

RESEARCH ARTICLE

Harmonizing genotype array data to understand genetic risk for brain amyloid burden in the AMYPAD PNHS Consortium

Emma S. Lockett^{1,2,3,4} | Yasmina Abakkouy⁴ | Luigi Lorenzini^{1,2} | Lyduine E. Collij^{1,2,5} | David Vallez Garcia^{1,2,6} | Pieter Jelle Visser^{7,8,9,10} | Anouk den Braber^{7,8,11} | Craig Ritchie^{12,13} | Mercè Boada^{14,15} | Patricia Genius^{6,16,17} | Natàlia Vilor-Tejedor^{6,16,17,18} | Juan Domingo Gispert^{6,15,16,19} | Rik Vandenberghe^{3,20} | Frederik Barkhof^{1,21} | Isabelle Cleynen⁴ | on behalf of the AMYPAD Consortium

¹Department of Radiology and Nuclear Medicine, Amsterdam UMC location VUmc, Amsterdam, the Netherlands

²Amsterdam Neuroscience, Brain Imaging, Amsterdam, the Netherlands

³Laboratory for Cognitive Neurology, KU Leuven, Leuven, Belgium

⁴Laboratory for Complex Genetics, KU Leuven, Leuven, Belgium

⁵Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Faculty of Medicine, Lund University, Lund, Sweden

⁶Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Carrer de Wellington, Barcelona, Spain

⁷Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

⁸Amsterdam Neuroscience, Neurodegeneration, Amsterdam, the Netherlands

⁹Alzheimer Center Limburg, School for Mental Health and Neuroscience, Maastricht University, Maastricht, the Netherlands

¹⁰Department of Neurobiology, Care Sciences and Society, Division of Neurogeriatrics, Karolinska Institutet, Stockholm, Sweden

¹¹Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

¹²Scottish Brain Sciences, Redheughs Rigg, South Gyle, UK

¹³University of St. Andrews, School of Medicine, North Haugh, St. Andrews, UK

¹⁴Ace Alzheimer Center Barcelona, Universitat Internacional de Catalunya, Barcelona, Spain

¹⁵Networking Research Center on Neurodegenerative Diseases (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain

¹⁶Neurosciences Programme, IMIM - Hospital del Mar Medical Research Institute, Barcelona, Spain

¹⁷Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain

¹⁸Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands

¹⁹Departament de Medicina i Ciències de la Vida, Universitat Pompeu Fabra, Ciutat Vella, Spain

²⁰Alzheimer Research Centre, Leuven Brain Institute, Leuven, Belgium

²¹Queen Square Institute of Neurology and Centre for Medical Image Computing, University College London, London, UK

Correspondence

Emma S. Lockett, Amsterdam UMC location VUmc Amsterdam, Department of Radiology and Nuclear Medicine, De Boelelaan 1117, 1081 HV, Amsterdam, the Netherlands.
Email: e.s.lockett@amsterdamumc.nl

Abstract

INTRODUCTION: We sought to harmonize genotype data from the predementia AMYPAD (Amyloid Imaging to Prevent Alzheimer's Disease) Consortium, compute

Emma S. Lockett and Yasmina Abakkouy contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Alzheimer's & Dementia* published by Wiley Periodicals LLC on behalf of Alzheimer's Association.

AMYPAD Consortium data used in the preparation of this article were obtained from the Prognostic and Natural History Study (PNHS), provided by the Amyloid Imaging to Prevent Alzheimer's Disease Consortium (AMYPAD). As such, investigators within the AMYPAD PNHS and AMYPAD Consortium contributed to the design and implementation of AMYPAD and/or provided data but did not participate in the analysis or writing of this report. A complete list of AMYPAD investigators can be found at <https://doi.org/10.5281/zenodo.7962737>

Funding information

Innovative Medicines Initiative, Grant/Award Number: No 115952; European Union's Horizon 2020 research; innovation programme; GE Healthcare; Springer Healthcare; MSCA, Grant/Award Number: #101108819; Alzheimer Association Research Fellowship, Grant/Award Number: #23AARF-1029663; Spanish Research Agency, Grant/Award Numbers: MICIU/AEI/10.13039/501100011033, RYC2022-038136-I; European Union FSE+, Grant/Award Number: PID2022-143106OA-I00; European Union FEDER; Alzheimer's Disease Data Initiative; 23S06083-001; Stichting Alzheimer Onderzoek; Internal Funds KU Leuven; Flemish Research Foundation, Grant/Award Numbers: GOG1519N, G094418N; VLAIO, Grant/Award Number: HBC.2019.2523; NIHR biomedical research centre

polygenic risk scores (PRS), and determine their association with global amyloid deposition.

METHODS: Genetic data from five AMYPAD parent cohorts were harmonized, and PRS were computed for Alzheimer's disease (AD) susceptibility, cerebrospinal fluid (CSF) amyloid beta ($A\beta$)42, and CSF phosphorylated tau181. Cross-sectional amyloid (Centiloid [CL]) burden was available for all participants, and regression models determined if PRS were associated with CL burden.

RESULTS: After harmonization, data for 867 participants showed that high CL burden was most strongly predicted by CSF $A\beta$ 42 PRS compared to traditional AD susceptibility PRS.

DISCUSSION: This work emphasizes the importance of data harmonization and pooling of cohorts for large-powered studies. Findings suggest a genetic predisposition to amyloid pathology that may predispose individuals early in the AD continuum. This validates the potential use of PRS in clinical (trial) settings as a non-invasive tool to assess AD risk.

KEYWORDS

Alzheimer's disease, amyloid, Amyloid Imaging to Prevent Alzheimer's Disease, genotype data harmonization, polygenic risk scores, predementia

Highlights

- We developed a robust harmonization pipeline for multi-cohort genotype array data.
- Cerebrospinal fluid amyloid beta ($A\beta$)-specific polygenic risk scores (PRS) more strongly predicted global $A\beta$ positron emission tomography burden than other PRS.
- Results suggest a strong genetic predisposition to early $A\beta$ pathology.
- This work highlights the need for robust data harmonization and data pooling.
- This work also validates the potential use of PRS as a non-invasive tool to assess Alzheimer's disease risk.

1 | BACKGROUND

Sporadic Alzheimer's disease (AD) is a complex heterogeneous disease influenced by both genetic and environmental risk factors, which contribute to its clinical manifestations.¹ Despite substantial advances in identifying modifiable risk factors, the biological heterogeneity driving the disease remains partially unexplored. Unravelling the genetics of AD offers opportunities to enhance the precision of clinical trial methodologies and results, and advance personalized medicine initiatives. An enhanced understanding of the genetic landscape, particularly across the AD continuum (from preclinical to clinical stages) and of its dynamic endophenotypes (such as amyloid and tau accumulation) in the earliest disease stage will provide further information regarding the pathophysiological mechanisms occurring in AD. In turn, this may allow for the generation of genetic profiles that can identify individuals in the early preclinical phase. Moreover, it can allow for delineating susceptible subpopulations of individuals or AD subtypes, thereby enriching our understanding of AD heterogeneity, and deter-

mining individuals suitable for prevention trials as part of a stratified recruitment process.

In this context, it is of utmost importance to have endeavors to enable large-powered studies to be performed. This in turn will likely facilitate the detection of additional AD risk variants or loci, and those with low minor allele frequency (MAF) or low effect size. Multi-center collaborative studies provide an opportunity for pooling data and generating large datasets. However, genetic data sources are heterogeneous, and harmonization processes must be implemented to enable data utility, especially when sharing large-scale data with the broader scientific community. The Amyloid Imaging to Prevent Alzheimer's Disease (AMYPAD) Prognostic and Natural History Study (PNHS) is a notable example of such an initiative.²⁻⁴ This pan-European collaboration comprises 10 parent cohorts, in which all individuals were older than 50 years without a dementia diagnosis at inclusion. Furthermore, the study was designed to acquire amyloid positron emission tomography (PET) and magnetic resonance imaging (MRI) scans of individuals captured over all stages of the AD risk spectrum (negative, gray zone,

and positive AD biomarker profiles). Thus, this cohort represents a large, deeply phenotyped, heterogeneous population that is well suited to study the early genetic determinants of AD, with higher statistical power than other available smaller studies of similar populations.

The aim of this study was 2-fold. First, we present a detailed methodology for the harmonization and genetic characterization of diverse AMYPAD PNHS parent cohorts to establish a large cohort of individuals with genetic, demographic, and imaging data. After this, as an application for the harmonized dataset, we computed polygenic risk scores (PRSs) using summary statistics from the Kunkle et al.⁵ case-control genome-wide association study (GWAS), and the Jansen et al.⁶ cerebrospinal fluid (CSF) amyloid beta 42 (A β 42) and CSF phosphorylated tau 181 (p-tau181) GWAS⁶ to assess the association of PRS and amyloid PET burden along the AD risk continuum.

2 | METHODS

2.1 | Study sample

The AMYPAD PNHS (EU Clinical Trials Register AMYPAD-02 EudraCT Number 2018-002277-22) is a pan-European multicenter study population of non-demented (Clinical Dementia Rating [CDR] ≤ 0.5) older adults ≥ 50 years old at inclusion. For full recruitment and study details, see Lopes Alves et al.² and Bader et al.³ Briefly, participants were recruited from 10 parent cohorts across seven countries with similar characteristics. These parent cohorts include: (1) the European Prevention of Alzheimer's Disease Longitudinal Cohort Study (EPAD LCS), (2) the twins subset of the European Medical Information Framework for Alzheimer's Disease 60++ (EMIF-AD 60++), (3) EMIF-AD 90+, (4) the Alzheimer's and Families study (ALFA+, Spain), (5) the Fundació ACE Healthy Brain Initiative (FACEHBI, Spain), (6) the Flemish Prevent AD Cohort KU Leuven (F-PACK, Belgium), (7) the Université Catholique de Louvain (UCL-2010-412 cohort, Belgium), (8) the Microbiota cohort (Switzerland), (9) the AMYPAD Diagnostic Patient Management Study (DPMS, VUmc only, Netherlands), and (10) the DZNE-Longitudinal Cognitive Impairment and Dementia Study (DELCODE, Germany).

