Escott-Price, Valentina and Sims, Rebecca and Bannister, Christian and Harold, Denise and Vronskaya, Maria and Majounie, Elisa and Badarinarayan, Nandini and Morgan, Kevin and Passmore, Peter and Holmes, Clive and Powell, John and Lovestone, Simon and Brayne, Carol and Gill, Michael and Mead, Simon and Goate, Alison and Cruchaga, Carlos and Lambert, Jean-Charles and van Duijn, Cornelia and Maier, Wolfgang and Ramirez, Alfredo and Holmans, Peter and Jones, Lesley and Hardy, John and Seshadri, Sudha and Schellenberg, Gerard D. and Amouyel, Philippe and Williams, Julie (2015) Common polygenic variation can predict risk of Alzheimer’s disease. Brain . ISSN 1460-2156

Common polygenic variation can predict risk of Alzheimer's disease

Valentina Escott-Price*[1], Rebecca Sims[1], Christian Bannister[1], Denise Harold[2], Maria Vronskaya[1], Elisa Majounie[1], Nandini Badarinarayan[1], GERAD/PERADES, IGAP consortia, Kevin Morgan, Peter Passmore, Clive Holmes, John Powell, Simon Lovestone, Carol Brayne, Michael Gill, Simon Mead, Alison Goate, Carlos Cruchaga, Jean-Charles Lambert, Cornelia van Duijn, Wolfgang Maier, Alfredo Ramirez, Peter Holmans[1], Lesley Jones[1], John Hardy[3], Sudha Seshadri[4], Gerard D Schellenberg[5], Philippe Amouyel[6,7,8,9], Julie Williams*[1]

*Corresponding Authors

1. Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, UK
2. School of Medicine, Trinity College Dublin, College Green, Dublin 2, Ireland
3. Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, Institute of Neurology, London, UK.
4. Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA
5. Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 19104, USA
6. Inserm U744, Lille, 59000, France
7. Université Lille 2, Lille, 59000, France
8. Institut Pasteur de Lille, Lille, 59000, France
9. Centre Hospitalier Régional Universitaire de Lille, Lille, 59000, France

Address of corresponding authors:

Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, CF24 4HQ, UK

E-mails: EscottPriceV@cardiff.ac.uk & WilliamsJ@cardiff.ac.uk

Running title: Prediction of Alzheimer’s disease risk

Abstract

Background: The identification of subjects at high risk for Alzheimer’s disease is important for prognosis and early intervention. We investigated the polygenic architecture of Alzheimer’s disease (AD) and the accuracy of AD prediction models, including and excluding the polygenic component in the model.

Methods: This study used genotype data from the powerful dataset comprising 17,008 cases and 37,154 controls obtained from the International Genomics of Alzheimer’s Project (IGAP). Polygenic score analysis tested whether the alleles identified to associate with disease in one sample set were significantly enriched in the cases relative to the controls in an independent sample. The disease prediction accuracy was investigated by means of sensitivity, specificity, Area Under the receiver operating characteristic Curve (AUC) and positive predictive value (PPV).

Results: We observed significant evidence for a polygenic component enriched in Alzheimer’s disease ($p=4.9x10^{-26}$). This enrichment remained significant after *APOE* and other genome-wide associated regions were excluded ($p=3.4x10^{-19}$). The best prediction accuracy AUC=78% was achieved by a logistic regression model with *APOE*, the polygenic score as predictors and age. When looking at the genetic component only, the PPV was 81%, increasing to 82% when age was added as a predictor. Setting the total normalised polygenic score of greater than 0.91, the positive predictive value has reached 90%.

Conclusion: Polygenic score has strong predictive utility of Alzheimer’s disease risk and is a valuable research tool in experimental designs, e.g. for selecting Alzheimer’s disease patients into clinical trials.

Key words: Alzheimer's disease, polygenic score, predictive model.

