Genetic stratification to identify risk groups for Alzheimer's disease

Citation for published version:

Digital Object Identifier (DOI):
10.3233/JAD-161070

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published in:
Journal of Alzheimer's Disease

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Genetic Stratification to Identify Risk Groups for Alzheimer’s Disease

Riccardo E. Marioni\textsuperscript{a,b,c,*}, Archie Campbell\textsuperscript{b}, Saskia P. Hagenaars\textsuperscript{a,d,e}, Reka Nagy\textsuperscript{f}, Carmen Amador\textsuperscript{f}, Caroline Hayward\textsuperscript{f}, David J. Porteous\textsuperscript{a,b}, Peter M. Visscher\textsuperscript{a,c} and Ian J. Deary\textsuperscript{a,d}

\textsuperscript{a}Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK
\textsuperscript{b}Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
\textsuperscript{c}Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia
\textsuperscript{d}Department of Psychology, University of Edinburgh, Edinburgh, UK
\textsuperscript{e}Division of Psychiatry, University of Edinburgh, Edinburgh, UK
\textsuperscript{f}Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

Handling Associate Editor: M. Arfan Ikram

Accepted 12 January 2017

Abstract. Stratification by genetic risk factors for Alzheimer’s disease (AD) may help identify groups with the greatest disease risk. Biological changes that cause late-onset AD are likely to occur years, if not decades prior to diagnosis. Here, we select a subset of the Generation Scotland: Scottish Family Health Study cohort in a likely preclinical age-range of 60–70 years (subset \(n\) = 3,495 with cognitive and genetic data). We test for cognitive differences by polygenic risk scores for AD. The polygenic scores are constructed using all available SNPs, excluding those within a 500 kb distance of the APOE locus. Additive and multiplicative effects of APOE status on these associations are investigated. Small memory decrements were observed in those with high polygenic risk scores for AD (standardized beta –0.04, \(p\) = 0.020). These associations were independent of APOE status. There was no difference in AD polygenic scores across APOE haplotypes (\(p\) = 0.72). Individuals with high compared to low polygenic risk scores for AD (top and bottom 5% of the distribution) show cognitive decrements, albeit much smaller than for APOE/\(e4\) compared to \(e3\) individuals (2.3 versus 3.5 fewer points on the processing speed test, and 1.8 versus 2.8 fewer points on the memory test). Polygenic risk scores for AD may help identify older individuals at greatest risk of cognitive decline and preclinical AD.

Keywords: Alzheimer’s disease, apolipoprotein E, cognitive function, genetics, polygenic traits

INTRODUCTION

It is widely acknowledged that the neuropathological hallmarks of Alzheimer’s disease (AD) present many years prior to diagnosis [1]. Cognitive decrements are expected to be observed closer to clinical diagnosis [1]. Targeting individuals who are likely to be in the earliest stages of the disease is therefore a key focus for clinical trials and interventions [2–4].

Age is the biggest risk factor for AD although there are also genetic components to the disease. The apolipoprotein gene, APOE, which is involved in lipid transportation, confers the greatest known genetic risk of AD [5, 6]. APOE/\(e4\) homozygotes...
have a 14.9 increased odds of developing dementia compared to those with the ε3ε3 reference haplotype [7]. The ε4 allele has a frequency in the general population of around 15% [8], implying that just over 2% of the population are ε4ε4 homozygotes. Despite the well-replicated association between APOE and AD, relatively little is known about its functional role in the disease process [5], although many biological processes including neuroinflammation, neurotoxicity, and lipid metabolism among others have been highlighted [6].

In addition to APOE, several other genes have been implicated in the pathogenesis of AD [9]. As with many other diseases, AD is a polygenic trait whereby many genetic polymorphisms of small effect are likely to contribute to the disease process [9]. One method that incorporates many of these variants into a single measure is polygenic risk scoring [10]. This method uses existing results from genome-wide association studies (GWAS) to provide weights specific to each genetic polymorphism, which can then be applied to independent cohorts. Thus, each individual in an independent cohort can be assigned a genetic risk score that is based on potentially thousands of genetic variants that individually explain some fraction of the risk of AD. For example, polygenic scores for AD predict around 2% of the variance of AD in an independent cohort [11]. AD polygenic risk scores were also shown to discriminate best between cases and controls between the ages of 60 and 70 years [11].