As part of recruitment into the AMYPAD PNHS, participants received a static or dual-window amyloid PET scan of either [18F]Flutemetamol or [18F]Florbetaben, as well as a structural MRI scan. Because participants were recruited from existing parent cohorts, data collection and availability were comparable across parent cohorts, which mainly included historical PET and MRI scans, neuropsychological assessment, and fluid biomarkers. Furthermore, genotyping data were also available for a subset of parent cohorts, forming the subgroup of individuals for the present study: EPAD LCS, ALFA, EMIF-AD 60++, F-PACK, and FACEHBI. Note that genetic data from all participants within EPAD LCS, ALFA, and F-PACK were harmonized prior to subsetting based on AMYPAD participation and amyloid PET scan availability. ALFA participants included in the final AMYPAD PNHS subset were part of the ALFA+ subset of ALFA.

RESEARCH IN CONTEXT

- Systematic review:** Few studies describe harmonization of multi-cohort genotype array data for Alzheimer's disease (AD) to assess polygenicity. Although many polygenic risk score (PRS) studies exist in the context of AD, few explore PRS beyond those computed using AD susceptibility genome-wide association study data, particularly in large predementia cohorts.
- Interpretation:** We present a robust pipeline to harmonize genotype array data from the predementia Amyloid Imaging to Prevent Alzheimer's Disease (AMYPAD) Consortium that can be applied to other multi-cohort studies. Global amyloid beta (A β) positron emission tomography burden was significantly associated with computed cerebrospinal fluid A β -specific PRS, suggesting a strong genetic predisposition to early A β pathology, validating the potential use of PRS in clinical (trial) settings as a primary non-invasive tool to assess AD risk.
- Future directions:** Expanding AMYPAD with other cohorts to further increase power will be essential to discover underlying disease mechanisms, as well as investigating PRS associations with longitudinal A β trajectories and other risk factors, eventually leading to advancements in precision medicine.

2.2 | Neuroimaging data acquisition and harmonization

For each participant, an amyloid PET scan was acquired using either [18F]Flutemetamol or [18F]Florbetaben (90 minutes post-injection), and a T1-weighted MRI. A PET harmonization protocol was applied to allow for the comparability of derived PET metrics.⁷ PET images were processed using IXICO's in-house fully automated MR-based PET pipeline to produce Centiloids (CL) in the Global Alzheimer's Association Interactive Network (GAAIN, <http://www.gaain.org/centiloid-project>) cortical region of interest using the whole cerebellum as the reference region.⁸

Since individuals in the AMYPAD PNHS were scanned for amyloid PET at inclusion, individuals were stratified by low amyloid burden (CL < 10), gray zone amyloid burden (10 < CL < 30), and high amyloid burden (CL > 30).⁹

2.3 | DNA extraction and genotyping

Whole-blood samples were collected for DNA at each parent cohort site, per local procedures, and DNA was extracted according to standard protocols. For details on procedures from each parent cohort, see de Rojas et al.,¹⁰ Vilor-Tejedor et al.,¹¹ Adamczuk et al.,^{12,13} Bos

et al.,¹⁴ and Access | EPAD.¹⁵ Genome-wide genotyping was performed on all extracted DNA for each cohort or batch separately. For ALFA this was done in four batches using the Illumina Infinium Neuro Consortium (Neurochip) Array version 1.0 (batch 1 and batch 2) and version 1.2 (batch 3 and batch 4, Illumina Inc.).¹¹ EPAD LCS, F-PACK (batch 1 and batch 2), and EMIF-AD 60++ were genotyped using the Illumina Infinium Global Screening Array with Shared Custom Content (Illumina Inc.).¹⁶ For FACEHBI, genotyping was performed using the Axiom 815K Spanish biobank array (Thermo Fisher).⁸ Genotype calling for ALFA, EPAD LCS, F-PACK, and EMIF-AD 60++ was performed on the raw intensity data with the Illumina GenomeStudio 2.0 software^{11,16} and using the Affymetrix power tool 1.15.0 for FACEHBI.¹⁰

2.4 | Genetic data quality control

Quality control (QC) was performed independently on each cohort or batch, unless otherwise specified, using PLINK (version 1.9, www.cog-genomics.org/link/1.9).¹⁷ The following steps were undertaken to ensure the accuracy and reliability of genetic data.

First, duplicate single nucleotide polymorphisms (SNPs) were removed based on a chromosome:basepair identifier, followed by the removal of indels, monomorphic, and mitochondrial variants. Next, MAF were calculated based on extracting overlapping SNPs from the European Phase 3 of the 1000 Genomes Project (1000G, $N = 503$).¹⁸ A/T and G/C variants with a high MAF (≥ 0.45) were removed, and those below this threshold were flipped to correct for potential strand misalignments. Then, bcftools was used to remove any ambiguous SNPs, and flip and swap alleles to align data to the positive strand of the reference human genome assembly GRCh37/hg19.¹⁹ After these steps for F-PACK, both batches were merged, given the small number of participants in batch 2.

Next, we performed standard QC as detailed in Luckett et al.²⁰ and Marees et al.²¹ These steps included removing SNPs and individuals with a low call rate (≤ 0.98); removing mismatched sex samples (genetically determined sex from the X chromosome versus reported demographic sex: individuals a priori determined as females F statistic < 0.2 , and individuals a priori determined as males F statistic > 0.8); removing SNPs with low MAF (≤ 0.01); removing SNPs deviating from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$); removing individuals with an outlying heterozygosity rate (± 5 standard deviations); and checking for related or duplicated samples using identity by descent (PI-HAT > 0.2 indicating at least second-degree relatives). All related individuals were retained, but duplicate samples were removed (PI-HAT close to 1). Last, we performed a principal component analysis after merging with the Phase 3 1000G to detect any ethnic outliers (1000G $N = 2504$).¹⁸

2.5 | Genetic data imputation, harmonization, and integration

Following the above-mentioned QC procedures, each parent cohort or batch underwent pre-imputation QC, during which variant call

format (VCF) files were generated for each individual chromosome. Each cohort or batch was imputed independently using the TOPMed Imputation Server (<https://imputation.biodatacatalyst.nhlbi.nih.gov/>), TOPMed r3 panel (all populations), and Eagle phasing (version 2.4).²² After imputation, data were first filtered with imputation information score > 0.7 and MAF ≥ 0.01 . Next, SNP and sample missingness were assessed (≤ 0.98), Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) was applied, and any duplicate SNPs were removed.

Batches and cohorts were integrated to generate a single harmonized dataset. Before merging, datasets first had mono- and multi-allelic alleles removed if PLINK flagged these as present. For each pair of datasets to be merged, a random sample of overlapping SNPs was checked to ensure that the locations matched. All overlapping SNPs were subsequently extracted from each dataset, and the merge-mode 2 command in PLINK was used to merge the datasets. This command was chosen such that genotype calls in dataset 2 were included if they were missing in dataset 1 for a given individual. For ALFA, batch 1 was merged with batch 2, then with batch 3, and finally with batch 4. Parent cohorts were merged with this ALFA dataset, including a new parent cohort each time: F-PACK, FACEHBI, EPAD LCS, and EMIF-AD 60++. Post-merging, ambiguous SNPs were removed, and identity by descent was checked to note related samples (PI-HAT > 0.2) and to remove duplicates (PI-HAT close to 1). To note, after merging with EPAD LCS, if there was a duplicate sample with an EPAD LCS ID and another parent cohort ID, the EPAD LCS sample was removed, as each parent cohort provided data to the AMYPAD PNHS under the original parent cohort ID and not the EPAD LCS ID. Finally, the EMIF-AD 60++ participants were integrated with the merged dataset, and all twin pairs were retained for further steps. Once all datasets had been integrated, the standard QC steps were performed for a final check of outlying heterozygosity (± 5 standard deviations), SNP and individual missingness (≤ 0.98), and MAF ≤ 0.01 prior to performing a principal component analysis to generate genetic principal components.

2.6 | PRS calculations

PRS were computed using PRSice-2²³ using summary statistics from the Kunkle et al.⁵ case-control GWAS, and the Jansen et al. CSF A β 42 and CSF p-tau181 GWAS.⁶ Before calculating PRS using the Jansen et al. summary statistics, the AMYPAD PNHS dataset genome build was lifted from GRCh37 to GRCh38 to match that of the summary statistics using the LiftOver tool (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>). The European individuals from 1000G minus the Fins ($N = 404$) were used as an external reference panel for clumping (clumping window = 250 kilobases, $r^2 = 0.1$). We computed genome-wide PRS with (PRS_{amyloid}, PRS_{tau}, and PRS_{Kunkle}) and without (PRS_{amyloid-noAPOE}, PRS_{tau-noAPOE}, and PRS_{Kunkle-noAPOE}) the apolipoprotein E (APOE) region (19:44.4-45.5Mb), for each set of summary statistics at three thresholds for SNP inclusion (pT): 5×10^{-8} , 1×10^{-5} , and 0.1. For each score, the following approach was used: $PRS_j = (\sum_i S_{ij} G_{ij}) / M_j$ (where j is each PRS, i is the individual, S is the effect size for the reference allele, G is the number of reference alleles observed, and M_j is the number of SNPs included). All PRS

were adjusted for the first five genetic principal components and then standardized against 1000G minus the Fins.

Note that a lower PRS_{amyloid} value denotes a higher genetic predisposition, as this refers to genetically predicted lower levels of CSF A β 42.

2.7 | Statistical analysis

Statistical analyses were performed in R version 4.4.0 (2024-04-24; The R Foundation for Statistical Computing; <https://cran.r-project.org/>). Shapiro–Wilk tests were used to determine data normality, and Bonferroni correction was applied with $\alpha^* < 0.05$ considered significant.

Prior to analysis, related individuals were removed, so the dataset did not contain pairs of individuals with a PI-HAT > 0.2. Cohort descriptors were then assessed between low, gray zone, and high amyloid burden groups using Kruskal–Wallis and post hoc Dunn tests for continuous data, and χ^2 and pairwise proportion tests for categorical data.

PRS distributions were visually assessed by constructing density plots, and the SNP set size per PRS using a bar chart with a log scale to allow for the visualisation of all scores on a single axis. The variability of each PRS was assessed by calculating the interquartile range. Last, a correlation matrix was constructed, and Spearman correlations between each pair of PRS were performed.