Introduction

Genome-wide association (GWA) studies have proved a powerful method to identify susceptibility alleles for complex diseases. The most powerful currently undertaken study, provided by the International Genomics of Alzheimer's Project (IGAP), has identified over twenty AD susceptibility loci (Lambert *et al.*, 2013). GWA study datasets can be used to determine a polygenic contribution of common SNPs that show disease association but fail to meet the accepted P-value threshold for genome-wide significance ($p<5x10^{-8}$). Recent studies confirm that the estimated heritability detected in AD GWA studies (24-35%) (Lee *et al.*, 2013) increases substantially when weak effect loci are also considered. This strongly implies that a large proportion of the genetic signal must lie below the genome wide significance threshold.

The Polygenic score (PS) approach encompasses more of the causal variance, as a genetic risk score is calculated based not solely on genome-wide significant polymorphisms, but on all nominally associated variants at a defined significance threshold (typically thousands of variants). This type of analysis has recently shown significant polygenic contribution in other complex genetic diseases. For example in Parkinson disease, a polygenic basis was confirmed and shown to correlate with age at disease onset (Escott-Price *et al.*, 2014). The method can also be used to identify overlap in genetic determinants between related disorders, e.g. schizophrenia and bipolar disorder; depression and anxiety (Demirkan *et al.*, 2011). While the polygenic method undoubtedly introduces noise by including some variants that are not involved in disease susceptibility (i.e. false positives), this is more than offset by the increased power to identify those at highest/lowest risk of disease. Trait differences between those with highest/lowest polygenic risk scores have also been identified. For example, in a study of the Lothian Birth Cohort, increased polygenic risk of schizophrenia was associated with lower cognitive ability at age 70 and greater relative decline in general cognitive ability between the ages of 11 and 70 (McIntosh *et al.*, 2013).

We investigated the polygenic architecture of Alzheimer's disease using the powerful IGAP GWA dataset (Lambert *et al.*, 2013). The IGAP dataset was split into two independent subsets before the polygenic contribution to AD was investigated by assessing whether score alleles identified in one subset were significantly enriched in cases from another subset.

We also investigated the prediction accuracy of the model, which includes the number of ε4 and ε2 alleles at the *APOE* gene, a PS component based upon genome-wide significant (GWS) loci, and a PS component constructed using all independent markers within the dataset including statistically not-significant SNPs. Furthermore we looked at the utility of the PS when the analysis was restricted to subjects with ε2 and ε3 alleles only. As age is a strong predictor of AD, we tested the prediction models in samples stratified by age. To test the sensitivity of the prediction models to population differences we ran the same analyses for subjects from UK, USA and Germany separately.

Materials and Methods

We used the discovery dataset reported by the IGAP consortium (Lambert *et al.*, 2013) , comprising of 17,008 AD cases and 37,154 controls. This sample of AD cases and controls comprises 4 data sets taken from GWA studies performed by GERAD, EADI, CHARGE and

ADGC (Lambert *et al.*, 2013). Full details of each study including the samples and methods utilised are provided elsewhere (Harold *et al.*, 2009, Lambert *et al.*, 2009, Seshadri *et al.*, 2010, Hollingworth *et al.*, 2011, Naj *et al.*, 2011). Each of the 4 datasets were imputed with either Impute2(Howie *et al.*, 2009) or MACH(Li *et al.*, 2010) software, using the 1000 genomes data (release Dec2010) as a reference panel.

Polygenic score analysis

We followed the approach previously described by the International Schizophrenia Consortium (International Schizophrenia *et al.*, 2009). The PS analysis requires two independent datasets. For the first, result data is sufficient as this dataset is used to select the SNPs, the risk score alleles and their genetic effects. The second dataset is used to test whether the polygenic risk scores differ in cases and controls and requires the genotypes for each individual. The meta-analysed results data of the EADI, CHARGE and ADGC consortia (13,831 cases and 29,877 controls, hereafter referred to as IGAP.noGERAD) was used for SNP selection. We used the individual genotypes of the GERAD consortium (Harold *et al.*, 2009) data (3,177 cases and 7,277 controls), we used the GERAD data as the test sample.

We included only autosomal SNPs that passed stringent quality control criteria, i.e. minor allele frequencies (MAF) ≥ 0.01 and imputation quality score greater than or equal to 0.5 in each study. This resulted in 6,928,531 SNPs, which were present in at least 40% of the AD cases and 40% of the controls, being included in the analysis. The summary statistics across the 3 datasets were combined using fixed-effects inverse variance-weighted meta-analysis.

