YearsOfEducation	Numeric	Demographic/Socioeconomic	Years of education
Handgrip	Numeric	Clinical/Functional	Handgrip strength
Digit_span_Forward	Numeric	Other clinical/cognitive	Digit_span_Forward
Digit_span_Backward	Numeric	Other clinical/cognitive	Digit_span_Backward
Digit_span_Forward_Backward	Numeric	Other clinical/cognitive	Digit_span_Forward_Backward
Repetition	Numeric	Other clinical/cognitive	Repetition
Naming_K_BNT	Numeric	Other clinical/cognitive	Naming_K_BNT
Praxis_Ideomotor	Numeric	Other clinical/cognitive	Praxis_Ideomotor
Calculation_add	Numeric	Other clinical/cognitive	Calculation_add

Feature	Data_type	Group	Description
Calculation_sub	Numeric	Other clinical/cognitive	Calculation_sub
Calculation_mul	Numeric	Other clinical/cognitive	Calculation_mul
Calculation_div	Numeric	Other clinical/cognitive	Calculation_div
Calculation_total_score	Numeric	Other clinical/cognitive	Calculation_total_score
Rey_CFT_copy_score	Numeric	Other clinical/cognitive	Rey_CFT_copy_score
Rey_CFT_copy_time	Numeric	Other clinical/cognitive	Rey_CFT_copy_time
SVLT_recall_trial1	Numeric	Neuropsychological	SVLT_recall_trial1
SVLT_recall_trial2	Numeric	Neuropsychological	SVLT_recall_trial2
SVLT_recall_trial3	Numeric	Neuropsychological	SVLT_recall_trial3
SVLT_recall_total_score	Numeric	Neuropsychological	SVLT_recall_total_score

Feature	Data_type	Group	Description
SVLT_Delayed_recall	Numeric	Neuropsychological	SVLT_Delayed_recall
SVLT_recognition_TP	Numeric	Neuropsychological	SVLT_recognition_TP
SVLT_recognition_FP	Numeric	Neuropsychological	SVLT_recognition_FP
SVLT_recognition_score	Numeric	Neuropsychological	SVLT_recognition_score
SVLT_recognition_discriminability_index	Numeric	Neuropsychological	SVLT_recognition_discriminability_index
RCFT_immediate_recall	Numeric	Neuropsychological	RCFT_immediate_recall
RCFT_delayed_recall	Numeric	Neuropsychological	RCFT_delayed_recall
RCFT_recognition_TP	Numeric	Neuropsychological	RCFT_recognition_TP
RCFT_recognition_FP	Numeric	Neuropsychological	RCFT_recognition_FP
RCFT_recognition_score	Numeric	Neuropsychological	RCFT_recognition_score

Feature	Data_type	Group	Description
RCFT_recognition_discriminability_index	Numeric	Neuropsychological	RCFT_recognition_discriminability_index
Contrasting_program	Numeric	Other clinical/cognitive	Contrasting_program
Go_No_Go	Numeric	Other clinical/cognitive	Go_No_Go
COWAT_animal	Numeric	Neuropsychological	COWAT_animal
COWAT_supermarket	Numeric	Neuropsychological	COWAT_supermarket
COWAT_giut	Numeric	Neuropsychological	COWAT_giut
COWAT_iung	Numeric	Neuropsychological	COWAT_iung
COWAT_siut	Numeric	Neuropsychological	COWAT_siut
COWAT_phonemic_total_score	Numeric	Neuropsychological	COWAT_phonemic_total_score
StroopTest_Wordreading_correct	Numeric	Neuropsychological	StroopTest_Wordreading_correct

Feature	Data_type	Group	Description
StroopTest_Wordreading_error	Numeric	Neuropsychological	StroopTest_Wordreading_error
StroopTest_Wordreading_time	Numeric	Neuropsychological	StroopTest_Wordreading_time
StroopTest_Wordreading_Correct_response_rate	Numeric	Neuropsychological	StroopTest_Wordreading_Correct_response_rate
StroopTest_Wordreading_Time_per_item	Numeric	Neuropsychological	StroopTest_Wordreading_Time_per_item
StroopTest_Colorreading_correct	Numeric	Neuropsychological	StroopTest_Colorreading_correct
StroopTest_Colorreading_error	Numeric	Neuropsychological	StroopTest_Colorreading_error
StroopTest_Colorreading_time	Numeric	Neuropsychological	StroopTest_Colorreading_time
StroopTest_Colorreading_Correct_response_rate	Numeric	Neuropsychological	StroopTest_Colorreading_Correct_response_rate