As an application for the computed PRS, we assessed their association with global amyloid burden. First, PRS distributions were evaluated across three distinct amyloid burden groups—low, gray zone, and high—using Kruskal–Wallis tests to identify overall differences, where post hoc pairwise comparisons were conducted using Dunn tests to assess specific group differences. The analyses across amyloid burden groups were further refined by stratifying participants based on their APOE ϵ 4 status (carriers vs. non-carriers). Statistical significance and the robustness of these findings were assessed using 1000 bootstrap replications to estimate confidence intervals, ensuring the stability of our results across multiple samples.

Next, we performed a series of linear regressions to assess the association between each PRS and global amyloid burden (CL as a continuous variable), adjusting for chronological age, sex, and years of education. For the primary models, PRS was the main predictor. To explore the influence of APOE ϵ 4 status, we introduced it as an additional covariate for PRS_{noAPOE} . We further assessed interaction effects between PRS_{noAPOE} and APOE ϵ 4 status to determine whether genetic impacts on amyloid burden varied by APOE ϵ 4 status. Linear regression results are reported by means of standardized betas ($\beta_{\text{standardized}}$), the Bonferroni-corrected P values, and the model-adjusted R^2 . To determine the influence of sample size, the primary models were performed for each AMYPAD PNHS parent cohort individually, as well as for the harmonized AMYPAD PNHS cohort.

Finally, PRS were split into low-, medium-, and high-risk tertiles, and the global amyloid burden (CL as a continuous variable) was assessed between these groups using Kruskal–Wallis tests to identify overall dif-

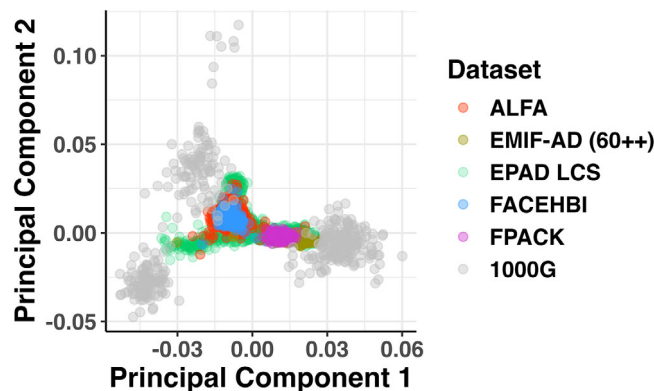


FIGURE 1 Principal component analysis plot including parent cohorts and European individuals minus the Fins from 1000G. Each cohort is represented by a different color with 1000G in gray. For visualization purposes, the AMYPAD PNHS data were projected onto the principal component analysis space from the 1000G dataset after post-imputation QC was performed prior to the calculation of PRS. 1000G, 1000 Genomes Project; ALFA+, Alzheimer's and Families study; AMYPAD, Amyloid Imaging to Prevent Alzheimer's Disease consortium; EMIF-AD 60++, European Medical Information Framework for Alzheimer's Disease 60++; EPAD LCS, European Prevention of Alzheimer's Disease Longitudinal Cohort Study; FACEHBI, Fundació ACE Healthy Brain Initiative; FPACK, Flemish Prevent AD Cohort KU Leuven; PNHS, Prognostic and Natural History Study; PRS, polygenic risk score; QC, quality control.

ferences and post hoc pairwise comparisons using Dunn tests to assess specific group differences. Then, logistic regression was applied using the PRS risk groups to predict high amyloid (CL > 30) with three contrasts of interest: high risk versus low risk; high risk versus medium risk; medium risk versus low risk. Results are reported by means of odds ratios (ORs) and the Bonferroni-corrected P values.

3 | RESULTS

3.1 | Genetic data harmonization and integration

After genotyping, there were $\geq 486,137$ SNPs present in each cohort or batch, with a total of 5511 participants. After the data preparation QC steps, there were $\geq 395,932$ SNPs remaining for standard QC, with all individuals retained at this stage. Following standard QC, there were $\geq 266,323$ SNPs remaining, with an overall loss of 420 individuals due to: high individual-level SNP missingness $N = 311$, sex discrepancy $N = 36$, heterozygosity $N = 41$, duplicate samples $N = 1$, and ethnic outliers $N = 31$. See Figure 1 for a principal component analysis plot with the integrated datasets and European individuals from 1000G (minus the Fins) for illustrative purposes, and Figure S1 in supporting information for individual principal component analyses performed with each cohort or batch with all ethnicities from 1000G. After the pre-imputation QC, there were 5088 participants and $\geq 262,018$ SNPs available for imputation. After filtering the imputed data, there were $\geq 7,792,995$ SNPs available for cohort integration. For more details on

the number of SNPs present in each cohort for each of the steps, see Table S1 in supporting information.

Before integrating all parent cohorts, ALFA batches were first merged to generate a single ALFA cohort, resulting in a dataset of 7,232,606 SNPs and 2513 individuals. The ALFA dataset was merged with F-PACK (6,652,627 variants and 2646 individuals), FACEHBI (6,584,287 variants and 2856 individuals), EPAD LCS (6,575,222 variants and 4492 participants), and EMIF-AD 60+, resulting in a final dataset totalling 6,383,175 variants and 4690 participants. Of these individuals, 957 had an amyloid PET scan as part of the AMYPAD PNHS. Of those individuals from ALFA, only those from the ALFA+ subset were included here. Last, for pairs of individuals that had a PI-HAT > 0.2 from these 957, one individual was removed to ensure no related samples were present, resulting in a final dataset of 867 individuals for further analyses.

3.2 | Participant demographics

Participants were predominantly cognitively unimpaired (CDR = 0, 85.1%), had a low amyloid PET burden (CL < 10, 60%), and 42.4% carried at least one APOE ε4 allele. Forty-five percent of individuals were scanned with [18F]Florbetaben and 55% with [18F]Flutemetamol. Compared to gray zone or low amyloid burden groups, individuals with high amyloid burden were significantly older ($P = 6.7 \times 10^{-11}$, $P = 1.2 \times 10^{-20}$, respectively), had significantly higher global CDR scores ($P = 1.1 \times 10^{-6}$, $P = 2.4 \times 10^{-11}$, respectively), and had significantly lower Mini-Mental State Examination scores ($P = 5.6 \times 10^{-6}$, $P = 6.3 \times 10^{-4}$, respectively). There were significantly fewer APOE ε4 carriers in the low amyloid burden group (32.8%) compared to the gray zone (48.1%) and high (66.7%) amyloid burden groups ($P = 2.5 \times 10^{-4}$ and $P = 3.8 \times 10^{-14}$, respectively), and significantly more APOE ε4 carriers in the high amyloid burden group compared to the gray zone group ($P = 7.2 \times 10^{-4}$). More individuals were scanned with [18F]Flutemetamol in each of the groups, which was significantly higher in the gray zone (69.3%) compared to low (50.6%) amyloid burden individuals ($P = 1.4 \times 10^{-5}$), and in high (50.6%) compared to gray zone individuals ($P = 0.005$). For cohort characteristics on the full AMYPAD PNHS cohort see Bader et al.³ and see Table 1 for cohort characteristics on the subset used in the present study.

3.3 | Use of harmonized genetic data: PRS and their characteristics

Figure 2 displays the distribution plots for PRS at the genome-wide significance threshold for SNP inclusion (P value threshold, $pT = 5 \times 10^{-8}$). These plots illustrate that the PRS distributions across individual parent cohorts are largely overlapping and generally follow a normal distribution.

The number of SNPs included in each PRS across the SNP inclusion thresholds is shown in Figure 3. In general, PRS_{Kunkle} has the largest SNP set sizes at each pT , except at 0.1, where PRS_{amyloid} has the largest

SNP set size ($N = 53,050$), followed by PRS_{tau} ($N = 52,739$). Excluding the APOE region from the PRS results in a reduction in the number of SNPs, ranging from 93.75% (PRS_{amyloid} and PRS_{amyloid-noAPOE}) to 0.1% (PRS_{tau} and PRS_{tau-noAPOE}).

In Figure 4, the highest variability was observed for PRS_{amyloid-noAPOE} and PRS_{amyloid} at $pT = 5 \times 10^{-8}$ (interquartile range [IQR] = 1.83 and 1.68, respectively), whereas the lowest variability was for PRS_{Kunkle-noAPOE} and PRS_{Kunkle} at $pT = 0.1$ (IQR = 1.05 and 1.05, respectively). Figure 4 shows that individuals carrying more APOE ε4 alleles exhibit a higher genetic predisposition, indicated by lower scores for PRS_{amyloid}, and higher scores for PRS_{tau} and PRS_{Kunkle}. Despite differences in PRS set sizes and score variability, the strongest significant negative correlations were identified between PRS_{amyloid} and PRS_{Kunkle} at $pT = 5 \times 10^{-8}$ ($\rho = -0.84$) and between PRS_{amyloid} at $pT = 1 \times 10^{-5}$ and PRS_{Kunkle} at $pT = 5 \times 10^{-8}$ ($\rho = -0.83$). The strongest positive correlations were observed between PRS_{amyloid} and PRS_{amyloid-noAPOE} at $pT = 0.1$ ($\rho = 0.99$), and between PRS_{Kunkle} at $pT = 5 \times 10^{-8}$ and PRS_{Kunkle} at $pT = 1 \times 10^{-5}$ ($\rho = 0.97$). All correlations are illustrated in Figure 5 and detailed in Table S2 in supporting information.

3.4 | Application of PRS: associations with a global amyloid burden phenotype

All PRS, including the APOE region, were significantly different between low, gray zone, and high amyloid burden groups ($P < 0.04$, Figure S2 in supporting information). When excluding the APOE region, only PRS_{tau-noAPOE} at $pT = 0.1$ and PRS_{Kunkle-noAPOE} at $pT = 5 \times 10^{-8}$ ($P = 0.02$ and 2.2×10^{-3} , respectively) were significantly different between amyloid burden groups (Figure S3 in supporting information).

When further stratifying for APOE ε4 carriership (carrier vs. non-carrier), similar differences were observed for PRS, including the APOE region, where APOE ε4 carriers presented with higher genetic predisposition for amyloid burden than non-carriers, that is, APOE ε4 carriers had higher scores for PRS_{tau} and PRS_{Kunkle}, and lower scores for PRS_{amyloid} (Figure S4 in supporting information). However, fewer significant comparisons were observed between APOE ε4 and amyloid burden groups with the more flexible SNP inclusion threshold of $pT = 0.1$.