Using GERAD study data we performed a) random linkage disequilibrium (LD) pruning using $r^2>0.2$, and b) “intelligent” pruning (--clump option in PLINK (Purcell *et al.*, 2007) genetic analysis tool) using the same r^2 parameter and a physical distance threshold for clumping SNPs of 1Mb. The random LD pruning resulted in 401,584 SNPs that are in relative linkage equilibrium ($r^2 \leq 0.2$) and common between GERAD and IGAP.noGERAD datasets. The "intelligent" pruning allows to capture SNPs which are most (even if not-significantly) associated with the disease in an LD block. This "intelligent" pruning identified 538,363 independent SNPs that were most significantly associated with AD in IGAP.noGERAD data. We selected markers, based upon significance thresholds, to construct a polygenic score in the GERAD data. The PS was calculated from the effect size (β)-weighted sum of associated alleles within each subject. PS were normalised by subtracting the mean and dividing by the standard deviation.

We assessed a variety of significance thresholds for the selection of markers for PS construction; overlapping panels of markers were used (e.g. significant at $p \leq 0.01$, 0.05, 0.1, …, 1 in the IGAP.noGERAD) in the construction of a subject-level score in GERAD case/control sample. The ability of each panel-based score distribution to distinguish those with disease from cognitively normal individuals was assessed using logistic regression analysis while adjusting for three principal components (Harold *et al.*, 2009), reflecting underlying stratification in the sample due to population and/or genotyping technique differences. Age was not included as a covariate in the logistic regression models as it had already been accounted for as a covariate in the IGAP.noGERAD meta-analyses.

Analysis of predictive accuracy

To find the best predictors of the AD, we tested a variety of regression models. For this analysis we used the genotyped (rather than imputed) SNP data as we note that the prediction accuracy is sensitive to the number of missing genotypes, which is often exacerbated by the uncertainty of imputation.

Since the genotyped data at the *APOE* locus contained only proxy SNPs for the *APOE*-ε4 and *APOE*-ε2 variants (rs429358 and rs7412), we limited our analysis to those individuals (3,049 cases and 1,554 controls) for whom we had *APOE* genotype data. For the other 21 GWS SNPs (Lambert *et al.*, 2013), proxies with r^2 greater than 0.8 were available for 11 SNPs in the GERAD data, for an additional 7 loci we had genotyped markers that were in modest LD (r^2 between 0.5 and 0.8) with a GWS marker. Two GWS SNPs in the *SLC24A4/RI*N3 and *CD33* loci had proxies with r^2~0.3 (Supplementary Table 1). We excluded the *DSG2* gene as this association did not replicate in IGAP stage 2(Lambert *et al.*, 2013), and the best proxy to the putative GWS SNP was in low LD (r^2=0.06) in the GERAD sample.

We calculated sensitivity, specificity, area under the receiver operating characteristic curve (AUC) and positive and negative predictive values (PPV and NPV) by comparing the observed case/control status and the predicted probability estimated by logistic regression models using the *prediction()* and *performance()* functions in R-statistical software. We used as predictors a number of explanatory variables including *APOE*-ε4, *APOE*-ε2, age, PS based upon 20 GWS SNP proxies, and PS calculated using SNPs with AD association p-values ranging from 0.0001 till 0.9 in the IGAP.noGERAD sample (APOE and GWAS loci were excluded, see Supplementary Table 1). We performed similar analyses on imputed data however the prediction accuracy using this dataset was marginally lower due to noise introduced through a number of missing values as a result of genotypes imputed with low certainty (results are not shown). To test the sensitivity of our results to possible bias due to age and population stratification, we ran the same models in subsamples stratified by geographical region (UK, USA and Germany), and age groups <60, 60-69, 70-79, 80-89 and 90+.