Feature	Data_type	Group	Description
StroopTest_Colorreading_Time_per_item	Numeric	Neuropsychological	StroopTest_Colorreading_Time_per_item
StroopTest_Interference_score	Numeric	Neuropsychological	StroopTest_Interference_score
K_MMSE_orientation_to_time	Numeric	Neuropsychological	K_MMSE_orientation_to_time
K_MMSE_orientation_to_place	Numeric	Neuropsychological	K_MMSE_orientation_to_place
K_MMSE_Registation	Numeric	Neuropsychological	K_MMSE_Registation
K_MMSE_attention_and_calculation	Numeric	Neuropsychological	K_MMSE_attention_and_calculation
K_MMSE_recall	Numeric	Neuropsychological	K_MMSE_recall
K_MMSE_language	Numeric	Neuropsychological	K_MMSE_language
K_MMSE_drawing	Numeric	Neuropsychological	K_MMSE_drawing
K_MMSE_total_score	Numeric	Neuropsychological	K_MMSE_total_score

Feature	Data_type	Group	Description
CDR	Numeric	Clinical/Functional	Clinical Dementia Rating
B_ADL	Numeric	Clinical/Functional	Basic activities of daily living score
Sum_of_boxes	Numeric	Clinical/Functional	CDR sum of boxes
Education_Level	Numeric	Other clinical/cognitive	Education_Level
Comprehension_Total_Score	Numeric	Clinical/Functional	Comprehension_Total_Score
Naming_S_K_BNT	Numeric	Clinical/Functional	Naming_S_K_BNT
Digit_span_Forward_plus_Backward	Numeric	Clinical/Functional	Digit_span_Forward_plus_Backward
COWAT_animal_giut	Numeric	Neuropsychological	COWAT_animal_giut
CDT_Contour	Numeric	Clinical/Functional	CDT_Contour
CDT_Numbers	Numeric	Clinical/Functional	CDT_Numbers

Feature	Data_type	Group	Description
CDT_time_setting	Numeric	Clinical/Functional	CDT_time_setting
CDT_Total_score	Numeric	Clinical/Functional	CDT_Total_score
DSC_correct	Numeric	Clinical/Functional	DSC_correct
DSC_error	Numeric	Clinical/Functional	DSC_error
K_TMT_E_A_time	Numeric	Neuropsychological	K_TMT_E_A_time
K_TMT_E_A_error	Numeric	Neuropsychological	K_TMT_E_A_error
K_TMT_E_B_time	Numeric	Neuropsychological	K_TMT_E_B_time
K_TMT_E_B_error	Numeric	Neuropsychological	K_TMT_E_B_error
SNSB_II_Domain_Attention	Numeric	Neuropsychological	SNSB_II_Domain_Attention
SNSB_II_Domain_Language	Numeric	Neuropsychological	SNSB_II_Domain_Language

Feature	Data_type	Group	Description
SNSB_II_Domain_Visuospatial	Numeric	Neuropsychological	SNSB_II_Domain_Visuospatial
SNSB_II_Domain_Memory	Numeric	Neuropsychological	SNSB_II_Domain_Memory
SNSB_II_Domain_Frontal	Numeric	Neuropsychological	SNSB_II_Domain_Frontal
SNSB_C_Ani_giut	Numeric	Clinical/Functional	SNSB_C_Ani_giut
rs7946	Numeric	Genetic	<i>PEMT</i> SNP dosage
rs25489	Numeric	Genetic	<i>XRCC1</i> SNP dosage
rs440446	Numeric	Genetic	<i>APOE</i> -region SNP dosage
rs429358	Numeric	Genetic	<i>APOE</i> SNP dosage
rs28469095	Numeric	Genetic	<i>NKPD1</i> -region SNP dosage
<i>APOE</i>	Categorical	Genetic	<i>APOE</i> genotype string

Feature	Data_type	Group	Description
Education	Categorical	Demographic/Socioeconomic	Education
Job	Categorical	Demographic/Socioeconomic	Job
Letter_cancellation	Categorical	Other clinical/cognitive	Letter_cancellation
Spontaneous_speech_fluency	Categorical	Other clinical/cognitive	Spontaneous_speech_fluency
Spontaneous_speech_contents	Categorical	Other clinical/cognitive	Spontaneous_speech_contents
Comprehension	Categorical	Other clinical/cognitive	Comprehension
Reading	Categorical	Other clinical/cognitive	Reading
Writing	Categorical	Other clinical/cognitive	Writing

Feature	Data_type	Group	Description
Finger_naming	Categorical	Other clinical/cognitive	Finger_naming
Right_left_orientation	Categorical	Other clinical/cognitive	Right_left_orientation
Body_part_identification	Categorical	Other clinical/cognitive	Body_part_identification
Praxis_buccofacial	Categorical	Other clinical/cognitive	Praxis_buccofacial
Motor_impersistence	Categorical	Other clinical/cognitive	Motor_impersistence
Alternating_hand_movement	Categorical	Other clinical/cognitive	Alternating_hand_movement
Alternating_square_and_triangle	Categorical	Other clinical/cognitive	Alternating_square_and_triangle
Luria_loop	Categorical	Other clinical/cognitive	Luria_loop
K_TMT_E_A_success	Categorical	Neuropsychological	K_TMT_E_A_success
K_TMT_E_B_success	Categorical	Neuropsychological	K_TMT_E_B_success