In contrast, significance was only observed for PRS_{amyloid-noAPOE} at $pT = 1 \times 10^{-5}$ ($P = 0.02$) between APOE ε4 non-carriers in the gray zone group and APOE ε4 non-carriers in the high amyloid burden group, and for PRS_{Kunkle-noAPOE} at $pT = 5 \times 10^{-8}$ ($P = 0.03$) between APOE ε4 non-carriers in the gray zone amyloid burden group and APOE ε4 carriers in the high amyloid burden group (Figure S5 in supporting information).

For the primary regression models, all PRS, including the APOE region, were significantly associated with amyloid burden ($P < 1 \times 10^{-3}$, Figure 6) except PRS_{Kunkle} at $pT = 0.1$, with the most significant predictor being PRS_{amyloid} at $pT = 5 \times 10^{-8}$ ($\beta_{\text{standardized}} = -0.29$, $P = 7.2 \times 10^{-18}$, adjusted $R^2 = 0.16$). PRS_{amyloid-noAPOE} ($\beta_{\text{standardized}} = 0.10$, $P = 0.03$, adjusted $R^2 = 0.08$) and PRS_{tau-noAPOE} ($\beta_{\text{standardized}} = 0.12$, $P = 6.0 \times 10^{-3}$, adjusted $R^2 = 0.09$) at

TABLE 1 Cohort demographics stratified by amyloid PET status.

	Low (CL < 12) (N = 516)	Gray-zone (10 < CL < 30) (N = 189)	High (CL > 30) (N = 162)	Overall (N = 867)	Statistics
Sex (female, N, %)	319 (61.8%)	105 (55.6%)	89 (54.9%)	513 (59.2%)	$\chi^2 = 3.7$, $P = 0.2$
Age (years, median, range)	64 (49-89)	66 (50-93)	72 (54-88)	66 (49-93)	$\chi^2 = 88.2$, $P = 7.2 \times 10^{-20}$
APOE $\epsilon 4$ carriers (N, %)	169 (32.8%)	91 (48.1%)	108 (66.7%)	368 (42.4%)	$\chi^2 = 61.3$, $P = 5.0 \times 10^{-14}$
Years of education (years, median, range)	15 (5-28)	15 (6-32)	15 (6-25)	15 (5-32)	$\chi^2 = 1.2$, $P = 0.6$
Global CDR 0 (N, %)	460 (89.1%)	167 (88.4%)	111 (68.5%)	738 (85.1%)	$\chi^2 = 46.0$, $P = 1.0 \times 10^{-10}$
MMSE (median, range)	29 (24-30)	30 (20-30)	29 (16-30)	29 (16-30)	$\chi^2 = 21.8$, $P = 1.8 \times 10^{-5}$
PET tracer [18F]Flutemetamol (N, %)	261 (50.6%)	131 (69.3%)	88 (54.3%)	480 (55.4%)	$\chi^2 = 19.7$, $P = 5.2 \times 10^{-5}$
Cohort (N, %)					
ALFA+	113 (21.9%)	46 (24.3%)	18 (11.1%)	177 (20.4%)	–
F-PACK	34 (6.6%)	7 (3.7%)	4 (2.5%)	45 (5.2%)	–
FACEHBI	147 (28.5%)	20 (10.6%)	23 (14.2%)	190 (21.9%)	–
EPAD LCS	178 (34.5%)	81 (42.9%)	102 (63%)	361 (41.6%)	–
EMIF-AD (60++)	44 (8.5%)	35 (18.5%)	15 (9.3%)	94 (10.8%)	–

Abbreviations: ALFA+, Alzheimer's and Families study; APOE, apolipoprotein E; CDR, Clinical Dementia Rating; CL, Centiloid; EMIF-AD 60++, European Medical Information Framework for Alzheimer's Disease 60++; EPAD LCS, European Prevention of Alzheimer's Disease Longitudinal Cohort Study; FACE-HBI, Fundació ACE Healthy Brain Initiative; F-PACK, Flemish Prevent AD Cohort KU Leuven; MMSE, Mini-Mental State Examination; PET, positron emission tomography.

pT = 0.1 were also significantly associated with global amyloid burden (Table S3 in supporting information). Note that when relaxing the threshold for SNP inclusion, the variance explained decreased for all PRS, including the APOE region, but increased for all PRS_{noAPOE} except for PRS_{Kunkle-noAPOE} (Figures S6 and S7 in supporting information). Furthermore, most significance was lost when analyzing these relationships in the individual parent cohorts that comprise the AMYPAD PNHS (Figure 6).

When APOE $\epsilon 4$ status was included as a covariate in the models for PRS_{noAPOE}, significance remained only for PRS_{tau-noAPOE} at pT = 0.1 ($\beta_{\text{standardized}} = 0.11$, $P = 6.5 \times 10^{-3}$, adjusted $R^2 = 0.17$, Figure S8, Table S4 in supporting information).

When APOE $\epsilon 4$ status was included as an interaction term with PRS_{noAPOE}, PRS remained a significant predictor for global amyloid burden only for PRS_{tau-noAPOE} at pT = 0.1 ($\beta_{\text{standardized}} = 0.16$, $P = 0.02$, adjusted $R^2 = 0.17$, Figure S9, Table S5 in supporting information). The interaction term was not a significant predictor for any of the models (Figure S10, Table S6 in supporting information).

3.5 | Application of PRS: PRS risk stratification in a clinical context

Amyloid burden was significantly different between low-, medium-, and high-risk PRS tertiles for all PRS, including the APOE region ($P < 0.01$),

except for PRS_{Kunkle} at pT = 0.1 (Figure S11 in supporting information). Amyloid burden was not significantly different between risk tertiles for PRS_{noAPOE}, except for PRS_{amyloid} at pT = 5×10^{-8} ($P = 7.6 \times 10^{-5}$, Figure S12 in supporting information). For those significant associations observed, individuals in the high PRS risk group had significantly higher amyloid burden compared to the medium PRS risk and low PRS risk groups.

When global amyloid burden was dichotomized into low and high (CL > 30) burden, eight models were significant for the comparison between high versus low PRS risk, two models for the comparison between high versus medium PRS risk, and three for medium versus low PRS risk (Figure 7, Table S7 in supporting information). The highest ORs for having high amyloid burden were observed for high versus low PRS risk for PRS_{amyloid} at pT = 1×10^{-5} (OR = 6.2 [3.8–10.5], $P = 3.5 \times 10^{-11}$) and pT = 5×10^{-8} (OR = 5.9 [3.6–10.0], $P = 1.0 \times 10^{-10}$).

4 | DISCUSSION

We harmonized genotype data from diverse arrays across multiple deeply phenotyped cohorts within the pan-European AMYPAD PNHS consortium spanning the AD risk continuum, resulting in a unified multimodal dataset suitable for large-scale analyses. Using this harmonized dataset, we computed several PRS and demonstrated their