Given the low frequency of the ε4ε4 haplotype, large sample population-based cohorts are required to study its effects with precision. A previous study utilizing one such cohort, Generation Scotland (n = 18,337), investigated cognitive ability by APOE status [12]. It found evidence for poorer memory and processing speed in ε4ε4 homozygotes (compared to ε3ε3 homozygotes) in a sub-sample of participants aged over 60 years. These age-stratified findings coincide with the theoretical predictions of Sperling et al. [1]. Furthermore, given the prediction models of AD development, it is plausible that cognitive decrements predictive of AD will be most notable in populations between the ages of 60 and 70, i.e., the decade prior to an exponential increase in AD diagnosis.

The primary aim of this study is to test if there are cognitive decrements in those with a high polygenic risk of AD and to see how these effects compare with APOE ε4ε4 status. The analysis will focus on a subgroup from the Generation Scotland cohort in the age range of 60 to 70 years.

**MATERIALS AND METHODS**

**Generation Scotland: Scottish Family Health Study**

Data came from Generation Scotland: Scottish Family Health Study (hereafter referred to as Generation Scotland), a large population-based cohort sampled from five regional centers across Scotland [13, 14]. Initial recruitment focused on 7,953 individuals aged between 35 and 65 years, who were registered with a participating General Practice surgery; around 96% of the UK population is registered with a general medical practitioner. Relatives of these probands were then recruited. There were up to three generations of ~7,000 participating families in the study, recruited between 2006 and 2011, yielding a cohort of over 24,000 subjects.

There was no intended recruitment enrichment for any disease or health condition. Details on cognitive, anthropometric, and health measures were recorded. A full description of the cohort and the data collected have been reported elsewhere [13, 14] and at http://www.generationscotland.org.

**Cognitive data**

As previously described, four domains of cognitive function were assessed by single tests in nearly all Generation Scotland participants (n = 21,524): processing speed (Wechsler Digit Symbol Substitution Test [15]), verbal declarative memory (Wechsler Logical Memory Test; sum of immediate and delayed recall of one paragraph [16]), verbal fluency (the phonemic Verbal Fluency Test; using the letters C, F, and L, each for one minute [17]), and vocabulary (the Mill Hill Vocabulary Scale; junior and senior synonyms combined [18]). As a previous Generation Scotland study showed evidence for age-related cognitive decrements in processing speed and verbal declarative memory but not verbal fluency or vocabulary [12], we focused here on the former two outcomes only.

**Genetic data**

Genome wide genotyping and APOE haplotyping details have been described previously [12]. Briefly, Generation Scotland participants were genotyped with either the HumanOmníExpressExome8v1-2_A or HumanOmníExpressExome-8v1_A. Quality control was carried out in PLINK version 1.9b2c [19, 20].
SNPs were removed if they had a missingness rate >2% or a Hardy-Weinberg Equilibrium test \( p < 10^{-6} \). Duplicate samples were removed. Individuals were removed based on gender mismatch and missingness (>2% of genotypes missing). The subsequent data were combined with the 1,092 individuals of the 1000 Genomes population [21] prior to principal components being calculated in GCTA [22]. Outliers, defined by being more than six standard deviations away from the mean of the first two principal components, were removed [23]. This left a sample of 20,032 participants.

**APOE** haplotype status depends on the genotypes of two single nucleotide polymorphisms (SNPs), rs429358 and rs7412 that can form three possible haplotypes: e2, e3, and e4 [24]). Array genotyping of these SNPs is technically difficult and, as a result, they are not available on the majority of commercial arrays. SNP genotypes were thus obtained using Taqman technology at the Wellcome Trust Clinical Research Facility Genetics Core, Edinburgh. Blood samples from Generation Scotland participants were collected, processed, and stored using standard operating procedures and managed through a laboratory information management system at the Wellcome Trust Clinical Research Facility Genetics Core, Edinburgh [25]. APOE genotyping data were available on 21,039 individuals.

**Analysis cohort**

After merging the APOE, GWAS, and cognitive data, and after excluding individuals with self-reported AD (or a missing value) and restricting the cohort to individuals aged between 60 and 70 years, inclusive, the analysis population contained 3,495 participants. A flowchart documenting the selection process is provided in Fig. 1.