Feature	Data_type	Group	Description
Fist_edge_palm_right	Categorical	Other clinical/cognitive	Fist_edge_palm_right
Fist_edge_palm_left	Categorical	Other clinical/cognitive	Fist_edge_palm_left
FAMILY_ID	Identifier	Identifier/linkage variable	Subject/linkage identifier in this analytic export; excluded before model fitting and never used as a predictor
ID	Identifier	Identifier/linkage variable	Subject identifier; excluded before model fitting and never used as a predictor
Ages	Categorical	Demographic/Socioeconomic	Five-year age-band variable derived exactly from Test_Age; excluded before model fitting to avoid redundant age encoding

This table summarizes the WGS feature inventory before model fitting. Genetic features are encoded as allele dosage (0/1/2), and c

ategorical variables are one-hot encoded within training folds. ID and FAMILY_ID are linkage identifiers and PNTTM is a procedural field; all three are excluded before model fitting. Age denotes age at Alzheimer’s disease diagnosis and Test_Age denotes age at neuropsychological testing. Ages is a categorical five-year age-band variable derived from Test_Age; this relationship was verified in all 995 participants. Ages is excluded from all models to avoid redundant age encoding. Thus, not every field listed in the raw feature inventory is used as a predictor.

Table S6. Pairwise LD among the five nominated candidate SNPs in the WGS cohort.

SNP1	SNP2	N_pairwise_complete	R	R2
rs429358	rs440446	995	0.563605	0.317651
rs429358	rs28469095	995	-0.03932	0.001546

rs429358	rs7946	995	-0.05379	0.002894
rs429358	rs25489	995	-0.00284	0.000008
rs440446	rs28469095	995	0.03312	0.001097
rs440446	rs7946	995	-0.03828	0.001465
rs440446	rs25489	995	0.022779	0.000519
rs28469095	rs7946	995	0.010562	0.000112
rs28469095	rs25489	995	0.021493	0.000462
rs7946	rs25489	995	-0.00264	0.000007

LD was calculated as squared Pearson correlation (r^2) between genotype dosage values using pairwise complete observations. LD between rs429358 and rs7412 was estimated separately from *APOE* genotype data ($r^2 \approx 0.006$) and is not included in this table because rs7412 dosage was not part of the five-SNP candidate panel.

Table S7. Allele-frequency comparison of the five candidate variants in the WGS cohort and public reference datasets.

SNP	Compared allele	Present WGS	KRGDB	KGP4K	1000G	ExAC	UK10K	gnomAD Global
rs429358	C	0.2518	0.0917	0.091	0.104	0.171	0.1397	0.158091
rs440446	G	0.4673	0.3829	0.3701	0.578	0.643	0.6543	–
rs28469095	C	0.0688	0.0589	0.0662	0.091	0.128	0.0925	0.091221
rs7946	T	0.2437	0.2669	0.2708	0.556	0.588	0.7519	0.589714
rs25489	T	0.0945	0.1086	0.1005	0.065	–	0.0388	–

Allele frequencies in the WGS cohort were calculated directly from alternate-allele dosage as the total count of the compared allele divided by twice the number of non-missing diploid genotypes. All five candidate variants had complete genotype data in the WGS cohort (n=995). The compared allele is shown explicitly for each variant. External reference frequencies are presented for descriptive population context only and were not used for candidate nomination, association testing, model training, or feature attribution. Cross-dataset comparisons should be interpreted cautiously because reference resources differ in population composition, sampling frame, sequencing platform, release, and allele-frequency estimation procedures.

Table S8A. Genetic and variant-level evidence for exploratory candidate variants.

SNP	Gene	GRCh38 position	VEP consequence	Initial WES	Initial WES	ALT AF, Korean	LD with
				nominal P range	screening OR	WGS	rs429358 (r ²)
rs7946	PEMT	chr17:17506246	Missense	0.040–0.044	1.53	0.2437	0.003
rs25489	XRCC1	chr19:43552260	Missense	0.020–0.021	2.15	0.0945	<0.01

rs440446	APOE- region	chr19:44905910	Transcript-	0.036–0.040	1.49	0.4673	0.317
			dependent; intronic				
			on APOE MANE				
			Select				
rs429358	APOE	chr19:44908684	Missense	0.0049–0.005	2.36	0.2518	Reference

rs28469095	<i>NKPD1</i>	chr19:45152075	Missense	0.040–0.041	2.6	0.0688	<0.01
-------------------	--------------	----------------	----------	-------------	-----	--------	-------

Table S8A summarizes variant-level evidence for the five candidate variants, including GRCh38 genomic position, VEP consequence, WES association statistics, Korean WGS alternate-allele frequency, and cohort-derived linkage disequilibrium with APOE rs429358. Functional consequence annotations were obtained post hoc and were not used for candidate nomination, association testing, model training, or feature selection. WES association estimates are exploratory and should be interpreted together with the covariate-adjusted and multiple-testing-adjusted analyses reported in Supplementary Table S10.