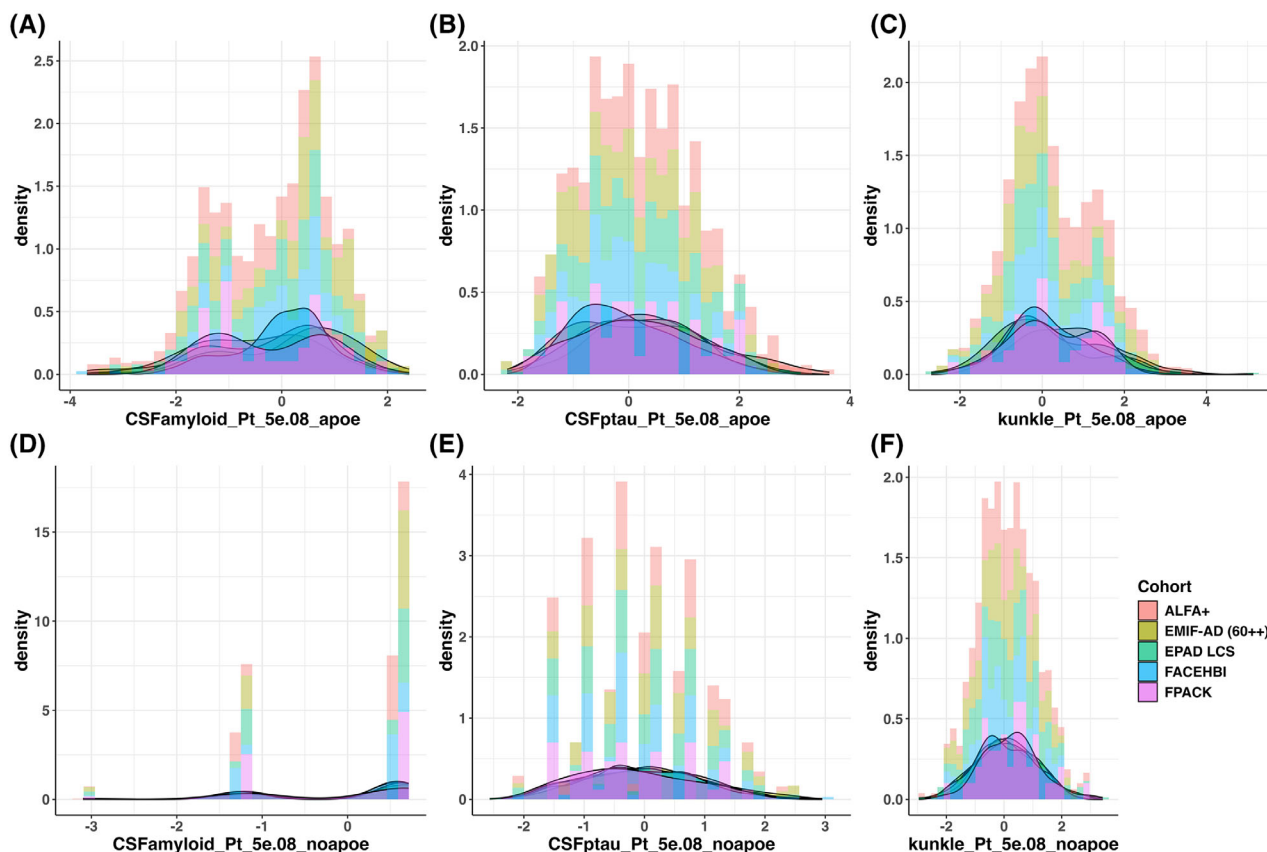


FIGURE 2 Representative PRS distributions for each AMYPAD PNHS Parent Cohort. All distributions show PRS at the genome-wide significance threshold for SNP inclusion ($pT = 5 \times 10^{-8}$). The top row shows PRS including the APOE region for (A) PRS_{amyloid}, (B) PRS_{tau}, and (C) PRS_{kunkle}. The bottom row shows PRS excluding the APOE region for (D) PRS_{amyloid-noAPOE}, (E) PRS_{tau-noAPOE}, and (F) PRS_{kunkle-noAPOE}. Note that a lower PRS_{amyloid(noAPOE)} is indicative of higher genetic predisposition to lower levels of CSF A β 42. A β , amyloid beta; ALFA+, Alzheimer's and Families study; AMYPAD, Amyloid Imaging to Prevent Alzheimer's Disease consortium; APOE, apolipoprotein E; CSF, cerebrospinal fluid; EMIF-AD 60++, European Medical Information Framework for Alzheimer's Disease 60++; EPAD LCS, European Prevention of Alzheimer's Disease Longitudinal Cohort Study; FACEHBI, Fundació ACE Healthy Brain Initiative; FPACK, Flemish Prevent AD Cohort KU Leuven; PNHS, Prognostic and Natural History Study; PRS, polygenic risk score; p-tau, phosphorylated tau; SNP, single nucleotide polymorphism.

association with global amyloid PET burden. Overall, PRS computed using CSF A β 42 GWAS summary statistics at $pT = 5 \times 10^{-8}$ and $pT = 1 \times 10^{-5}$ showed the strongest associations with global amyloid burden.

Access to large datasets with multiple data modalities available is key to understanding AD and its biological underpinnings. This is not a new phenomenon, given initiatives such as the Alzheimer's Disease Sequencing Project Phenotype Harmonization Consortium (ADSP-PHC). Although such large-scale initiatives aim to harmonize imaging and genetic data from thousands of individuals across the disease continuum, our study complements these efforts by focusing on a deeply phenotyped, prospectively recruited European cohort in the earliest disease stages. The AMYPAD PNHS is uniquely characterized by harmonized amyloid PET acquisition and centralized image processing, coupled with closely aligned biomarker and genetic data, making this a strong dataset for carrying out the present and future genetic association analyses. To investigate early pathological amyloid burden at the genomic level, harmonization efforts were required for integrating the heterogeneous genotype array data from multiple parent cohorts.

Our pipeline produced a harmonized genetic dataset, from data collected using different methodologies, suitable for investigating such genetic-endophenotype associations. First, principal component analysis confirmed that, despite the diverse origins of the data, individuals within the AMYPAD PNHS have a genetic profile similar to the broader European population. Second, the PRS derived from this harmonized dataset are largely overlapping, critical for the validity of cross-cohort genetic analyses and supporting the generalizability of AMYPAD PNHS findings. Last, we standardized the AMYPAD PNHS PRS against those from the European 1000G Project individuals. This approach removes intra-cohort standardization bias and provides a consistent framework for standardization, which is an important consideration for study replicability.

The association between PRSs and AD susceptibility has been well documented over the last decade, for example,²⁴⁻²⁹ where studies use AD susceptibility GWAS summary statistics to generate their scores. Similarly, studies that investigate the association of PRS with amyloid burden also use AD susceptibility GWAS summary statistics, which often capture a low variance explained despite the PRS being

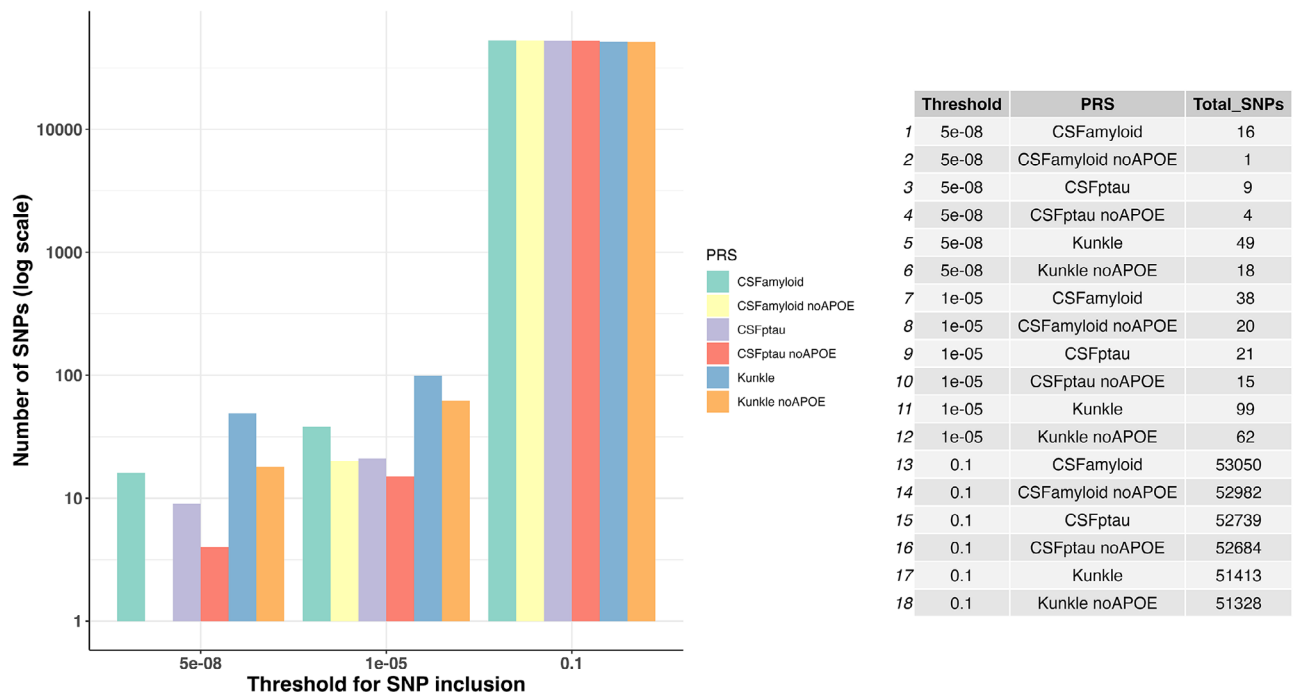


FIGURE 3 PRS SNP set size. A log scale is used on the y axis to enable all set sizes to be visualized given the wide range. The table on the right shows the actual SNP set size for each PRS. APOE, apolipoprotein E; CSF, cerebrospinal fluid; PRS, polygenic risk score; SNP, single nucleotide polymorphism.

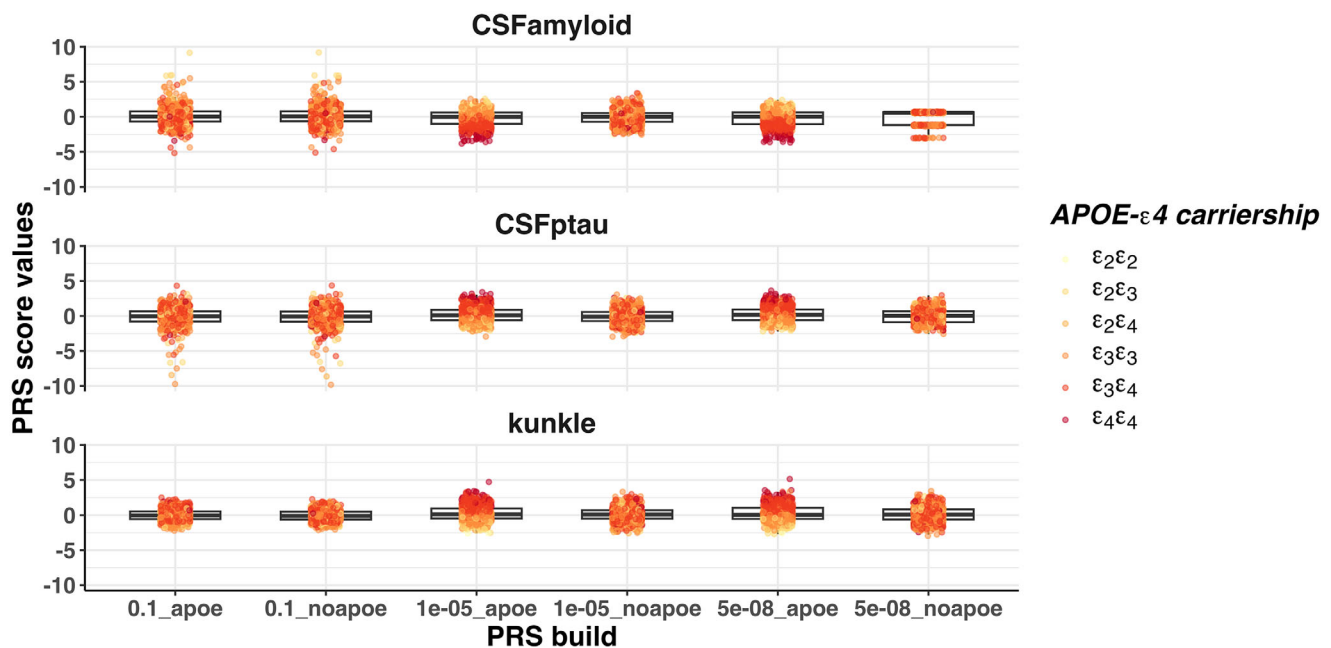


FIGURE 4 PRS distributions across phenotypes and P value thresholds. Each row represents PRS computed using a different set of summary statistics at the three thresholds for SNP inclusion. Scores with and without the APOE region are shown. Data points are colored based on the number of APOE $\epsilon 4$ alleles a participant carries, where darker colors represent the presence of more risk alleles. Note that a lower PRS_{amyloid(noAPOE)} is indicative of higher genetic predisposition to lower levels of CSF A β 42. A β , amyloid beta; APOE, apolipoprotein E; CSF, cerebrospinal fluid; PRS, polygenic risk score; SNP, single nucleotide polymorphism.