**Polygenic risk scores**

Polygenic risk scores for AD were calculated using the PRSice software program with LD clumping parameters set to \( R^2 > 0.25 \) over 250 kb sliding windows [26]. The discovery GWAS from which the SNP weights were extracted was the Stage I AD GWAS analysis by Lambert et al. [27]. The Generation Scotland polygenic scores were generated using all possible SNPs \( (p < 1) \) from the discovery GWAS [27] but excluding those within a 500 kb window of APOE. The \( p < 1 \) selection threshold was based on previous polygenic score models for AD, verbal-numerical reasoning (cognitive ability), and educational attainment [11, 28]. In these studies, while \( p < 1 \) was not the optimal threshold for AD and verbal-numerical reasoning \( (p < 0.5 \) and \( p < 0.05 \), respectively), there were negligible differences with the results for the \( p < 1 \) threshold. A total of 539,368 genotyped Generation Scotland SNPs (with MAF < 5%) were used to construct the score using weights from the Stage I analysis of Lambert et al. [27]. The Lambert et al. study was a meta-analysis GWAS of the 1000 Genomes imputed SNPs (nSNPs > 7,000,000). After excluding 2,581 SNPs within a 500 kb region of APOE, we mapped the remaining SNPs to the overlapping genotyped variants in Generation Scotland.
A summary of the methods and acknowledgements from the discovery GWAS [27] are presented in the Supplementary Material.

**Ethics**

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). Generation Scotland has also been granted Research Tissue Bank status by the Tayside Committee on Medical Research Ethics (REC Reference Number: 15/0040/ES), providing generic ethical approval for a wide range of uses within medical research.

**Statistical analyses**

Linear mixed modelling was used to test for differences in cognitive ability by AD polygenic risk scores and APOE status. A mixed modelling framework is necessary to account for potential relatedness between participants; familial relationships were fitted using a pedigree-based kinship matrix. The polygenic score was entered as either a continuous variable or as ventiles (5% groupings) of risk. A fully adjusted model added self-reported educational attainment, hypertension, stroke, diabetes, heart disease, and depression, along with a measure of social deprivation (Scottish Index of Multiple Deprivation) [12]. A sample size of 3,495 is sufficient to detect an effect size with an $R^2$ of 0.18% for a type-I error of $\alpha = 0.05$ at 80% power using a one-sided test. APOE was entered as a factor with e3 homozygotes as the reference category for all other haplotype combinations.

All analyses were conducted in R, using the ‘pwr’, ‘kinship2’, and ‘coxme’ packages [29–32].

**RESULTS**

**Description of the polygenic risk score cohort**

(n = 3,625, age-range 60–70 years)

A demographic summary of the target population aged between 60 and 70 years and with AD polygenic risk scores is presented in Table 1. The median age of the cohort was 63 (IQR 61–65) and 57% were female. The mean BMI of the cohort was 27.5 kg/m² (SD 5.0). The median educational attainment was 12–13 years (measured categorically). The self-reported health questionnaire identified 27% of participants with self-reported hypertension, 9% with depression, 6% with diabetes, 2% with stroke, and 8% with heart disease.

There was a statistically significant association between the polygenic score and memory (Table 2): effect size of –0.31 points per SD of the polygenic score, SE 0.24, $p = 0.025$. A similar effect size was observed for processing speed although it was not significantly different from the null (effect size –0.27, SE 0.24, $p = 0.25$). There was no difference in polygenic score by APOE genotype (age- and sex-adjusted ANOVA $p = 0.72$). Moreover, the effect size for the polygenic score in the memory model remained significant and was not attenuated after adjusting for APOE haplotype (effect size –0.30 points, SE = 0.14, $p = 0.025$); there was also no evidence for an APOE x polygenic score interaction (likelihood ratio test $P = 0.40$). Similarly, there was no evidence of an APOE x polygenic score interaction for the processing speed model (likelihood ratio test $p = 0.86$). In the fully adjusted models, which controlled for self-reported diabetes, stroke, heart disease, diabetes, and

<table>
<thead>
<tr>
<th>Variable</th>
<th>Polygenic risk cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years – median, IQR)</td>
<td>3,495 63 61–65</td>
</tr>
<tr>
<td>Digit Symbol Test</td>
<td>3,495 62.5 14.4</td>
</tr>
<tr>
<td>Logical Memory</td>
<td>3,495 29.5 8.0</td>
</tr>
<tr>
<td>SIMD (rank, median, IQR)*</td>
<td>3,318 4866 2924–5542</td>
</tr>
<tr>
<td>Educational attainment¹</td>
<td>3,365 4 3–5</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>1,998 57.2</td>
</tr>
<tr>
<td>Self-report hypertension (yes)</td>
<td>929 26.6</td>
</tr>
<tr>
<td>Self-report stroke (yes)</td>
<td>79 2.3</td>
</tr>
<tr>
<td>Self-report diabetes (yes)</td>
<td>194 5.6</td>
</tr>
<tr>
<td>Self-report heart disease (yes)</td>
<td>285 8.2</td>
</tr>
<tr>
<td>Self-report depression (yes)</td>
<td>298 8.5</td>
</tr>
<tr>
<td>$APOE$ ε2ε2</td>
<td>19 0.5</td>
</tr>
<tr>
<td>$APOE$ ε2ε3</td>
<td>437 12.5</td>
</tr>
<tr>
<td>$APOE$ ε2ε4</td>
<td>86 2.5</td>
</tr>
<tr>
<td>$APOE$ ε3ε3</td>
<td>2,081 59.5</td>
</tr>
<tr>
<td>$APOE$ ε3ε4</td>
<td>782 22.4</td>
</tr>
<tr>
<td>$APOE$ ε4ε4</td>
<td>90 2.6</td>
</tr>
</tbody>
</table>