Results

Polygenic risk score analysis

In this study we investigated whether the PS alleles identified in one AD GWA study were significantly enriched in the cases relative to the controls of an independent AD dataset. Our analysis revealed significant evidence for an overall enrichment of the AD polygenic risk score alleles of the IGAP.noGEARD data in the independent GEARD (Harold *et al.*, 2009) cohort of 3,177 AD cases and 7,277 controls from the UK, Europe and USA (Table 1). The pattern of the PS association was similar to those seen in studies of other complex diseases shown to have a polygenic signal (International Schizophrenia *et al.*, 2009, Stergiakouli *et al.*, 2012, Heilmann *et al.*, 2013, Michailidou *et al.*, 2013). Our most significant evidence for association was observed when SNPs with a selection threshold (P_T) of $p \leq 0.5$ in the IGAP.noGERAD sample were included. The p-values for a significant enrichment in the polygenic score ranged from $3.9x10^{-20}$ to $4.9x10^{-26}$ dependent on the P_T used (Table 1). For all significant associations the B-coefficients were positive, indicating that a higher polygenic score in the IGAP.noGERAD discovery dataset corresponds to a higher score in the independent GERAD replication dataset and provides evidence for a polygenic contribution to the development of Alzheimer’s disease.

Since the 538,363 independent SNPs that we used to identify AD polygenic risk score alleles included those most significantly associated with the disease, it is plausible that our results are artificially biased by SNPs whose evidence for association is a consequence of LD with a known genome-wide significant SNPs. To investigate this possibility we repeated our analysis using identical analysis thresholds but excluding all 5,006 SNPs that, after LD pruning, were present at the 24 genomic regions previously reported to be strongly associated with AD (Lambert *et al.*, 2013, Escott-Price *et al.*, 2014). The regions were defined as ±500KB of both sides of the GWA SNPs (Lambert *et al.*, 2013) or GWA genes (Escott-Price *et al.*, 2014) and between 44,400KB-46,500KB on chromosome 19 for the *APOE* locus (Supplementary Table 1). Given that each of these excluded regions is likely to contain at least one true AD susceptibility allele, this approach is highly conservative. Nevertheless, this analysis again revealed significant evidence that individuals with higher polygenic risk scores had greater probability of AD, with our most significant result $p=3.4x10^{-19}$ (Table 2). Moreover, we obtained analogous results when we used an alternative method of LD pruning, which ignores the strength to which SNPs are associated with AD, and thus excludes SNPs from the 24 associated regions (Supplementary Table 2). These analyses suggest that our findings are not dependent on either the previously identified susceptibility loci or the SNPs that are associated with AD merely as a consequence of LD with the GWS loci.

Analysis of predictive accuracy

The identification of subjects at high risk for Alzheimer's disease is important for prognosis and early intervention. We used logistic regression analysis to establish predictive values (sensitivity, specificity, AUC, PPV, NPV) of genetic risk factors in GERAD data. The results of this analysis are summarised in Table 3. A highly significant ($p<10^{-94}$) overall outcome was obtained for all measures of predictive accuracy (Table 3). The *APOE*-ε4 allele is the strongest known genetic risk factor for AD. In the presence of *APOE*-ε4 alleles, the sensitivity was 0.59, the specificity 0.75 and the AUC=0.678. Inclusion of the numbers of

APOE-ε2 alleles in the logistic regression model slightly increases all prediction accuracy values, in particular, the AUC increased to 0.688. As expected, prediction accuracy was further enhanced (AUC=0.715) when we added the polygenic score variable based upon proxies for the 20 GWS SNPs, where the weights of the SNP risk alleles were identified from the independent dataset IGAP.noGERAD (Supplementary Figure 1) .

We further investigated whether the PS based on risk alleles of small effect identified in one study (IGAP.noGERAD) were improving the prediction accuracy in an independent dataset (GERAD). For this we used PS calculated excluding the known AD associated regions (Supplementary Table 2). The best prediction accuracy (AUC=0.75) was achieved when we included the PS for SNPs with AD association p-values<0.5. The values of sensitivity and specificity (the proportion of cases and controls, respectively, which were correctly predicted) were about 0.69 when estimated with the minimized difference threshold MDT=0.64 (see Supplemental Figure 2). If we reduce the probability threshold to 0.47, the percentage of correctly identified cases increases to 0.9, at a cost of specificity (0.35) (see Supplemental Figure 2). To investigate possible population differences in the prediction of AD risk, we looked at UK, German and USA subjects separately. The pattern of predictive modelling results was similar to the main analyses results in all strata (Supplementary Table 3). Interestingly, the prediction in the USA strata was extremely good (the best AUC=0.95%). This might be due to the fact that the majority of subjects (about 80%) in the training set were of USA origin in contrast to 17% in the test set.