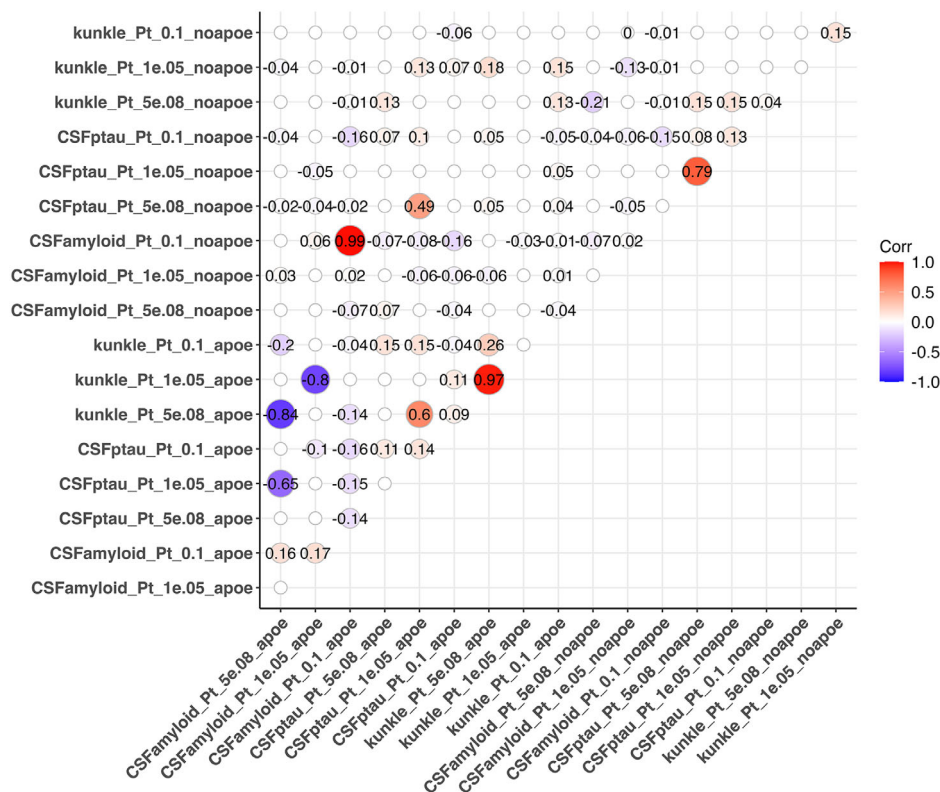


FIGURE 5 Correlation matrix illustrating the correlation coefficient for each pair of PRS. The color of the circles is based on the size of the correlation coefficient, and the size of the circles on the *P* value significance. Non-significant correlations are shown as blank circles. Note that a lower PRS_{amyloid(noAPOE)} is indicative of higher genetic predisposition to lower levels of CSF Aβ42. Aβ, amyloid beta; CSF, cerebrospinal fluid; PRS, polygenic risk score.

a significant predictor of amyloid deposition.^{20,30–33} We have shown the strength of combining data from individual cohorts in finding associations between PRS and amyloid burden in the predementia phase of AD, using both AD susceptibility and CSF A β 42 GWAS summary statistics. Indeed, the majority of the significant associations between PRS and amyloid burden emerged when analyzing the entire AMYPAD PNHS cohort as opposed to the individual parent cohorts, providing new insights into which PRS and corresponding summary statistics better predict amyloid burden.

The number of SNPs included in the PRS increased as the SNP inclusion threshold became more flexible. The overall smaller PRS set sizes for PRS_{amyloid(noAPOE)} and PRS_{tau(noAPOE)} compared to PRS_{Kunkle(noAPOE)} may reflect differences in genetic architecture or heritability of CSF A β 42 and CSF p-tau181 compared to AD diagnosis. However, it may also be due to the smaller sample sizes of the GWAS used to generate the summary statistics for CSF A β 42 and CSF p-tau181 ($N = 8074^{46}$) versus the Kunkle GWAS ($N = 21,982^{55}$). Larger sample sizes in future GWAS may enable the discovery of additional risk or protective variants currently undetected due to limited statistical power. Furthermore, the summary statistics from the Kunkle GWAS were derived from a traditional case-control design rather than for a specific outcome, such as levels of CSF A β 42 or CSF p-tau181. Consequently, the levels of CSF AD biomarkers are likely influenced by specific biological processes involving a limited number of contributing

SNPs, resulting in PRSs with smaller set sizes. These differences arising from using distinct summary statistics and PRS set size may also explain the PRS variability observed in Figure 4. Nonetheless, the high correlations observed between different builds of PRS, for example, PRS_{Kunkle} and PRS_{amyloid} at $pT = 5 \times 10^{-8}$ ($\rho = -0.84$), suggest that the PRS may be constructed using overlapping loci. Indeed, the two genome-wide significant loci for CSF A β 42 (Jansen) were also genome-wide significant for AD susceptibility (Kunkle).

Most PRS that included the *APOE* region were significantly associated with amyloid burden, with higher scores observed in *APOE* $\epsilon 4$ carriers, highlighting the well-established relationship between *APOE* $\epsilon 4$ and amyloid deposition in AD. However, $\text{PRS}_{\text{amyloid-noAPOE}}$ and $\text{PRS}_{\text{tau-noAPOE}}$ at $pT = 0.1$ were also significantly associated, providing evidence for non-*APOE* $\epsilon 4$ pathways or genetic variants that likely contribute to this pathological process during the earliest stages of the disease continuum. Further support for this was provided by the regression results when *APOE* $\epsilon 4$ status was included as a covariate; significance persisted for $\text{PRS}_{\text{tau-noAPOE}}$ at $pT = 0.1$. Notably, this PRS remained a significant predictor in the models that included an interaction term for $\text{PRS}_{\text{noAPOE}} \times \text{APOE } \epsilon 4$ status, indicating that the PRS independently contributes to amyloid burden. Our findings are consistent with previous studies, reaffirming the strong effect of *APOE* $\epsilon 4$ and the more modest aggregate contributions of other variants. However, we add value by validating these associations in a harmonized

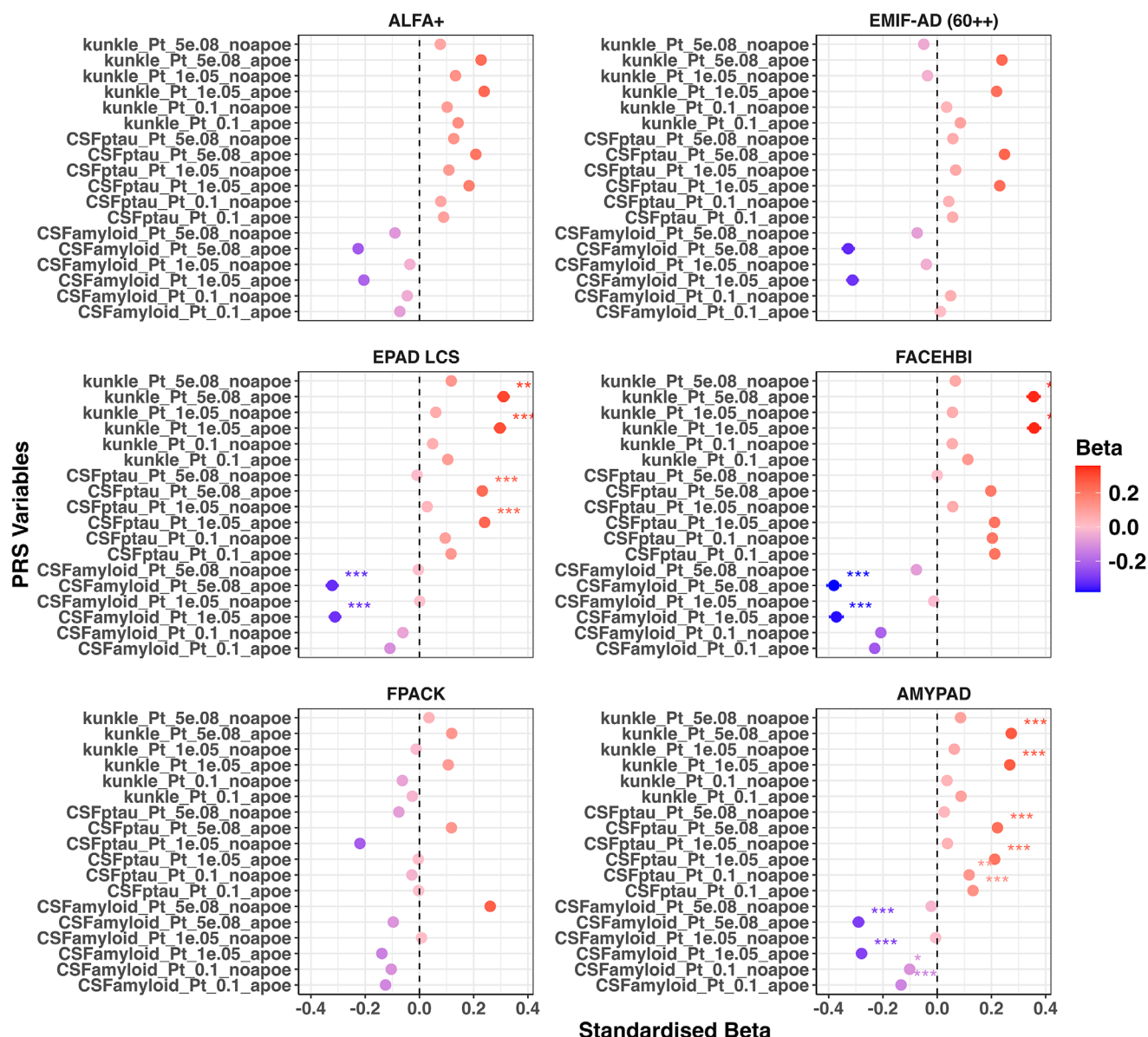


FIGURE 6 Association between PRS and global amyloid burden. The forest plots illustrate the standardized betas and confidence intervals from the primary linear regression models, with corresponding Bonferroni-corrected P values. Each panel is an individual parent cohort, with the bottom right panel being the harmonized AMYPAD PNHS cohort. * = $P < 0.05$; ** = $P < 0.005$; *** = $P < 0.001$. Note that a lower $PRS_{\text{amyloid}(noAPOE)}$ is indicative of higher genetic predisposition to lower levels of CSF $A\beta_{42}$. $A\beta$, amyloid beta; ALFA+, Alzheimer's and Families study; AMYPAD, Amyloid Imaging to Prevent Alzheimer's Disease consortium; CSF, cerebrospinal fluid; EMIF-AD 60++, European Medical Information Framework for Alzheimer's Disease 60++; EPAD LCS, European Prevention of Alzheimer's Disease Longitudinal Cohort Study; FACEHBI, Fundació ACE Healthy Brain Initiative; FPACK, Flemish Prevent AD Cohort KU Leuven; PNHS, Prognostic and Natural History Study; PRS, polygenic risk score.

and independent imaging cohort with robust preclinical phenotyping. In contrast to larger efforts, the AMYPAD PNHS provides a distinct setting for early-stage analysis, offering harmonized imaging and biomarker data that enable future modelling of brain amyloid burden in combination with other features (e.g., MRI, cognition, and further genetic analyses). This is particularly relevant given the more modest sample size, as the deeply phenotyped AMYPAD PNHS cohort allows for detailed subgroup analysis, longitudinal follow-up, and multimodal integration. Importantly, by focusing on cross-cohort harmonization and early amyloid PET burden, our study bridges a gap between genetic

risk prediction and early AD pathology—an area that is not always the central aim of larger GWAS-based PRS analyses.

