Endophenotype-based polygenic risk scores: Prediction of biomarker and clinical progression and dementia

Danai Chasioti
  Indiana University-Purdue University Indianapolis

Tanner Jacobson
  Indiana University School of Medicine

Kwangsik Nho
  Indiana University School of Medicine

Shannon L. Risacher
  Indiana University School of Medicine

Sujuan Gao
  Indiana University School of Medicine

Jingwen Yan
  Indiana University-Purdue University Indianapolis

Andrew J. Saykin (✉ asaykin@iupui.edu)
  Indiana University School of Medicine

Research Article

Abstract

BACKGROUND: Biomarkers provide a framework for a biological diagnosis of Alzheimer's disease (AD) whereas polygenic risk scores (PRS) provide method to estimate genetic risk. We derive biomarker-based PRS by incorporating endophenotype genetic risk relevant to amyloid, tau, neurodegeneration and cerebrovascular (A/T/N/V) pathology.

METHODS: Endophenotype-PRSs (PRS_A, PRS_T, PRS_N, PRS_V) and combined-PRSs (PRS_{AT}, PRS_{ATNV}) were generated using the Alzheimer's Disease Neuroimaging Initiative (ADNI) data. Prediction performance of the PRSs was assessed in terms of dementia risk, age at onset (AAO) and longitudinal change of 14 important AD biomarkers.

RESULTS: PRS_A and PRS_T explained more amyloid and tau variability than combined PRSs (CSF-amyloid: R^2_{PRSA} = 9.22%; CSF-tau: R^2_{PRST} = 6.37%; CSF-ptau: R^2_{PRST} = 7.10%). Combined-PRSs explained more neurodegeneration-related variability ($R^2_{PRSATNV}$ range: 1.22%-4.20%) and were strong predictors of dementia risk (HR and OR p-value<8.3e-03) and AAO ($AAO_{(predicted\_vs\_observed)}$: r_{AT}=0.76).

CONCLUSIONS: PRS_A and PRS_T are AD-specific, while combined-PRSs are linked to neurodegeneration in general. Biomarker-derived PRSs provide mechanistic insights beyond aggregate disease susceptibility, supporting development of precision medicine for dementia.

1. Background

According to the multifactorial etiology (1) for complex diseases, the phenotypic variability can be explained by the additive effect of multiple genetic factors. A PRS is the mathematical formulation of this hypothesis, being a single combinatorial measure of multiple individual genetic effects that express an individual's overall genetic liability (2, 3). In Alzheimer's disease (AD), PRS studies have focused primarily on risk and prognosis (3–11), with the majority focusing on late onset Alzheimer's disease (LOAD) based on large case-control genome wide association studies (GWAS). However, the increased phenotypic and genetic heterogeneity among LOAD patients calls for more personalized solutions and thus, for approaches that integrate biologically relevant genetic information (5, 8, 10, 12–14). Here we employed AD endophenotype-specific GWAS to develop individual and combined endophenotype-PRSs. Our goal was to investigate the potential of endophenotype-PRSs for prediction of biomarker progression and prognosis in dementia.

2. Materials And Methods

Briefly, in this work we first studied whether the individual endophenotype and combined endophenotype-PRSs can capture the risk of dementia (expressed as both odds ratio and hazard ratio) as well as the age of dementia onset. Next, we tested the potential of endophenotype-PRS to capture the genetic risk beyond *APOE* by studying the differences in the dementia risk and median survival among ε3/ε3 participants. Finally, we computed linear mixed models to determine the relationship with longitudinal trajectories of known AD endophenotypes.

2.1 Study population

Data used in the preparation of this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) (15). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For the purpose of developing endophenotype-PRSs, we focused on 11 biomarkers from ADNI1,GO/2, each biomarker representing either amyloid, tau, neuronal or vascular pathology (Fig. 1). From the 1,550 individuals available, only 585 participants had complete baseline information on all 11 biomarkers of interest. Of these 585 participants, 80% were used for training, 20% for validation, and the

remaining 965 participants were used for testing (Table 1). Diagnosis was based on clinical criteria and consisted of five different categories: cognitively normal (CN), significant memory concerns (SMC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) and demented (Dem), with demented being characterized as participants whose diagnosis was based on clinical rather than pathological evidence (16).