Table S8B. Functional and clinical-genomic context of exploratory candidate variants.

SNP	In silico / functional annotation	GTEx/eQTL annotation	ClinVar status	External AD genetics context	Functional/pathway context	Study interpretation
rs7946	Missense; SIFT deleterious (0.01); PolyPhen possibly damaging (0.55); CADD 19.3	No significant brain-tissue eQTL was detected in the GTEx query performed in this study	No pathogenic AD- specific ClinVar assertion identified	Not established as a primary AD GWAS locus in large-scale AD GWAS	Phosphatidylcholine biosynthesis; phospholipid homeostasis; lipid metabolism relevant to <i>APOE</i> biology	Hypothesis- generating Non-APOE candidate with post hoc functional context; not a confirmed

susceptibility
locus,
independent
predictor, or
causal variant.

		No significant	No	Not	Base-excision DNA	
rs25489	Missense; SIFT tolerated (0.17);	brain-tissue	pathogenic	established as	repair; oxidative	Plausible but
	CADD 13.8	eQTL was	AD-	a primary AD	stress/genome	less AD-specific
		detected in	specific	GWAS locus	maintenance	DNA-repair

the GTEx ClinVar candidate; weak

query assertion model stability

performed in identified

this study

rs440446	Transcript-dependent VEP	Transcript	No			<i>APOE-region</i>
	consequences: intron_variant on	consequence	pathogenic	<i>APOE-region</i>	<i>APOE-region</i>	<i>marker with</i>
	the APOE MANE Select transcript	does not by	AD-	context;	<i>context with</i>	<i>transcript-</i>
	(ENST00000252486/NM_000041.4);	itself establish	specific	requires	<i>transcript-</i>	<i>dependent</i>
	overall most severe consequence	regulatory	ClinVar	LD/haplotype-	<i>dependent</i>	<i>annotation; not</i>
	missense_variant on an alternative	function; eQTL	assertion	aware	<i>functional</i>	<i>interpreted as</i>
	transcript (ENST00000434152;	evidence	identified	interpretation	<i>annotation.</i>	<i>an independent</i>

C>G, p.Asn14Lys); 5' UTR	should be	<i>functional or</i>
consequences on additional	interpreted	<i>susceptibility</i>
alternative transcripts.	with APOE-	<i>variant without</i>
	region caution.	<i>conditional/fine-</i>
		<i>mapping and</i>
		<i>functional</i>
		<i>validation.</i>

	Known <i>APOE</i>	Known	Established	<i>APOE</i> lipid	Primary <i>APOE</i> -
rs429358	Missense; known <i>APOE</i> functional	functional/risk-	<i>APOE</i> -	<i>APOE</i> AD risk	driven genetic
	variant; CADD 16.6	defining	associated	transport; ε4-	
		variant	variant	defining variant	signal and

positive-control

locus

			No		NKPD1 functional	Exploratory
			No significant	Not		
rs28469095	Missense variant; canonical NKPD1	brain-tissue	pathogenic	established as	relevance to AD	<i>NKPD1</i>
	transcript ENST00000686631.1	eQTL was	AD-	a primary AD	remains uncertain;	missense
	(NM_198478.4), p.Arg788Gly; SIFT	detected in	specific	GWAS locus;	the missense	candidate with
	tolerated (0.54); PolyPhen benign	the GTEx	ClinVar	cohort LD	annotation	weak model
	(0).	query	assertion	with rs429358	provides variant-	attribution;
			identified		level context but	statistical

	performed in	was minimal	does not establish	evidence did
	this study	($r^2 < 0.01$).	pathogenicity or	not survive
			disease mechanism.	multiple-testing
				correction and
				functional
				validation is
				required.

Table S8B summarizes post hoc functional and external genetic context for the five candidate variants, including in silico functional predictions, GTEx/eQTL evidence, ClinVar annotations, external Alzheimer's disease genetic evidence, and relevant biological context. These annotations were used only for interpretation and were not incorporated into candidate selection, association testing, or model training. Absence of a significant brain-tissue eQTL in the queried GTEx resource should not be interpreted as evidence of absent biological relevance. Study interpretations are hypothesis-generating and do not establish pathogenicity, causality, independent association, or predictive utility.