For PRS including the $APOE$ region, the variance explained was higher than that observed in the models with PRS_{noAPOE} . However, the variance explained decreased when relaxing the SNP inclusion threshold to 0.1 for all PRS including the $APOE$ region, to similar values as observed for PRS_{noAPOE} . This highlights that less relevant SNPs are included as the threshold for SNP inclusion is relaxed for PRS_{APOE} , suggesting that amyloid burden is better predicted by a smaller set of high-confidence SNPs that includes the $APOE$ region. Nonetheless,

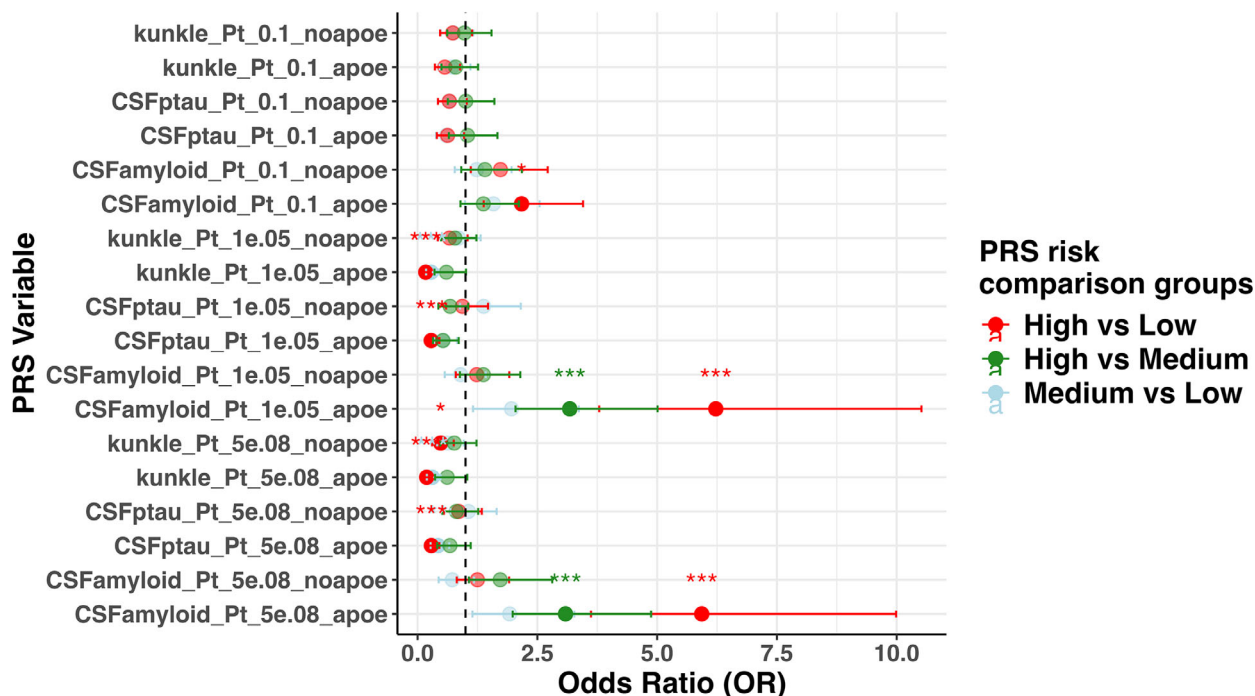


FIGURE 7 Effect of high, medium, and low PRS risk on high amyloid burden. The forest plot illustrates the odds ratios and confidence intervals from the logistic regression models, with corresponding Bonferroni-corrected *P* values. The dashed line at OR = 1 indicates no effect. High amyloid burden was defined as CL > 30. * = *P* < 0.05; ** = *P* < 0.005; *** = *P* < 0.001. CL, Centiloid; CSF, cerebrospinal fluid; OR, odds ratio; PRS, polygenic risk score; ptau, phosphorylated tau.

non-APOE $\epsilon 4$ influences should be further explored, given the results observed with PRS_{tau-noAPOE} as discussed above.

Among the PRS analyzed, PRS_{amyloid} showed the highest ORs, especially for high versus low and medium versus low PRS risk groups, and the strongest associations with brain amyloid burden. This suggests that PRS derived from CSF A β 42 GWAS summary statistics contain more relevant genetic variants contributing to amyloid deposition in the early stages of AD compared to those from a traditional case-control design. This validates the use of computing PRS using summary statistics beyond those of an AD case-control design when evaluating the association of PRS with amyloid burden. Furthermore, this complements published literature in which PRS are computed using different summary statistics^{25,34} or methods, for example, pathway-specific scores,^{35–38} providing further information regarding the genetic architecture of AD and its pathological processes. This highlights the potential for targeted genetic profiling to identify at-risk individuals, which is especially relevant given the ongoing clinical trials and regulatory approvals of amyloid-lowering therapies.^{37,38} The individuals most likely to benefit from these treatments are those predisposed to amyloid deposition or accumulation who are still in the earliest disease stages, prior to significant cognitive impairment. A targeted PRS capable of identifying such individuals presents as an ideal tool for use in a first-stage hierarchical approach, complementing established participant selection tools. Furthermore, the substantial overlap of PRS across cohorts and their significant predictive value for global amyloid PET burden highlight their comparability across a pan-European population, validating the potential use of these scores in clinical set-

tings. PRS could serve as a primary non-invasive tool to assess AD risk and inform treatment strategies without the immediate use of PET or CSF acquisition, as is currently performed in the initial stages of clinical evaluation.

We selected established GWAS summary statistics of AD susceptibility and CSF biomarkers for PRS computation, reflecting their common use in (preclinical) AD research and aligning with the primary focus of the AMYPAD PNHS on amyloid PET burden. However, future analyses could benefit from incorporating amyloid PET-specific GWAS summary statistics such as those from Ali et al.³⁹ This could refine the trait specificity of PRS estimates and enhance the interpretation of genetic influences on PET-derived amyloid burden. Nonetheless, our aim was to develop and demonstrate a reproducible genetic harmonization and PRS pipeline within this imaging-centric, preclinical cohort. However, looking ahead, incorporating summary statistics based on other AD-relevant endophenotypes, such as hippocampal volume or vascular pathology, may help strengthen the consistency and interpretability of findings. Although our study focused on European-ancestry data, we also recognize the importance of incorporating more diverse summary statistics—such as those from the recent multi-ethnic GWAS of amyloid imaging³⁹—to assess PRS performance across populations. This is a key area for future work aimed at improving generalizability and enhancing predictive power.

In conclusion, we successfully harmonized pan-European genotype array data for a predementia AD population, enabling the identification of specific associations between derived PRS and cortical amyloid PET burden. This work highlights the importance of robust

data harmonization procedures and pooling of cohort data in facilitating large-powered studies and ensuring their accessibility to the broader research community, and validates the potential use of PRS in clinical or clinical trial settings as a primary non-invasive tool to assess AD risk.

ACKNOWLEDGMENTS

The authors would like to express their thanks to the AMYPAD PNHS participants, without whom this research would have not been possible. For a full list of AMYPAD PNHS contributors see <https://doi.org/10.5281/zenodo.7962737>. This work used data from AMYPAD PNHS that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115952. This Joint Undertaking received support from the European Union's Horizon 2020 research and innovation programme and EFPIA. F.B. and E.S.L. have received support for data curation and storage from the Alzheimer's Disease Data Initiative (ADDI; paid to institution). L.E.C. has acquired research support from GE Healthcare and Springer Healthcare (paid by Eli Lilly), both paid to institution. L.E.C. is supported by the MSCA Postdoctoral fellowship (#101108819) and Alzheimer Association Research Fellowship (#23AARF-1029663) grants. N.V.T. receives funding from the Spanish Research Agency MICIU/AEI/10.13039/501100011033 (grant RYC2022-038136-I co-funded by the European Union FSE+, and grant PID2022-143106OA-I00 co-funded by the European Union FEDER). Additionally, N.V.T. is supported by the William H. Gates Sr. Fellowship from the Alzheimer's Disease Data Initiative, and grant 23S06083-001 funded by "la Caixa" Foundation and Barcelona City council. The F-PACK study was funded by the Stichting Alzheimer Onderzoek, Internal Funds KU Leuven, and Flemish Research Foundation (G0G1519N, G094418N), VLAIO (HBC.2019.2523, for R.V., Y.A., and I.C.). F.B. is supported by the NIHR biomedical research centre at UCLH.