Table 1
Data description

Characteristic	Full		Training		Validation		Testing	
Count	1550		468		117		965	
Age								
Mean	73.4		72.2		71.6		74.4	
Range	47−91		47−91		55−88		54−90	
Gender (%)								
Male	59.3		53.5		50.0		58.8	
Female	40.7		46.5		50.0		41.2	
Diagnosis (%)	Baseline Diagnosis	Final Diagnosis	Baseline Diagnosis	Final Diagnosis	Baseline Diagnosis	Final Diagnosis	Baseline Diagnosis	Final Diagnosis
CN	23.7	22.1	21.8	22.1	19.5	22.0	24.2	21.5
SMC	6.0	5.7	12.0	11.6	11.9	11.9	2.4	2.2
EMCI	18.0	17.5	29.8	26.6	32.3	26.3	10.4	11.9
LMCI	33.1	20.0	18.6	10.9	18.6	11.0	41.5	25.8
Dem	19.2	34.7	17.8	28.9	17.8	28.8	21.5	38.7

2.2 Biomarker PCA

We focused on integrating information from 11 biomarkers that fall under the A/T/N/V framework (16). This is an expansion of the A/T/N framework (17), which was developed to reflect the pathophysiology progress of the disease and thus, provide a better understanding of its clinical stages. The set of biomarkers that we selected for PRS development is presented in Fig. 1. These include CSF and PET amyloid (A), CSF tau (T), MRI and FDG-PET (N) from selected regions of interest (ROI), as well as white matter hyperintensity (V). To summarize the information from these biomarkers, we performed principal components analysis (PCA) simultaneously on the residuals of all 11 biomarkers that were first pre-adjusted for age, sex, years of education, and the first two genetic PCs that controlled for population stratification. PCA was applied on the 585 individuals with full baseline biomarker data (Fig. 1). For each participant, the baseline was defined as the first time point with available measurements for all 11 biomarkers. The analysis returned 4 components, each representing one biomarker group (A, T, N, and V).

2.3 Single Nucleotide Polymorphism (SNP) filtering

For each of the 4 obtained endophenotype components, we ran GWASs on the same 585 participants that had been used for the PCA step (Fig. 1). The genotype data were HRC imputed, with a total number of 5,406,481 SNPs. The GWASs results have been filtered based on a range of p-value cut-offs (5e-05, 8e-05, 1e-04, 8e-04, 1e-03, 5e-03, 1e-02). To address the linkage disequilibrium problem (LD), we performed clumping using PLINK on SNPs with MAF$\geq$5%, $r^2 = 0.1$ and window=1K kb. From the *APOE* region (defined as 1Mb up and downstream of the gene position, 44,409,039 to 46,412,650) only *rs429358* and *rs7412* were retained.

2.4 Further SNP filtering and SNP weight calculation

In addition to p-thresholding, we further filtered the SNPs by applying Lasso (18), a type of penalized regression. At each p-threshold and for each biomarker component, Lasso returned a list of SNPs and their corresponding weights. The Lasso penalty was determined by tuning the lambda parameter using 10-fold cross-validation. The criterion for optimal lambda selection was minimization of the mean square error (MSE). While shrinkage was applied to SNPs, the baseline age, sex, years of education, as well as the two *APOE* SNPs, *rs429358 and rs7412*, were not subject to penalization. To increase the stability of the result, Lasso was bootstrapped 100 times on the training set, returning at each iteration a list of SNPs and SNP weights. The final SNP list was obtained by retaining the most frequently selected SNPs (selection frequency $\geq 80\%$, Fig. 1).

According to the literature, re-weighted SNP coefficients may achieve improved PRS performance (5, 19) compared to the traditional case/control GWAS SNP effects. Because Lasso estimates tend to be biased (20), we followed a two-step procedure by refitting a linear regression model on the Lasso selected SNPs. The regression model was adjusted for the covariates, age, sex, and years of education and was performed separately for each of the four endophenotype components. The process was bootstrapped 100 times on the training set, and the final PRS SNP weight was calculated by averaging the corresponding regression coefficient over the 100 bootstrap iterations, as described in Eq. (1). Here, w_s is the new weight for SNP s, i is the bootstrap iteration index, N is the total number of bootstrap iterations (in this case $N = 100$), and $coef_{si}$ is the regression coefficient for the SNP s at the ith iteration.

$$w_s = \sum_{i=1}^{N} coef_{si} / N$$

1

2.5 Individual and Combined Biomarker-PRS

At each p-threshold and for every participant j, we calculated four individual endophenotype-PRSs (PRS$_A$, PRS$_T$, PRS$_N$, PRS$_V$) based on Eq. (2). The PRS was expressed as the sum over the weighted number of alleles per SNPs. Specifically, for the jth individual, the endophenotype b PRS (PRS_{bj}) was obtained by multiplying the minor allele count d_{sj} of the SNP s by the SNP weight w_s (described in Eq. 1).

$$PRS_{bj} = \sum_{s=1}^{S_b} w_s d_{sj}$$

2

Finally, we generated two combined endophenotype-PRSs (PRS$_{ATNV}$, PRS$_{AT}$) for each individual. The PRS$_{ATNV}$ was expressed as the weighted sum of the individual biomarker-PRSs as shown in (3). In Eq. (3), PRS_b is the individual endophenotype-PRS as described in Eq. (2). To obtain the weights w_b, we used the training set to regress each of the four endophenotype components on the corresponding PRS_b while controlling for the age, sex and years of education. The final weight w_b for each PRS_b was the average coefficient over 100 bootstrap iterations. A similar approach was followed for generating the PRS$_{AT}$.