Table S9. Exploratory gain-based feature-ranking stability across model configurations.

feature	mean_ rank	sd_ra nk	best_r ank	worst_ rank	mean_ gain	detected_mode l_configs	top20_model _configs	top20_frequency_acro ss_all_configs
num__Amyloid_PET	1	0	1	1	34.568	5	5	0.833

num_SNSB_II_Domain_Mem	2.8	1.303	2	5	17.365	5	5	0.833
ory		84						
num_SVLT_Delayed_recall	3.8	1.303	3	6	16.305	5	5	0.833
		84						
num_SVLT_recall_trial3	6.2	3.346	2	10	11.683	5	5	0.833
		64						
num_SVLT_recall_trial2	7.2	4.324	3	14	10.717	5	5	0.833
		35						
num_SVLT_recall_total_score	9.6	2.302	7	12	8.221	5	5	0.833
		173						

num__Praxis_Ideomotor	15	1.414	13	17	5.694	5	5	0.833
		214						
num__StroopTest_Wordreadi	13.4	7.300	6	23	6.553	5	4	0.667
ng_error		685						
cat__Fist_edge_palm_right_No	14.4	6.804	8	24	5.634	5	4	0.667
rmal		41						
num__Sum_of_boxes	16.8	3.834	13	22	4.798	5	4	0.667
		058						
num__B_ADL	22.4	21.33	9	60	4.837	5	4	0.667
		776						

num_Contrasting_program	14.5	18.33	5	42	8.854	4	3	0.5
		939						
num_RCFT_delayed_recall	15.4	8.142	7	26	5.982	5	3	0.5
		481						
num_CDR	18	3.807	14	23	4.535	5	3	0.5
		887						
cat_Fist_edge_palm_left_Nor	19	5.354	12	25	4.018	4	3	0.5
mal		126						
num_SVLT_recognition_discr	31.2	27.72	4	65	6.077	5	3	0.5
minability_index		544						

num_rs7946	51.167	50.99	4	102	4.504	6	3	0.5
		575						
num_rs429358	15.75	11.17	1	27	3.719	4	2	0.333
		661						
num_K_MMSE_drawing	16.5	0.707	16	17	3.543	2	2	0.333
		107						
cat_Comprehension_Normal	18.333	13.57	10	34	4.456	3	2	0.333
		694						
num_Rey_CFT_copy_score	24.2	5.761	17	29	2.876	5	2	0.333
		944						

cat_<i>APOE</i>_E3/E4	24.4	8.961	14	36	3.318	5	2	0.333
		027						
num_Naming_K_BNT	43.2	33.71	6	78	4.693	5	2	0.333
		498						
num_K_MMSE_total_score	44.6	24.54	18	67	1.67	5	2	0.333
		18						
num_rs28469095	2	NA	2	2	0.403	1	1	0.167
num_rs25489	3	NA	3	3	0.347	1	1	0.167
num_rs440446	59.2	30.93	4	76	0.327	5	1	0.167
		865						

Gain-based feature rankings were summarized across exploratory model configurations and are presented as supplementary descri

ptive analyses of model behavior. Some configurations included interaction terms that were not part of the prespecified primary models; these configurations were therefore not used for candidate nomination, association inference, biological interpretation, or the primary feature-attribution results. Gain-based rankings are not directly comparable with fold-wise validation permutation importance and do not represent genetic association, biological effect size, causal influence, or independent predictive utility.

Table S10. Covariate-adjusted WES candidate-variant association results and multiple-testing correction.

Gene_or_region	SNP	Model_gro up	Primary_mo del	Covariat es	NMIS S	OR CI	95% CI	P	FDR_BH_primary_ 5	Bonferroni_primar y_5	Interpretation
----------------	-----	-----------------	-------------------	----------------	-----------	----------	-----------	---	----------------------	--------------------------	----------------

<i>PEMT</i>	rs7946	Non-APOE candidate	Additive	age, sex, <i>APOE</i> ε4	331	1.43 5	–	0.1363	0.1363	0.6815	Attenuated after
											covariate
											adjustment;
											exploratory candidate only.
<i>XRCC1</i>	rs25489	Non-APOE candidate	Additive	age, sex, <i>APOE</i> ε4	332	1.99 1	–	0.0768 3	0.12805	0.38415	Borderline/attenu
											ated after
											covariate
											adjustment; exploratory candidate only.

APOE-region	rs4404	APOE-region candidate	Additive	age, sex	333	1.385	0.912	0.1265	0.1363	0.6325	APOE-region marker; not interpreted as independent of APOE-related mechanisms.
	46					5	2.102				
APOE	rs4293	APOE-region candidate	Additive	age, sex	332	2.087	1.096	0.0251	0.08703	0.1258	Expected APOE-region positive-control signal; not a novel candidate locus.
	58					7	3.975	6			

<i>NKPD1</i>	rs2846	Non-APOE	Additive	age, sex, <i>APOE</i> ε4	332	3.01	1.082	0.0348	0.08703	0.17405	Nominal									
											exploratory									
											candidate; not									
	9095	candidate									significant after									
												multiple-testing								
												correction.								