*Scottish Index of Multiple Deprivation. ¹Education was measured as an ordinal variable, so median and quartiles are reported. 0: 0 years, 1: 1–4 years, 2: 5–9 years, 3: 10–11 years, 4: 12–13 years, 5: 14–15 years, 6: 16–17 years, 7: 18–19 years, 8: 20–21 years, 9: 22–23 years, 10: ≥24 years.
depression, along with educational attainment and a social deprivation index, there was a slight increase in the effect size of the polygenic score on both the memory and processing speed measures: effect sizes of −0.34, SE 0.14, \( p = 0.014 \) and −0.31, SE 0.24, \( p = 0.20 \), respectively.

**Cognitive differences in the top versus bottom 5% of the polygenic score distribution (age-range 60–70 years)**

A significant association was observed in the age- and sex-adjusted analyses that compared the top and bottom ventile (5%) of the polygenic distribution for memory differences. Those in the top (highest AD risk) ventile scored a mean of 1.8 points (SE 0.8, \( p = 0.028 \)) lower than those in the bottom ventile on the memory test; for processing speed, those in the top ventile scored a mean of 2.3 points (SE 1.5, \( p = 0.13 \)) lower than the bottom ventile.

**Cognitive differences by APOE status (n = 3,625, age-range 60–70 years)**

In a regression of cognitive ability on age, sex, and APOE, \( e4e4 \) homozygotes scored a mean of 2.8 and 3.5 points lower on memory and processing speed (\( p = 0.001 \) and \( p = 0.022 \), respectively) compared to \( e3e3 \) homozygotes.

**Sensitivity and secondary analyses**

While a kinship matrix was included to model relatedness between participants, a sensitivity analysis on only unrelated individuals was performed. A genetic relationship matrix was created in GCTA and unrelated individuals (relationship coefficient <0.025) were retained (\( n = 2,677 \)). In this sub-group, we observed results consistent with the primary analysis (Supplementary Table 1).

A second sensitivity analysis was run after excluding those with fewer than 5 years of education (\( n = 12 \)) or a missing value for education (\( n = 130 \)). These results were consistent with the primary analysis (Supplementary Table 2).

To determine if cognitive decrements by AD polygenic scores were present at younger ages, we selected an analysis sub-cohort in the age range of 45 to 60 years (\( n = 6,853 \)). We observed generally smaller effect sizes to the 60 to 70 sub-group that were all non-significant (Supplementary Table 3). Similarly, we observed null associations between the polygenic score and cognitive decrements in a sub-group of participants aged over 70 years (Supplementary Table 4).

**DISCUSSION**

In a group of over 3,000 individuals aged between 60 and 70 years, polygenic risk scores for AD were associated with decrements for memory but not processing speed. This was the case when considering polygenic risk on a continuum and also when comparing the extremes (top and bottom 5%) of the distribution. Furthermore, a higher AD polygenic risk score was associated with an increased odds of family history of AD in an extended sample of 6,724 unrelated participants of all ages. A significant association was only present when comparing the extremes of the distribution rather than a continuous polygenic score. This increased risk was independent of APOE status. Relative to \( e3e3 \) homozygotes (59.5% of the study population), APOE \( e4e4 \) homozygotes (2.6% of the study population) carried a lower risk of familial AD than those in the top 5% of the AD polygenic

---

Table 2

Comparison of cognitive outcomes by genetic risk for AD and APOE status. All models adjust for age, sex, and pedigree-based relatedness