$$PRS_{ATNV} = \sum_{b \in \{A,T,N,V\}} w_b PRS_b$$

3

2.6 Best PRS threshold selection

To select the optimal p-threshold and thus the final PRS for the remainder of the analysis, we assessed the prediction performance of the biomarker PRSs on the validation set for each of the seven p-thresholds. Specifically, we obtained the adjusted variance explained (Adj.R^2) by regressing each biomarker component on the corresponding biomarker-PRS while controlling for baseline age, sex and years of education. The average Adj.R^2 over 100 bootstrap iterations indicated the best overall p-threshold.

2.7 Odds of AD in relation to PRS

To study the association between the six PRSs and the risk of dementia, we ran a logistic regression model, treating the PRS as predictor while adjusting for the centered covariates of age, sex, and years of education. In the model described here, age was defined as either the age of clinical diagnosis of dementia or the age at the last clinical visit for the non-demented participants. To simplify the interpretation, the PRSs, originally ranging from 0 to 1 with values closer to 1 indicating higher risk, were multiplied by 10. Among the 585 participants of the training set, 367 individuals that were either CN, SMC or Dem were used for model training. Having estimated the odds of AD for each PRS, we replicated the results on 712 individuals from the testing set, after excluding MCI patients. As an additional step, to assess the predictive ability of SNPs beyond *APOE*, we obtained the risk of developing dementia among $\epsilon3/\epsilon3$ participants.

2.8 AD hazard and time to AD onset in relation to PRS

Other statistical measures of interest in AD research include the hazard of dementia and the age of dementia onset. To assess the strength of the relationship between these measures and the biomarker PRSs, we ran a Cox proportional hazard (PH) model, which was trained using 367 individuals from the training set. As "event" we considered the onset of dementia (clinical manifestation), and we treated the age at dementia diagnosis as the survival time in the model. PRS was used as a predictor in the model, while adjusting for the years of education and sex. To simplify interpretation, the PRSs were multiplied by 10 and education was centered. The PH assumption was tested using the cox.zph() function in R. To get predictions of the age to dementia onset among the Dem cases, we predicted the survival curves using the Cox model that was previously applied on the training data. The actual and the predicted age to dementia were divided into deciles. The relationship between the predicted age and actual age of dementia onset was assessed using Pearson correlation (*r*). The analysis was replicated on 712 non-MCI individuals from the test set as well as on the $\epsilon3/\epsilon3$ participants.

2.9 PRS for baseline levels and longitudinal trajectories of responses of interest

In addition to dementia risk prediction, which may be useful at the prevention stage, information about disease progression and key outcomes are also important. Here, we assessed the baseline and longitudinal effects of PRSs on 14 responses of interest. These included three cognitive measures (ADNI-MEM, ADNI-EF and FAQ), as well as 11 biomarkers that were described earlier (Fig. 1). For each of the 14 responses, we applied a random intercept linear mixed model to account for the correlation between repeated measurements. The data were aligned for all participants, with time 0 representing the first visit when a measurement was available for each biomarker. All biomarkers were transformed to range between 0 to 1, while MRI biomarkers were pre-adjusted for intracranial volume (ICV). Whenever necessary, the biomarkers were log_{10} transformed. The model adjusted for sex and centered covariates, including years of education and baseline age. Fixed effects included years since baseline, as well as the PRS and their interaction. To simplify interpretation, the PRSs were multiplied by 10. The random intercept term allowed for varying intercepts among the participants. The performance was assessed by the Nakagawa's marginal pseudo-R^2 on the testing set. The significance of the increase in the pseudo-R^2 was assessed by ANOVA, which compared the (full) model, PRS and its interaction with time, to the (base) model, which contained covariates only. The p-values of the main PRS effect (baseline effect) and the interaction effect (longitudinal change), were corrected for multiple comparisons. Specifically, for each endophenotype, a Bonferroni correction was applied to account for testing against six PRSs (Bonferroni p-value = 8.3e-03).

3. Results

3.1 PRS Calculation

The best PRS performance was achieved for the GWAS p-value threshold of 8e-04, based on the average Adj.R^2 over 100 bootstrap iterations. Because that was the best threshold for all biomarkers, except for vascular, we considered it to be the overall optimal p-threshold. At this optimal threshold, the number of PRS SNPs (including the *APOE* SNPs rs429358 and rs7412) was 145 for PRS_A, 166 for PRS_T, 160 for PRS_N and 159 for PRS_V. The PRSs calculated at the specific threshold were used in steps for the remaining analysis.