Another way to look at the utility of the PS as a predictor for AD, is to exclude the strongest predictor, namely the ε4 allele, from the analysis. There were 1242 cases and 1160 controls in the sample without ε4 allele. When looking at these individuals only, the AUC was 65.0% when we included the PSs based upon proxies for the 20 GWS SNPs and for SNPs with AD association p-values<0.5, increasing to 65.8% when the number of ε2 alleles was added as a predictor. Similar accuracy was achieved (64.5% and 65.8%) when we ran the analysis on the whole sample without ε4 as a predictor.

As expected, our results show that inclusion of age in the regression model further improved the prediction accuracy (AUC=0.78), see Table 3 and Supplementary Figure 2. In the context of practical application, e.g. in experimental designs comparing cases with high or low polygenic risk AD, age has to be taken into account. Supplementary Table 4 presents the results of the genetic predictive modelling stratified by age groups. The results of the stratified analyses have shown similar pattern of prediction accuracy. As before, the best accuracy in each strata was achieved when the numbers of *APOE*-ε4, *APOE*-ε2 alleles, the PS variable based upon proxies for the 20 GWS SNPs, and the PS for SNPs with AD association p-values<0.5 were included as predictors. The AUC value was ranging from 73% to 79%, with the highest in the 60-69 age group (Supplementary Table 4). The best prediction in this age group might indicate that this particular age group has the strongest common genetic effect, with the younger age group (<60) potentially due to Mendelian forms of the disorder, and the older age groups confounded by general ageing effects.

With regard to the practical use of PS in the identification of subjects at high risk for AD, we investigated the prediction accuracy of the genetic component in terms of positive predictive value (PPV), the percentage of patients with a positive prediction who actually have the disease. To achieve PPV of 0.9, i.e. have 90% of predicted cases to actually be cases, the prediction probability threshold has to be set to 0.87. This prediction probability threshold captured cases with normalised total PS of greater than 0.91. The total PS combines effects of ε4, ε2, 20 GWAS proxy SNPs and AD associated SNPs (p<0.5), which comprised the best prediction model in our analysis.

Discussion

The molecular genetic data reported in this study provides strong support for a large polygenic contribution to the overall heritable risk of Alzheimer's disease. This implies that the genetic architecture of AD includes many common variants of small effect that is likely to reflect a large number of susceptibility genes and a complex set of biological pathways related to disease. The AD PS alleles identified in the GERAD cohort are not significantly enriched (minimum p=0.14) in an independent GWA study for Parkinson's disease (Moskvina *et al.*, 2013) indicating that the identified polygenic component of AD is disease specific.

Further studies are required if we are to progress from the knowledge that there is a polygenic contribution to AD, to understanding the specific genetic factors that comprise the polygenic component. Increasing the discovery sample size will allow more loci with increasingly small individual effect sizes to pass the threshold of genome-wide significance, and should substantially refine the polygenic scores derived here. Moreover, as we have previously shown, using approaches such as gene pathways analyses it is possible to utilise the captured polygenic signal and identify genes or biological systems relevant to AD (International Genomics of Alzheimer's Disease, 2014).

It is possible that our findings are influenced by rare AD susceptibility variants that are in LD with the common alleles analysed in this study. The ongoing efforts of studies performing exome and whole genome sequencing in large numbers of AD case/control cohorts will allow us to establish the haplotype structure of common and rare alleles an in turn, to understand which loci are subject to 'synthetic association'(Dickson *et al.*, 2010). Moreover, as previously demonstrated in other complex diseases (Purcell *et al.*, 2014), future PS analysis of variants identified by exome/genome sequencing are expected to further inform our understanding of the genetic underpinnings of AD.

One possible limitation of this study that the population structure in the training set is only moderately representative of the test set, due to differences in proportions of subjects from different countries.