Primary additive candidate-variant association results are shown after covariate adjustment. For Non-APOE candidate variants, logistic regression models included age, sex, and *APOE* ε4 carrier status. APOE-region variants were analyzed separately using age and sex adjustment because *APOE* ε4 carrier status is partly determined by *APOE* genotype and may introduce collinearity. Benjamini–Hochberg FDR and Bonferroni-adjusted P values were calculated across the five primary additive candidate tests. These results were interpreted as exploratory candidate-nomination evidence only and not as confirmation of Alzheimer’s disease susceptibility loci.

Table S11. Full fold-wise validation permutation-importance results for the strict leakage-reduced model.

Rank	Feature	Feature_group	Permutation_importance_mean_AUROC_decrease	Permutation_importance_SD	Interpretation
1	Test_Age	Age-related demographic feature	0.086317	0.030968	Descriptive demographic model reliance
2	rs429358	APOE/APOE-region genetic feature	0.075857	0.030636	Known APOE-region model reliance; not novel

Rank	Feature	Feature_group	Permutation_importance_mean_AUROC_decrease	Permutation_importance_SD	Interpretation
3	YearsOfEducation	Education-related demographic feature	0.035303	0.020655	Descriptive demographic model reliance
4	APOE	APOE/APOE-region genetic feature	0.034571	0.017022	APOE proxy model reliance
5	Age	Age-related demographic feature	0.026704	0.015162	Descriptive demographic model reliance
6	Job	Non-diagnostic epidemiologic/physical feature	0.014297	0.011404	Model reliance only
7	rs440446	APOE/APOE-region genetic feature	0.008493	0.011849	APOE-region marker; not interpreted as APOE-independent

Rank	Feature	Feature_group	Permutation_importance_mean_AUROC_decrease	Permutation_importance_SD	Interpretation
8	Sex	Sex demographic feature	0.005724	0.010562	Descriptive demographic model reliance
9	rs25489	Non-APOE candidate SNP	0.002302	0.005959	Exploratory model-reliance signal only
10	rs7946	Non-APOE candidate SNP	0.002066	0.006520	Exploratory model-reliance signal only
11	Education_Level	Education-related demographic feature	0.001893	0.005285	Descriptive demographic model reliance
12	Handgrip	Non-diagnostic epidemiologic/physical feature	-0.000063	0.000563	Model reliance only

Rank	Feature	Feature_group	Permutation_importance_mean_AUROC_decrease	Permutation_importance_SD	Interpretation
13	rs2846909	Non-APOE candidate S	-0.000321	0.005368	Near-zero exploratory mod
	5	NP			el reliance
14	Education	Education-related dem	-0.004250	0.007463	Descriptive demographic m
		ographic feature			odel reliance

Permutation importance was computed only in outer validation folds and aggregated across five folds and 50 repeated permutations per fold. Ages was excluded as a redundant age-band variable.

Table S12. Candidate-variant rationale and claim boundaries.

Variant	Gene	Candidate_role	Selection_rationale	Primary_interpretation	Claim_boundary
	/region				
rs429358	APOE	Positive-control	Included because APOE-region	Known-region contribution;	No novelty or
		APOE-region	variation is an established	not interpreted as a novel	causality claim.
		variant	contributor to AD risk and	discovery.	
			provides an internal		
			benchmark.		

rs440446	APOE-region	APOE-region	Included as an APOE-region	APOE-region marker; not	No claim of
		marker	candidate marker for	interpreted as independent of	independence from
			exploratory evaluation with	APOE-related mechanisms.	APOE-related
			<i>APOE</i> rs429358.		mechanisms.

rs7946	<i>PEMT</i>	Exploratory Non-	Nominated from WES	Initial nominal signal	Not a confirmed
		APOE candidate	screening and carried forward	attenuated after age, sex, and	susceptibility locus;
			for exploratory model-based	<i>APOE</i> ε4 adjustment; retained	not an independent
			evaluation.	only as hypothesis-	association claim.
				generating.	

rs25489	XRCC1	Exploratory Non-	Nominated from WES	Borderline/attenuated after	Not a confirmed
		APOE candidate	screening and carried forward for exploratory model-based evaluation.	covariate adjustment; interpreted as exploratory.	susceptibility locus; no causal inference.
rs28469095	NKPD1	Exploratory Non-	Nominated from WES	Nominal exploratory	Not a confirmed
		APOE candidate	screening and retained nominal evidence after covariate adjustment but did not survive multiple-testing correction.	candidate requiring replication.	susceptibility locus; no clinical utility claim.

The table summarizes the rationale for carrying forward each nominated candidate variant and the corresponding interpretive bou

ndary applied in this study. APOE-region variants were treated as known-region or APOE-region benchmark features, whereas *PEMT*, *XRCC1*, and *NKPD1* variants were treated as exploratory Non-APOE candidates. Candidate status does not imply corrected statistical significance, causal inference, biological validation, clinical utility, or confirmation of Alzheimer’s disease susceptibility loci.