3.2 Odds of AD in relation to PRS

On the training set, the strongest association between clinically diagnosed dementia and PRS was observed for PRS_N followed by PRS_{ATNV}. For the former, a 0.1 unit increase in the PRS_N increased the odds of dementia by 4.5 times (OR = 4.5, p = 1.28e-20), whereas for the latter, the OR of Dem was 3.38 (p = 9.96e-24). The results were validated on the testing set including all *APOE* groups (PRS_N: OR = 1.29, p = 4.8e-04; PRS_{ATNV}: OR = 1.52, p = 1.03e-07). To demonstrate the information provided by the SNPs beyond *APOE*, we examined the strength of the association among $\epsilon3/\epsilon3$ carriers, observing a significant OR of 4.7 (p = 1.45e-08) for PRS_N and 2.76 for PRS_{ATNV} (p = 7.58e-10). On the $\epsilon3/\epsilon3$ testing group, neither PRS_N nor PRS_{ATNV} effects were significant (PRS_N: OR = 1.13, p > 0.1; PRS_{ATNV}: OR = 1.19, p > 0.1).

3.3 AD hazard and time to AD onset in relation to PRS

On the training set, the strongest association between dementia onset and PRS was observed for PRS_{AT} followed by PRS_{ATNV}. For the former, the rate of being clinically diagnosed with dementia at any time point was increased by 67% for each 0.1 unit increase of the PRS_{AT} (p = 3.52e-26), whereas for the latter, the hazard ratio (HR)of AD was 1.62 (p = 2.04e-25). Both associations were replicated on the test set (PRS_{AT}: HR = 1.24, p = 5.97e-07; PRS_{ATNV}: HR = 1.20, p = 2.08e-05). Among $\epsilon3/\epsilon3$ carriers of the training set, both PRS effects were significant (PRS_{AT}: HR = 1.53, p = 4.74e-06; PRS_{ATNV}: HR = 1.58, p = 3.49e-08). We additionally obtained a 10-year difference in the median AAO between the extreme PRS_{AT} quartiles ($PRS_{AT,Q1} \leq 0.29$, $PRS_{AT,Q4} \geq 0.60$) of the $\epsilon3/\epsilon3$ in the training set (Fig. 2.A; AAO: 76 for Q4 and 86 for Q1). For PRS_{ATNV} the median AAO was 86 in Q1 and 74 in Q4 ($PRS_{ATNV, Q1} \leq 0.33$, $PRS_{ATNV, Q4} \geq 0.63$) (Fig. 2.B). For all the results reported here, the proportional hazard (PH) assumption was met. Finally, we evaluated the scores' performance in predicting the AAO by observing its association with the actual AAO. The association was observed among deciles of the predicted and observed AAO. For the training set, the Pearson correlations were r_{AT} = 0.83 (p = 2.7e-03) and r_{ATNV} = 0.78 (p = 7.4e-03). For the testing set, the correlations were r_{AT} = 0.76 (p = 1.2e-02) and r_{ATNV} = 0.64 (p = 4.6e-02).

3.4 PRS and longitudinal trajectories of cognitional and biomarker responses

We compared the percentage variance explained for the mixed model with and without PRS and assessed the change using ANOVA to test between the two models. The results on the test set are presented in Table 2. For CSF- amyloid, tau, and p-tau, the individual endophenotype-PRSs for amyloid, and tau (PRS_A, PRS_T) resulted in the highest improvement of the explained variance for the corresponding biomarkers. This improvement was statistically significant after correcting for multiple comparisons (Table 2).

Endophenotype	Base model	PRS$_A$	PRS$_T$	PRS$_N$	PRS$_V$	PRS$_{AT}$	PRS$_{ATNV}$
ATNV-related endophenotypes							
Roche CSF Abeta	0.93%	**9.22%**	1.98%	1.12%	–	8.82%	7.96%
AV45	0.70%	11.25%	4.08%	2.48%	0.58%	12.38%	**12.76%**
Roche CSF Tau	2.06%	1.55%	**6.37%**	–	0.09%	5.47%	4.50%
Roche CSF pTau	1.23%	2.40%	**7.10%**	–	0.07%	7.00%	5.41%
Mean Lat. Temp. (Thx Avg)	11.19%	0.26%	0.68%	0.74%	0.09%	0.69%	**1.22%**
Mean Med. Temp	12.30%	0.77%	1.29%	1.18%	0.11%	1.57%	**1.78%**
Hippocampus Vol.	13.72%	1.86%	1.48%	0.52%	0.20%	**2.78%**	2.17%
FDG Temp. Lobe	3.10%	1.65%	2.82%	1.11%	–	3.51%	**4.20%**
FDG Ang. Gyrus	3.62%	1.41%	2.21%	0.82%	–	2.88%	**3.58%**
FDG Cingulate	6.00%	1.41%	1.74%	1.19%	0.54%	2.57%	**3.90%**
WMHI	5.84%	0.30%	–	0.13%	**0.40%**	0.28%	**0.40%**
Not ATNV-related endophenotypes							
ADNI MEM	8.14%	1.71%	–	1.08%	–	1.79%	**2.29%**
ADNI EF	8.62%	–	**0.65%**	–	–	0.40%	0.51%
FAQ	4.52%	1.12%	1.35%	0.95%	–	2.02%	**2.28%**