<table>
<thead>
<tr>
<th>Variable</th>
<th>beta</th>
<th>SE</th>
<th>( p )</th>
<th>FDR Adjusted ( p^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect per SD of PGRS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digit Symbol Test</td>
<td>−0.28</td>
<td>0.24</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Logical Memory</td>
<td>−0.31</td>
<td>0.14</td>
<td>0.020</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Top versus Bottom 5% of PGRS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digit Symbol Test</td>
<td>−2.32</td>
<td>1.54</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>Logical Memory</td>
<td>−1.84</td>
<td>0.83</td>
<td>0.028</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>APOE e4e4 versus e3e3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digit Symbol Test</td>
<td>−3.51</td>
<td>1.53</td>
<td>0.022</td>
<td>0.04</td>
</tr>
<tr>
<td>Logical Memory</td>
<td>−2.78</td>
<td>0.86</td>
<td>1.2 × 10^{-3}</td>
<td>0.007</td>
</tr>
</tbody>
</table>

PGRS, Polygenic risk score; SD, standard deviation; SE, standard error. *False discovery rate adjusted \( p \)-values after applying a Benjamini-Hochberg correction to the six empirical \( p \)-values.
burden compared to those in the bottom 5%. Furthermore, although a relatively large point estimate was observed in the expected direction, the $e4e4$ association was not significantly associated with family history of AD, unlike $e4$ presence (versus absence). This is likely to be due to a lack of statistical power.

The main limitation of the current study is the sample size. The post hoc power calculations showed that the total number of participants in the 60–70 age range was only just sufficient to detect relatively small memory decrements by AD polygenic score status. The relatively modest association $p$-values for the primary analyses (Table 2) reflect this lack of power. The associations remained significant after a FDR correction; only the $APOE$ association with Logical Memory would remain significant after a Bonferroni correction ($p < 0.05/6$).

Another possible limitation is the construction of the AD polygenic risk predictor. As the number of cases and controls increases in the discovery GWAS [27], the precision and reliability of the SNP regression weights will improve. The cross-sectional design of the Generation Scotland analysis may also be a limitation, as might the lack of information on subjective memory complaints. One recent study showed that a high genetic score for AD (based on 22 top SNP hits from a GWAS study) was associated with steeper decline in memory, although the magnitude of the effect was reduced when the $APOE$ locus was removed from the score [33].

With sufficiently large sample sizes, it is likely that cognitive differences in processing speed will be present in the general population for those with high versus low polygenic risk of AD. Larger discovery GWAS studies will also help to identify the optimal number of SNPs (all SNPs in a truly polygenic architecture versus a smaller number of possibly more biologically informative SNPs) for a polygenic predictor. The genetic contribution to AD has been shown to overlap with the genetics of education, intelligence, and income but not other health, disease, or psychiatric outcomes [28, 34, 35]. Intuitively, we would therefore expect to see phenotypic differences across all ranges of the polygenic scores and more acutely with the extremes of the distribution.

The most comprehensive study to have examined the association between polygenic scores for AD with cognitive function [27] used a predictor based on the Lambert et al. discovery GWAS [11]. The independent target dataset in that study was the UK Biobank study. Small but significant associations, not explaining more than 0.05% of the variance in three cognitive traits and 0.07% of the variance in educational attainment [28].

In conclusion, there is potential clinical utility for the stratification of mid-to-late-life population-based cohorts into high and low risk groups (based on $APOE$ status and global polygenic risk) to better understand the pathophysiology of AD. However, large sample sizes for both the GWASs used to build the polygenic scores and to select at risk sub-groups of the population are likely to be necessary. By contrast, smaller sample sizes are likely to be required when stratifying by $APOE$ $e4e4$ status, as effect sizes are far greater in magnitude. Nonetheless, with increasingly powerful polygenic predictors—as a result of bigger baseline GWAS studies—it seems likely that the extremes of the distribution will provide high risk groups equivalent to those with two $e4$ alleles. However, the extremes of the polygenic score distribution will be of additional value as, by definition of their construction, they will tap into genome wide risk and multiple pathways that lead to AD. Longitudinal collection of cognitive test data in addition to biomarker panels and ‘omics data, such as methylyomics, which have been linked to AD pathology [36] may help illuminate biological signatures for AD, and improve long-term prediction of the disease.

ACKNOWLEDGMENTS

This work was supported by multiple sources. Generation Scotland received core funding from the Chief Scientist Office of the Scottish Government Health Directorates CZD/16/6 and the Scottish Funding Council HR03006. We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the UK Medical Research Council (MRC). The Quantitative Trait Locus team at the Human Genetics Unit are funded by the MRC. REM, SPH, DJP, PMV, and IJD undertook the work within The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (MR/K026992/1), part of the cross council Lifelong Health and...
SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-161070.

REFERENCES