In conclusion, the derived polygenic scores have demonstrated utility for calculating an individual level genetic risk profile that can predict disease development. Measures of polygenic burden could prove useful in distinguishing AD patients whose disease liability is most likely to carry a large or small genetic component. This utility of the developed polygenic score is increased among subjects of 60-69 years of age, which is a desirable target group for identification and preventative intervention of AD. Identifying these individuals would benefit study recruitment into clinical trials and could facilitate a better understanding of how gene-gene and gene-environment interactions increase risk for AD.

Acknowledgements

We thank the IGAP consortium for providing summary statistics for the train dataset. Cardiff University was supported by the Medical Research Council (MRC) grant (MR/K013041/1), the EU Joint Programme – Neurodegenerative Disease Research (JPND) grant (MR/L501517/1), the Alzheimer’s Research UK (ARUK) grant (ARUK-PG2014-1), and the NISCHR funded Centre for Ageing and Dementia Research (CADR). Christian Bannister and Elisa Majounie were supported by the Medical Research Council (MRC) via the UK Dementia’s Platform (UKDP, reference MR/L023784/1).

Table 1. Results of polygenic score analysis based upon a set of independent SNPs (at $r^2 \leq 0.2$) pruned to retain those most significantly associated with the disease.

P_T*	Effect	SE	p	R^2	NSNPs
0.01	0.283	0.0308	3.9E-20	0.016	16,749
0.05	0.311	0.0308	5.9E-24	0.019	61,552
0.1	0.321	0.0309	2.6E-25	0.020	107,834
0.2	0.327	0.0309	3.6E-26	0.021	185,737
0.3	0.317	0.0308	7.9E-25	0.020	251,850
0.4	0.323	0.0308	1.0E-25	0.020	308,780
0.5	**0.327**	**0.0310**	**4.9E-26**	**0.021**	**359,500**
0.6	0.326	0.0310	6.2E-26	0.021	404,626
0.7	0.325	0.0309	9.3E-26	0.020	444,663
0.8	0.328	0.0310	4.1E-26	0.021	480,271
0.9	0.323	0.0309	1.9E-25	0.020	511,297
1	0.321	0.0309	3.0E-25	0.020	538,362

*Selection threshold of ‘score’ SNPs taken from the IGAP.noGERAD discovery sample.

Table 2. Results of polygenic score analysis based upon a set of relatively independent SNPs (at $r^2 \leq 0.2$) pruned to retain those most significantly associated with the disease, excluding the genome-wide associated loci. (Exact positions of the excluded regions are given in Supplementary Table 1.)

P_T*	Effect	SE	p	R^2	NSNPs
0.01	0.154	0.0304	4.01×10^{-7}	0.005	16,412
0.05	0.232	0.0305	2.50×10^{-14}	0.011	60,750
0.1	0.256	0.0307	5.92×10^{-17}	0.013	106,587
0.2	0.270	0.0307	1.23×10^{-18}	0.014	183,808
0.3	0.263	0.0305	6.47×10^{-18}	0.014	249,314
0.4	0.271	0.0306	7.26×10^{-19}	0.014	305,741
0.5	**0.275**	**0.0307**	$\mathbf{3.45 \times 10^{-19}}$	**0.015**	**356,033**
0.6	0.274	0.0307	4.66×10^{-19}	0.015	400,785
0.7	0.273	0.0307	6.76×10^{-19}	0.014	440,473
0.8	0.276	0.0308	2.93×10^{-19}	0.015	475,769
0.9	0.271	0.0307	1.13×10^{-18}	0.014	506,532
1	0.269	0.0307	1.67×10^{-18}	0.014	533,356

*Selection threshold of ‘score’ SNPs taken from the IGAP.noGERAD discovery sample.

Table 3. Predictive accuracy for 3,049 AD cases vs 1,554 controls. The PS' were constructed using independent SNPs associated with AD in IGAP.noGERAD at different significance levels (MODEL column), excluding *APOE* and 20 GWAS regions (see Supplementary Table 2). Numbers of SNPs participating in the predictive model are given in column N SNPs.