* Longitudinal analysis results. The table presents the overall marginal R^2 increase compared to the base model

Base mixed model $Y = \beta_0 + \beta_1 t + covar + (1\,|\,ID)$

Full mixed model $Y = \beta_0 + \beta_1 t + \beta_2 PRS + \beta_3 (PRS * t) + covar + (1\,|\,ID)$

Overall increase represents the increase due to inclusion of PRS and its interaction with time all together

Dashes indicate that the overall R^2 increase compared to the base model (covariates only) was not significant (ANOVA p-value > 0.05)

By examining the significance of the PRS terms in these models, we noticed significant main effects for the PRS, but the interaction terms were insignificant for all three biomarkers (Table 3). Significant interaction for both tau biomarkers was achieved by PRS$_{ATNV}$ although this was not the optimal score in terms of variance explained (Table 2, CSF-tau: $R^2_{PRS_{ATNV}}$ = 4.50%; CSF-ptau: $R^2_{PRS_{ATNV}}$ = 5.41%). When studied individually, PRS$_A$ and PRS$_V$ showed a negative relation to both tau measures (CSF-tau: Time×PRS$_A$: p = 6.5e-03; Time×PRS$_V$: p = 5.5e-03; CSF-ptau: Time×PRS$_V$: p = 2.3e-03; Time×PRS$_{ATNV}$: p = 5.9e-02). On the other hand, PET-amyloid, MRI and FDG-PET biomarkers, as well as memory and FAQ had stronger associations to combined-PRSs (Table 2).

Table 3
Main and interaction PRS effects. Results on ADNI1,GO/2 test *

ATNV-related endophenotypes	Best PRS	Main effect (β_2)	Interaction effect (β_3)
Roche CSF Abeta	A	-0.04**	–
AV45	ATNV	0.05**	–
Roche CSF Tau	T	0.04**	–
Roche CSF pTau	T	0.05**	–
Mean Lat. Temp. (Thx Avg)	ATNV	-0.01**	-1.8e-03**
Mean Med. Temp	ATNV	-0.02**	-3.3e-03**
Hippocampus Vol.	AT	-0.01**	-1.2e-03**
FDG Temp. Lobe	ATNV	-0.03**	-4.1e-03**
FDG Ang. Gyrus	ATNV	-0.02**	-2.4e-03**
FDG Cingulate	ATNV	-0.02**	-2.7e-03**
WMHI	V	0.01**	–
Not ATNV-related endophenotypes			
ADNI MEM	ATNV	-0.02**	–
ADNI EF	T	-0.01*	–
FAQ	ATNV	0.05**	4.1e-03**

* Longitudinal analysis results. The table presents the main and interaction with time PRS effects (APOE included in the PRS)

Mixed model $Y = \beta_0 + \beta_1 t + \beta_2 PRS + \beta_3 (PRS * t) + covar + (1 \mid ID)$

Dashes indicate insignificant result (p-value > 0.05)

**p < 8.3e-03 (Bonferroni correction was done by biomarker (p-value: 0.05/6= 8.3e-03)

* p-value < 0.05

This is with the exception of PET-amyloid and MEM, where levels were strongly linked to the combined PRS levels both at the baseline and longitudinally (Table 3), even after correcting for multiple comparisons (Table 3). Overall, most of the associations studied here remained significant within the ε3/ε3 training set (Supplementary Table 1), but none of the neurodegeneration markers reached significance on the ε3/ε3 testing set (Supplementary Table 2). FAQ was the only cognition measure that remained significant, even among the ε3/ε3 individuals of the test set.

4. Discussion

We developed individual and combined endophenotype-PRSs and evaluated their association with dementia risk, age at dementia onset, and biomarker trajectories. Combined endophenotype-PRSs led to significantly higher dementia hazard and specifically accelerated the median AAO among $\epsilon3/\epsilon3$ participants up to 12 years. Finally, PRS_A and PRS_T were AD-specific, as they were better predictors of amyloid and tau biomarkers, while combined endophenotype-PRSs were better predictors of neurodegeneration.