Table S13. WES candidate-screening and PLINK genotype quality-control summary.

Item	Value
WES candidate-nomination cohort enrolled	346 participants

aMCI	263
------	-----

Controls	83
----------	----

PLINK BIM variant entries before filtering	142,740
--	---------

Candidate variants present in BIM	5/5
-----------------------------------	-----

PLINK QC filters	--mind 0.10, --geno 0.10, --maf 0.01, --hwe 1e-6
------------------	--

PLINK QC-filtered analytic samples	332
------------------------------------	-----

PLINK QC-filtered variants	53,858
----------------------------	--------

Candidate minor allele frequency range	0.0661–0.4189
--	---------------

Candidate HWE P-value range	0.0644–0.8937
-----------------------------	---------------

Candidate genotype missingness range	0.0263–0.0292
Candidate genotype call-rate range	97.08%–97.37%
Annotated WES screening records used for candidate review	5,215
PASS variants in annotated report	4,753
Candidate variants carried forward	5
Primary corrected statistical tests	5 additive candidate-variant tests
Correction applied	Benjamini–Hochberg FDR and Bonferroni
Covariates for Non-APOE candidates	age, sex, <i>APOE</i> ε4 carrier status

Covariates for APOE-region variants	age, sex
Ancestry principal components	External reference-projected PCs unavailable; internal WES genotype-derived PC-adjusted sensitivity analysis reported in Supplementary Table S14
Interpretation	exploratory candidate nomination only

The WES cohort included 346 enrolled participants. Candidate-level PLINK analyses were performed using the available genotype dataset, in which all five carried-forward candidate variants were present. The PLINK BIM file contained 142,740 variant entries before filtering. After standard genotype-level filtering, the candidate-level analytic dataset included 332 samples and 53,858 variants. Across the five candidate variants, MAF ranged from 0.0661 to 0.4189, HWE P values ranged from 0.0644 to 0.8937, and genotype call rates ranged from 97.08% to 97.37%. Corrected statistical interpretation was restricted to the five primary additive candidate-variant tests reported in Supplementary Table S10. These results were not interpreted as exome-wide discovery, corrected genome-wide association, causal inference, or confirmat

ion of Alzheimer's disease susceptibility loci. External reference-projected ancestry PCs were not available; therefore, residual population structure was acknowledged as a limitation. Internal genotype-derived PC-adjusted sensitivity analyses are reported separately in Supplementary Table S14.

Table S14. Internal principal-component-adjusted sensitivity analysis of WES candidate-variant associations.

Variant	Gene/reg ion	Mode l	Covariates	N	OR	95% CI	P	FDR_ BH	Bonferro ni P	Interpretation
rs7946	PEMT	Additi ve	age, sex, APOE ε4, P C1–PC5	33 1	1.66 1	0.996–2. 771	0.05 19	0.101 3	0.2595	Borderline nominal exploratory Non-APOE candi date; not corrected
rs25489	XRCC1	Additi ve	age, sex, APOE ε4, P C1–PC5	33 1	1.78 0	0.833–3. 806	0.13 69	0.171 1	0.6844	Exploratory Non-APOE candidate; not corrected

Variant	Gene/reg ion	Mode l	Covariates	N	OR	95% CI	P	FDR_ BH	Bonferro ni P	Interpretation
rs28469	<i>NKPD1</i>	Additi	age, sex, <i>APOE</i> ϵ 4, P	33	2.69	0.956–7.	0.06	0.101		Borderline nominal exploratory candidate; not c
095		ve	C1–PC5	1	2	578	08	3	0.3038	orrected
rs42935	<i>APOE</i>	Additi	age, sex, PC1–PC5	33	1.98	1.043–3.	0.03	0.101		<i>APOE</i> -region positive-control context; not correc
8		ve		1	5	776	67	3	0.1836	ted
rs44044	<i>APOE</i> -reg	Additi	age, sex, PC1–PC5	33	1.32	0.866–2.	0.19	0.195		<i>APOE</i> -region marker; not <i>APOE</i> -independent
6	ion	ve		1	2	018	55	5	0.9776	

Internal genotype-derived principal components were calculated from QC-filtered, LD-pruned autosomal WES variants after excluding the *APOE*-region and candidate loci. PCA was performed using 26,479 LD-pruned variants with a total genotyping rate of 0.9957. PC-adjusted

candidate-level logistic regression was then performed using the first five internal WES principal components. For Non-APOE candidate variants, models included SNP dosage, age, sex, *APOE* ϵ 4 carrier status, and PC1–PC5. APOE-region variants were analyzed using SNP dosage, age, sex, and PC1–PC5 to avoid collinearity with *APOE* ϵ 4 carrier status. No candidate variant survived candidate-panel FDR or Bonferroni correction. These analyses were interpreted as internal population-structure sensitivity analyses and not as external ancestry validation.

Table S15. WGS candidate-variant logistic regression sensitivity analysis.