CONFLICT OF INTEREST STATEMENT

This communication reflects the views of the authors and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. Research of Alzheimer Center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stichting Steun Alzheimer-centrum Amsterdam. EMIF-Twins-60+ is supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking EMIF grant agreement no. 115372, Alzheimer Nederland and Stichting Dioraphte. P.J.V. is a coinventor on a patent of CSF proteomic subtypes (published under patent no. US2022196683A1, owner VUmc Foundation). C.R. is the founder, CEO, and majority shareholder of Scottish Brain Sciences; has received compensation for study-related activities from AC Immune SA (paid to institution); has received consulting fees from Biogen, Eisai, MSD, Actinogen, Roche, Eli Lilly, and Novo Nordisk; has received payment or honoraria from Roche, Eisai, and Eli Lilly; and participates on a data safety monitoring board for Novo Nordisk. M.B. is an employee of the Ace Alzheimer Center and an advisory board member for Grifols, Roche, Eli Lilly, Araclon Biotech, Merck, Zambon, Biogen, Novo

Nordisk, Bioiberica, Eisai, Servier, and Schwabe Pharma. J.D.G. has received research support from GE HealthCare, Roche Diagnostics, Hoffmann La Roche, and Life-MI; has participated in symposia sponsored by Biogen, Philips Nederlands, Life-MI, and Esteve; acted as a consultant for Roche Diagnostics; and served on the molecular neuroimaging advisory board of Prothena Biosciences. J.D.G. is founder, co-owner, and member of the board of directors of Betascreen SL. J.D.G. is currently a full-time employee of AstraZeneca. R.V.'s institution has Clinical Research Agreements (RV as PI) with Alektor, AviadoBio, Biogen, BMS, J&J, and UCB. R.V.'s institution has consultancy agreements (R.V. as DSMB chair or member) with AC Immune and Novartis. F.B. is a steering committee or data safety monitoring board member for Biogen, Merck, Eisai, and Prothena; advisory board member for Combinostics, Scottish Brain Sciences, and Alzheimer Europe; and consultant for Roche, Celltrion, Rewind Therapeutics, Merck, and Bracco. F.B. has research agreements with ADDI, Merck, Biogen, GE Healthcare, and Roche, and is co-founder and shareholder of Queen Square Analytics LTD. E.S.L., Y.A., L.L., L.E.C., D.V.G., A.dB., M.B., P.G., N.V.T., and I.C. have no disclosures. Author disclosures are available in the [supporting information](#).

DATA AVAILABILITY STATEMENT

Post-publication, derived PRS will be integrated into the AMYPAD PNHS public release dataset on the Alzheimer's Disease Data Initiative WorkBench (<https://www.alzheimersdata.org/ad-workbench>) and will be made available to individuals after an approved data access request. The harmonization pipeline will be made available upon publication. Data used in the preparation of this article were obtained from the AMYPAD PNHS dataset located on the Alzheimer's Disease Data Initiative WorkBench: AMYPAD PNHS (Harmonized and Derived) v202306.

CONSENT STATEMENT

All AMYPAD participants provided written informed consent.

REFERENCES

- Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry*. 2006;63:168-174. doi:10.1001/archpsyc.63.2.168
- Lopes Alves I, Collij LE, Altomare D, et al. Quantitative amyloid PET in Alzheimer's disease: the AMYPAD prognostic and natural history study. *Alzheimers Dement*. 2020;16:750-758. doi:10.1002/alz.12069
- Bader I, Bader I, Lopes Alves I, et al. Recruitment of pre-dementia participants: main enrollment barriers in a longitudinal amyloid-PET study. *Alzheimers Res Ther*. 2023;15:1-14. doi:10.1186/s13195-023-01332-4
- Collij LE, Farrar G, Vallés García D, et al. The amyloid imaging for the prevention of Alzheimer's disease consortium: a European collaboration with global impact. *Front Neurol*. 2023;13:1063598. doi:10.3389/fneur.2022.1063598
- Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics*. 2019;51:414-430. doi:10.1038/s41588-019-0358-2
- Jansen IE, van der Lee SJ, Gomez-Fonseca D, et al. Genome-wide meta-analysis for Alzheimer's disease cerebrospinal fluid biomarkers. *Acta Neuropathol*. 2022;144:821-842. doi:10.1007/s00401-022-02454-z

7. Shekari M, Verwer EE, Yaqub M, et al. Harmonization of brain PET images in multi-center PET studies using Hoffman phantom scan. *EJNMMI Phys*. 2023;10:68. doi:10.1186/s40658-023-00588-x
8. Klunk WE, Koeppe RA, Price JC, et al. The Centiloid project: standardizing quantitative amyloid plaque estimation by PET. *Alzheimers Dement*. 2015;11:1-15.e4. doi:10.1016/j.jalz.2014.07.003
9. Collij LE, Bollaack A, La Joie R, et al. Centiloid recommendations for clinical context-of-use from the AMYPAD consortium. *Alzheimers Dement*. 2024;20:9037. doi:10.1002/ALZ.14336
10. de Rojas I, Moreno-Grau S, Tesi N, et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nat Commun*. 2021;12:1-16. doi:10.1038/s41467-021-22491-8
11. Vilor-Tejedor N, Genius P, Rodríguez-Fernández B, et al. Genetic characterization of the ALFA study: uncovering genetic profiles in the Alzheimer's continuum. *Alzheimers Dement*. 2023;18:19. doi:10.1002/alz.13537
12. Adamczuk K, De Weer AS, Nelissen N, et al. Functional changes in the language network in response to increased amyloid β deposition in cognitively intact older adults. *Cerebral Cortex*. 2016;26:358-373. doi:10.1093/cercor/bhu286
13. Adamczuk K, De Weer AS, Nelissen N, et al. Polymorphism of brain derived neurotrophic factor influences β amyloid load in cognitively intact apolipoprotein $\epsilon 4$ carriers. *Neuroimage Clin*. 2013;2:512-520. doi:10.1016/j.nicl.2013.04.001
14. Bos I, Vos S, Vandenberghe R, et al. The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. *Alzheimers Res Ther*. 2018;10:64. doi:10.1186/s13195-018-0396-5
15. Access | EPAD n.d. (accessed February 7, 2024) <https://ep-ad.org/open-access-data/access/>
16. Hong S, Prokopenko D, Dobricic V, et al. Genome-wide association study of Alzheimer's disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset. *Transl Psychiatry*. 2020;10:1-12. doi:10.1038/s41398-020-01074-z
17. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559-575. doi:10.1086/519795
18. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74. doi:10.1038/NATURE15393
19. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008
20. Luckett ES, Abakkouy Y, Reinartz M, et al. Association of Alzheimer's disease polygenic risk scores with amyloid accumulation in cognitively intact older adults. *Alzheimers Res Ther*. 2022;14:1-13. doi:10.1186/s13195-022-01079-4
21. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27:e1608. doi:10.1002/mpr.1608
22. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284-1287. doi:10.1038/ng.3656
23. Choi SW, O'Reilly PF. PRSice-2: polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019;8:1-6. doi:10.1093/gigascience/giz082
24. Escott-Price V, Sims R, Bannister C, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain*. 2015;138:3673-3684. doi:10.1093/brain/awv268
25. Bellou E, Kim W, Leonenko G, et al. Benchmarking Alzheimer's disease prediction: personalised risk assessment using polygenic risk scores across various methodologies and genome-wide studies. *Alzheimers Res Ther*. 2025;17. doi:10.1186/s13195-024-01664-9
26. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun*. 2021;12:1-10. doi:10.1038/s41467-021-24082-z
27. Bellou E, Baker E, Leonenko G, et al. Age-dependent effect of APOE and polygenic component on Alzheimer's disease. *Neurobiol Aging*. 2020;93:69-77. doi:10.1016/j.neurobiolaging.2020.04.024
28. Escott-Price V, Shoaib M, Pither R, Williams J, Hardy J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol Aging*. 2017;49:2141e7. doi:10.1016/j.neurobiolaging.2016.07.018
29. Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann Neurol*. 2017;82:311-314. doi:10.1002/ana.24999
30. Ramanan VK, Gebre RK, Graff-Radford J, et al. Genetic risk scores enhance the diagnostic value of plasma biomarkers of brain amyloidosis. *Brain*. 2023;146:4508-4519. doi:10.1093/brain/awad196
31. Ramanan VK, Heckman MG, Przybelski SA, et al. Polygenic Scores of Alzheimer's Disease Risk Genes Add Only Modestly to APOE in Explaining Variation in Amyloid PET Burden. *J Alzheimer's Disease*. 2022;88:1615-1625. doi:10.3233/JAD-220164
32. Leonenko G, Shoaib M, Bellou E, et al. Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition. *Ann Neurol*. 2019;86:427-435. doi:10.1002/ana.25530
33. Xicota L, Gyorgy B, Grenier-Boley B, et al. Association of APOE-Independent Alzheimer Disease Polygenic Risk Score With Brain Amyloid Deposition in Asymptomatic Older Adults. *Neurology*. 2022;99:E462-75. doi:10.1212/WNL.0000000000200544
34. Chaudhury S, Brookes KJ, Patel T, et al. Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Transl Psychiatry*. 2019;9, 154. doi:10.1038/s41398-019-0485-7
35. Tesi N, Van Der Lee SJ, Hulsman M, et al. Pathway-specific polygenic risk score of AD-associated genetic variants associated with AD risk, resilience against AD, and progression to AD. *Alzheimers Dement*. 2021;17:e053500. doi:10.1002/alz.053500
36. Harrison JR, Foley SF, Baker E, et al. Pathway-specific polygenic scores for Alzheimer's disease are associated with changes in brain structure in younger and older adults. *Brain Commun*. 2023;5. fcd229. doi:10.1093/braincomms/fcd229
37. Tesi N, van der Lee SJ, Hulsman M, et al. Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease. *Transl Psychiatry*. 2020;10, 332. doi:10.1038/s41398-020-01018-7
38. Darst BF, Kosciak RL, Racine AM, et al. Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid- β Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. *J Alzheimer's Disease*. 2017;55:473-484. doi:10.3233/JAD-160195
39. Ali M, Archer DB, Gorijala P, et al. Large multi-ethnic genetic analyses of amyloid imaging identify new genes for Alzheimer disease. *Acta Neuropathol Commun*. 2023;11:68. doi:10.1186/s40478-023-01563-4

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Luckett ES, Abakkouy Y, Lorenzini L, et al. Harmonizing genotype array data to understand genetic risk for brain amyloid burden in the AMYPAD PNHS Consortium. *Alzheimer's Dement*. 2025;21:e70376. <https://doi.org/10.1002/alz.70376>