We showed that PRSs based on specific AD biomarkers can be used to assess dementia risk and prediction of biomarker trajectory (21). In addition, we found that the progression of pathophysiological biomarkers and cognitive decline have a stronger association with the combined-PRSs, compared to the more biologically restricted individual endophenotype-PRSs. A possible explanation is that the combined-PRSs accounts for the effect of SNPs related to multiple endophenotypes and thus, better captures the multiple biological mechanisms implicated in these biomarkers. The insignificant interaction effects of time with PRS_A and PRS_T, when modeling the CSF amyloid and tau trajectories respectively (Table 3), may indicate that the observed increase in explained variance was likely driven by the strong association of these scores with the corresponding baseline biomarker levels. The significant $PRS_{ATNV} \times Time$ interaction for CSF-tau and CSF-ptau on the other hand, possibly emanates from the amyloid and vascular terms integrated in $PRS_{ATNV,}$ which seem to have a negative association to both tau trajectories. This could support the idea that different PRSs may be preferred, depending on whether we are interested in predicting cross-sectional differences or differences in the rate of change. Lastly, we provided significant evidence for genetic risk beyond *APOE* by replicating the previously observed differences in the age of dementia onset among $\epsilon3/\epsilon3$ participants (5).

In this study, we found that PRS accounts for biologically relevant information that may elucidate the level of genetic complexity of AD endophenotypes and related outcomes. Superiority in performance of combined-PRSs compared to individual-PRSs may indicate greater complexity in the underlying biological mechanisms of the response of interest, which may be indicative of the additional genetic information incorporated in the combined-PRS. While the development of endophenotype-PRS was aimed at improving PRS interpretation, other scores with the same goal have been developed. Pathway-PRS is one such score, which attempts to increase interpretability by inclusion of SNPs that are part of a specific biological pathway (10, 12, 21). Despite the seemingly similar rationale between the proposed individual endophenotype-PRS and pathway-PRS (10, 12, 21), there are also major differences. For example, pathway-PRS are developed based on existing case/control GWA studies that may fail to identify SNPs related to important biomarkers (22), particularly if the disease endophenotypes are closer to the molecular mechanism than the disease status, in which case endophenotype-GWASs may have greater utility in identifying biomarker-related SNPs (22). Pathway-PRS also requires *apriori* knowledge about the disease pathways and the SNPs belonging to that pathway, which may be problematic, as literature has shown that the number of informative SNPs in the PRS can significantly affect the results of the analysis (2). In contrast, the endophenotype-PRS identifies biomarker-related SNPs through endophenotype-GWAS, allowing multiple biological pathways associated with that biomarker to enter the score at the same time.

In this work, we provided a comprehensive comparison between individual and combined endophenotype-PRSs based on ADNI data. Despite the interesting findings, our work also has limitations. First, there was a limited sample size for PRS development, as ADNI is the only publicly available study that offers such an extensive collection of AD-related biomarkers. Limited sample size is also a barrier because the data had to be further split into training and testing conditions. However, as more participants are recruited, the power of the analysis will improve. The limited discovery sample may also explain our failure to observe a significant AAO difference among the $\epsilon3/\epsilon3$ participants in the testing set. Second, although part of the ADNI1,GO2 was kept apart and used solely for replication, it is still a part of the same cohort that was used for PRS development. Replication of these results in completely independent data sets is necessary and should be pursued when data availability permit. Third, ADNI is not ideal for building vascular-PRSs because individuals with more severe cerebrovascular disease are typically excluded. A better vascular-PRS should be derived using a more appropriate data set enriched for cerebrovascular disease.

To conclude, our study suggests that PRS_A and PRS_T are AD-specific as they have the best performance in predicting amyloid and tau biomarkers, whereas the combined PRSs are more generic and preferred for predicting neurodegeneration. Also, further analysis of the endophenotype-PRSs could offer functional insights and promote treatment development. Specifically, in the context of precision medicine, endophenotype-PRSs could be used for generating more specific genetic risk profiles for prospective trial enrollees that are specifically aligned with measurable biomarkers thought to reflect disease status. In the future, individual endophenotype scores could be utilized to study the genetic heterogeneity among individuals at risk for

dementia, which could potentially provide useful information about the observed variability in the disease's clinical manifestation and further support the necessity for individualized treatment.

Abbreviations

AAO: age at onset

AD: Alzheimer's Disease

ADNI: Alzheimer's Disease Neuroimaging Initiative

CN: Cognitively Normal

Dem: Demented

EMCI: Early Mild Cognitive Impairment

GWAS: Genome Wide Association Study

HR: Hazard Ratio

ICV: Intracranial Volume

LD: Linkage Disequilibrium

LMCI: Late Mild Cognitive Impairment

LOAD: Late Onset Alzheimer's Disease

MCI: Mild Cognitive Impairment

MRI: Magnetic Resonance Imaging

MSE: Mean Square Error

OR: Odds ratio

PCR: Principal Components Analysis

PET: Positron Emission Tomography

PH: Proportional Hazard

PRS: Polygenic Risk Scores

ROI: Regions of Interest

SMC: Significant Memory Concerns

Declarations

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

As per ADNI protocols, all procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later

amendments or comparable ethical standards. More details can be found at adni.loni.usc.edu. (This article does not contain any studies with human participants performed by any of the authors).