Variant	Gene/region	Model	Covariates	N	OR	95% CI	P	FDR_BH	Bonferroni P	Interpretation
rs7946	PEN1	Additive	age, sex, rs429358	995	0.974	0.782–1.214	0.8161	0.8161	1.0000	Exploratory Non-APOE candidate; no WGS association evidence after rs429358 adjustment
rs2548	XRCC1	Additive	age, sex	99	1.3	0.972–	0.0740	0.1234	0.3701	Exploratory Non-APOE candidate; not correct

9		ve	x, rs429	5	40	1.848					ed
			358								
rs2846	NKPD1	Additi	age, se	99	1.1	0.776–	0.5507	0.6883	1.0000		Exploratory Non-APOE candidate; no WGS as
9095		ve	x, rs429	5	17	1.607					sociation evidence after rs429358 adjustment
			358								
rs4293	APOE	Additi	age, sex	99	2.9	2.340–	2.14×	1.07×1	1.07×10 ⁻¹⁷		APOE-region positive-control signal
58		ve		5	91	3.824	10 ⁻¹⁸	0 ⁻¹⁷			
rs4404	APOE-re	Additi	age, sex	99	1.2	1.066–	0.0086	0.0216	0.0431		APOE-region marker; not interpreted as APO
46	gion	ve		5	85	1.549					E-independent

nosis), sex, and APOE rs429358 dosage. APOE-region variants were analyzed using SNP dosage, Age (age at diagnosis), and sex. Genotype missingness was 0% for all five candidate variants in the WGS dataset. Benjamini–Hochberg FDR and Bonferroni-adjusted P values were calculated across the five WGS candidate-variant tests. Non-APOE candidate variants did not show corrected WGS association evidence after rs429358 adjustment. *APOE* rs429358 showed the expected positive-control association, and rs440446 was interpreted as an APOE-region marker rather than an APOE-independent locus. These analyses were interpreted as exploratory candidate-level sensitivity analyses and not as validation of genetic association, causality, or clinical utility.

Table S16. XGBoost hyperparameters and reproducibility settings.

Parameter / setting	Reported value	Purpose / note
objective	binary:logistic	Binary Alzheimer’s disease classification
eval_metric	logloss	XGBoost internal evaluation metric; AUROC was computed separately for model-p

performance reporting

n_estimators	500	Number of boosting rounds/trees
max_depth	3	Maximum tree depth
learning_rate	0.03	Boosting shrinkage parameter
subsample	0.8	Row subsampling fraction
colsample_bytree	0.8	Column subsampling fraction per tree
min_child_weight	5	Minimum child weight
gamma	Not explicitly set; XGBoost default used	Minimum split-loss reduction
reg_alpha	Not explicitly set; XGBoost default used	L1 regularization term

reg_lambda	1.0	L2 regularization term
tree_method	Not explicitly specified in the analysis code; XGBoost default used	XGBoost default tree construction method used by the installed version
n_jobs	-1	Use available CPU threads
missing values	Native XGBoost missing-value handling	Sentinel missing codes were converted to missing before modeling
categorical variables	One-hot encoded within each training fold	Encoder fitted on training-fold data only and applied to the corresponding validation fold; unseen categories handled without refitting
data split	Stratified five-fold cross-validation with pooled OOF predictions	Each participant receives one prediction from a model not trained on that participant; identical fold assignments across model configurations

cross-validation	5 outer stratified folds; fixed hyper parameters	Preprocessing and one-hot encoding fitted within each training fold only; validation-fold outcomes not used for fitting or preprocessing
random seed	42	Fixed for reproducibility
classification thres	0.50 (prespecified)	Used for accuracy, sensitivity, specificity, PPV, NPV, and confusion matrices; not optimized
hold		on OOF validation outcomes
AUROC/AUPRC confidence intervals	co 2,000 class-stratified nonparametric bootstrap resamples of pooled OOF predictions	Percentile 2.5th and 97.5th limits; intervals summarize uncertainty in the fixed pooled OOF predictions
permutation importance	Outer validation folds only; 50 repeats per fold	Aggregated across five folds (250 estimates per feature)
global exclusions	ID; FAMILY_ID; PNTTM; Ages	Identifiers/procedural fields excluded; Ages excluded as a redundant age-band variable

The same fixed XGBoost configuration is used for the prespecified candidate SNP-only, full clinical–genomic, strict leakage-reduced, strict leakage-reduced

ed without rs429358/APOE proxy, and Non-APOE SNP-only model configurations. All models use the same stratified five-fold OOF partition. Preprocessing and one-hot encoding are fitted separately within each training fold and applied to the corresponding validation fold. The probability threshold is pre-specified at 0.50. AUROC and AUPRC confidence intervals are estimated from pooled OOF predictions using 2,000 paired class-stratified bootstrap resamples. Strict-model permutation importance is calculated only on outer validation folds.