CONSENT FOR PUBLICATION

Consent for publication has been granted by ADNI administrators.

AVAILABILITY OF DATA AND MATERIAL

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. The principal investigator of ADNI is Michael W. Weiner, MD (email: Michael.Weiner@ucsf.edu).

COMPETING INTERESTS

FUNDING

AUTHORS' CONTRIBUTIONS

Dr. Andrew J. Saykin, Dr. Danai Chasioti and Dr. Jingwen Yan conceived of the presented idea. Dr. Danai Chasioti and Dr. Jingwen Yan developed the methodology. Dr. Danai Chasioti performed the computations. Dr. Tanner Jacobson performed the genome wide association analysis. Dr. Shannon L. Risacher performed the quality control and the preparation of the clinical data. Dr. Kwangsik Nho performed the quality control and the preparation of the genetic data. Dr. Jingwen Yan and Dr.

References

1. Fisher R. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh. 1919;2(52):399–433.

2. Chasioti D, Yan J, Nho K, Saykin AJ. Progress in Polygenic Composite Scores in Alzheimer's and Other Complex Diseases. Trends Genet. 2019;35(5):371–82.

3. Shen L, Thompson PM. Brain Imaging Genomics: Integrated Analysis and Machine Learning. Proc IEEE Inst Electr Electron Eng. 2020;108(1):125–62.

4. Daunt P, Ballard CG, Creese B, Davidson G, Hardy J, Oshota O, et al. Polygenic Risk Scoring is an Effective Approach to Predict Those Individuals Most Likely to Decline Cognitively Due to Alzheimer's Disease. J Prev Alzheimers Dis. 2021;8(1):78–83.

5. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med. 2017;14(3):e1002258.

6. Banks SJ, Qiu Y, Fan CC, Dale AM, Zou J, Askew B, et al. Enriching the design of Alzheimer's disease clinical trials: Application of the polygenic hazard score and composite outcome measures. Alzheimers Dement (N Y). 2020;6(1):e12071.

7. Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain. 2015;138(Pt 12):3673–84.

8. Escott-Price V, Myers A, Huentelman M, Shoai M, Hardy J. Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. J Prev Alzheimers Dis. 2019;6(1):16–9.

9. Chaudhury S, Brookes KJ, Patel T, Fallows A, Guetta-Baranes T, Turton JC, et al. Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. Transl Psychiatry. 2019;9(1):154.

10. Darst BF, Koscik RL, Racine AM, Oh JM, Krause RA, Carlsson CM, et al. Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-beta Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. J Alzheimers Dis. 2017;55(2):473–84.

11. genoSCORE-LAB [Available from: https://www.cytoxgroup.com/alzheimers-overview.

12. Ahmad S, Bannister C, van der Lee SJ, Vojinovic D, Adams HHH, Ramirez A, et al. Disentangling the biological pathways involved in early features of Alzheimer's disease in the Rotterdam Study. Alzheimers Dement. 2018;14(7):848–57.

13. Desikan RS, Schork AJ, Wang Y, Thompson WK, Dehghan A, Ridker PM, et al. Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease. Circulation. 2015;131(23):2061–9.

14. Sierksma A, Escott-Price V, De Strooper B. Translating genetic risk of Alzheimer's disease into mechanistic insight and drug targets. Science. 2020;370(6512):61–6.

15. Alzheimer's Disease Neuroimaging Initiative [Available from: http://adni.loni.usc.edu.

16. Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018;14(4):535–62.

17. Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Feldman HH, Frisoni GB, et al. A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. Neurology. 2016;87(5):539–47.

18. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society. 1996;58(1):267–88.

19. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41(6):469–80.

20. Zhao Y, Dantony E, Roy P. Optimism Bias Correction in Omics Studies with Big Data: Assessment of Penalized Methods on Simulated Data. OMICS. 2019;23(4):207–13.

21. Harrison JR, Mistry S, Muskett N, Escott-Price V. From Polygenic Scores to Precision Medicine in Alzheimer's Disease: A Systematic Review. J Alzheimers Dis. 2020;74(4):1271–83.

22. Barber RC, Phillips NR, Tilson JL, Huebinger RM, Shewale SJ, Koenig JL, et al. Can Genetic Analysis of Putative Blood Alzheimer's Disease Biomarkers Lead to Identification of Susceptibility Loci? PLoS One. 2015;10(12):e0142360.
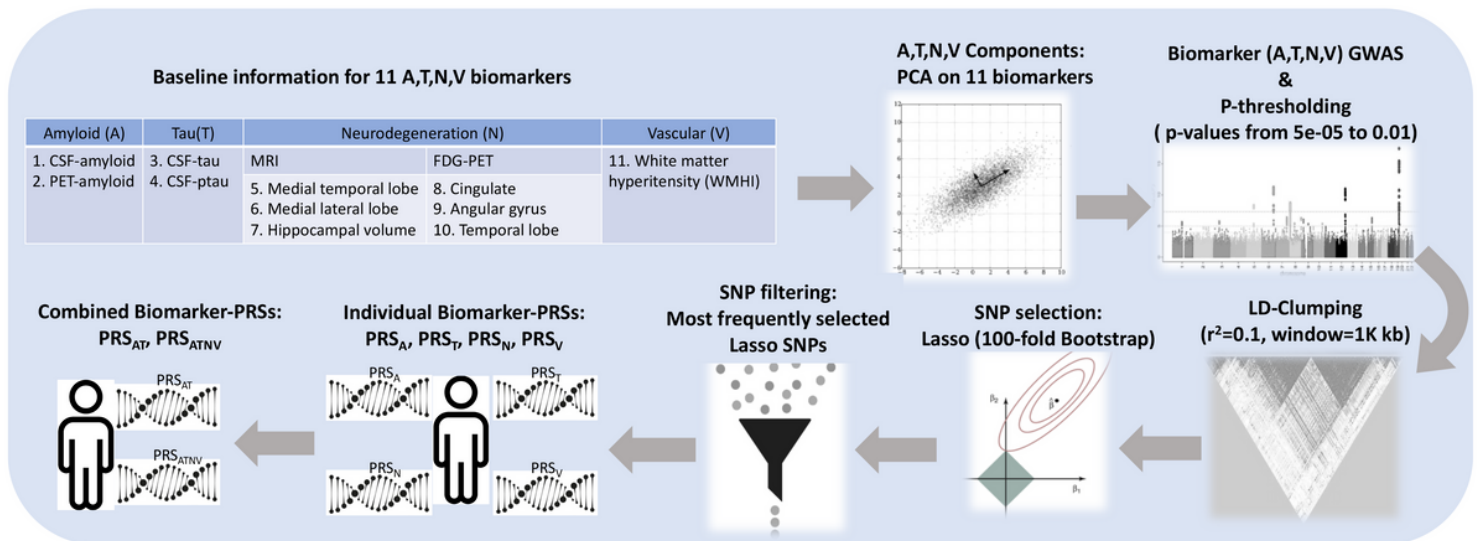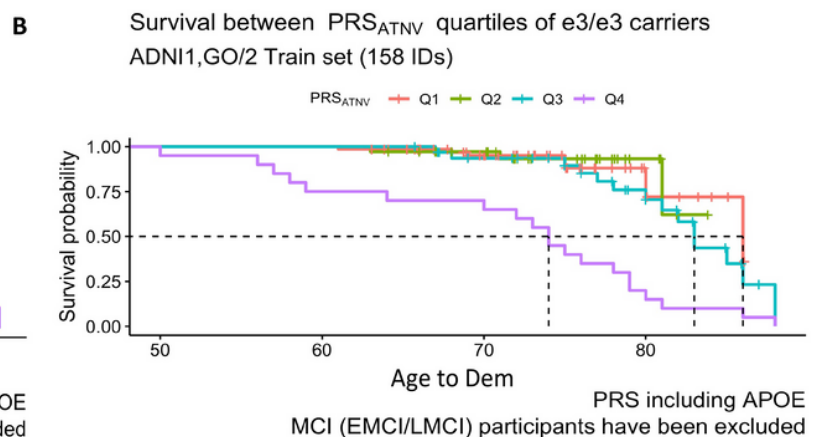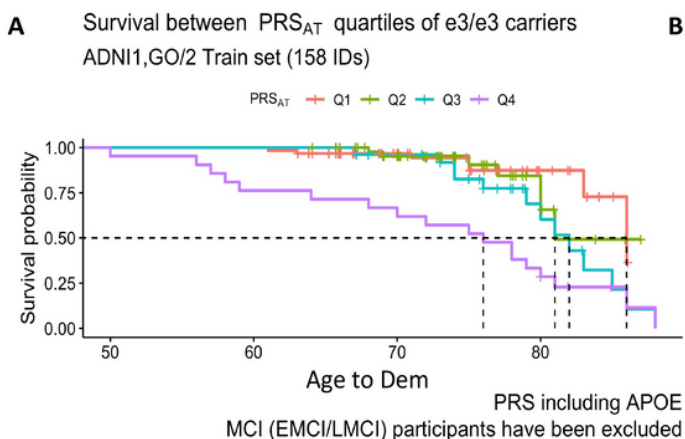
Figures

Figure 1

PRS calculation steps

Figure 2

Survival curves among ε3/ε3 individuals in the training set. Dashed lines represent the median age to dementia. **(A)** Results by PRS_{AT} quartiles. **(B)** Results by PRS_{ATNV} quartiles.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.pdf
- AdditionalFile2.pdf