MODEL	N SNPs	Sensitivity	Specificity	AUC	PPV*	NPV**
ε4	1	0.593	0.746	0.678	0.821	0.483
ε4 + ε2	2	0.593	0.746	0.688	0.821	0.483
ε4 + ε2+ 20 GWAS SNPs + PS p<0.0001	130	0.669	0.669	0.717	0.798	0.507
ε4 + ε2+ 20 GWAS SNPs + PS p<0.001	549	0.668	0.668	0.720	0.798	0.506
ε4 + ε2+ 20 GWAS SNPs + PS p<0.01	3388	0.672	0.672	0.729	0.801	0.511
ε4 + ε2+ 20 GWAS SNPs + PS p<0.05	13273	0.677	0.677	0.738	0.804	0.516
ε4 + ε2+ 20 GWAS SNPs + PS p<0.1	23676	0.682	0.682	0.740	0.808	0.522
ε4 + ε2+ 20 GWAS SNPs + PS p<0.2	42273	0.683	0.683	0.743	0.808	0.523
ε4 + ε2+ 20 GWAS SNPs + PS p<0.3	58963	0.684	0.683	0.744	0.809	0.524
ε4 + ε2+ 20 GWAS SNPs + PS p<0.4	73941	0.684	0.684	0.744	0.809	0.525
ε4 + ε2**+ 20 GWAS SNPs + PS p<0.5**	**87605**	**0.686**	**0.686**	**0.745**	**0.811**	**0.527**
ε4 + ε2+ 20 GWAS SNPs + PS p<0.6	99724	0.685	0.685	0.745	0.810	0.526
ε4 + ε2+ 20 GWAS SNPs + PS p<0.7	110431	0.685	0.685	0.745	0.810	0.525
ε4 + ε2+ 20 GWAS SNPs + PS p<0.8	119616	0.683	0.683	0.745	0.809	0.523
ε4 + ε2+ 20 GWAS SNPs + PS p<0.9	127585	0.684	0.684	0.745	0.809	0.524
ε4 + ε2+ 20 GWAS SNPs + PS p<0.5+age	**87605**	**0.702**	**0.701**	**0.781**	**0.822**	**0.545**

* Positive Predictive Value

** Negative Predictive Value

Reference

Demirkan A, Penninx BW, Hek K, Wray NR, Amin N, Aulchenko YS, et al. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. Molecular psychiatry. 2011;16(7):773-83.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS biology. 2010;8(1):e1000294.

Escott-Price V, Bellenguez C, Wang LS, Choi SH, Harold D, Jones L, et al. Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease. PloS one. 2014;9(6):e94661.

Escott-Price V, IPDGC, Nalls M, Morris H, Lubbe S, Brice A, et al. Common polygenic variation contributes to risk of Parkinson’s Disease and is correlated with disease age at onset. Annals of Neurology. 2014.

Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nature genetics. 2009;41(10):1088-93.

Heilmann S, Brockschmidt FF, Hillmer AM, Hanneken S, Eigelshoven S, Ludwig KU, et al. Evidence for a polygenic contribution to androgenetic alopecia. The British journal of dermatology. 2013;169(4):927-30.

Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nature genetics. 2011;43(5):429-35.

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009;5(6):e1000529.

International Genomics of Alzheimer's Disease C. Convergent genetic and expression data implicate immunity in Alzheimer's disease. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2014.

International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.

Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nature genetics. 2009;41(10):1094-9.

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature genetics. 2013;45(12):1452-8.

Lee SH, Harold D, Nyholt DR, Consortium AN, International Endogene C, Genetic, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Human molecular genetics. 2013;22(4):832-41.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology. 2010;34(8):816-34.

McIntosh AM, Gow A, Luciano M, Davies G, Liewald DC, Harris SE, et al. Polygenic risk for schizophrenia is associated with cognitive change between childhood and old age. Biological psychiatry. 2013;73(10):938-43.

Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013;45(4):353-61, 61e1-2.

Moskvina V, Harold D, Russo G, Vedernikov A, Sharma M, Saad M, et al. Analysis of genome-wide association studies of Alzheimer disease and of Parkinson disease to determine if these 2 diseases share a common genetic risk. JAMA neurology. 2013;70(10):1268-76.

Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nature genetics. 2011;43(5):436-41.

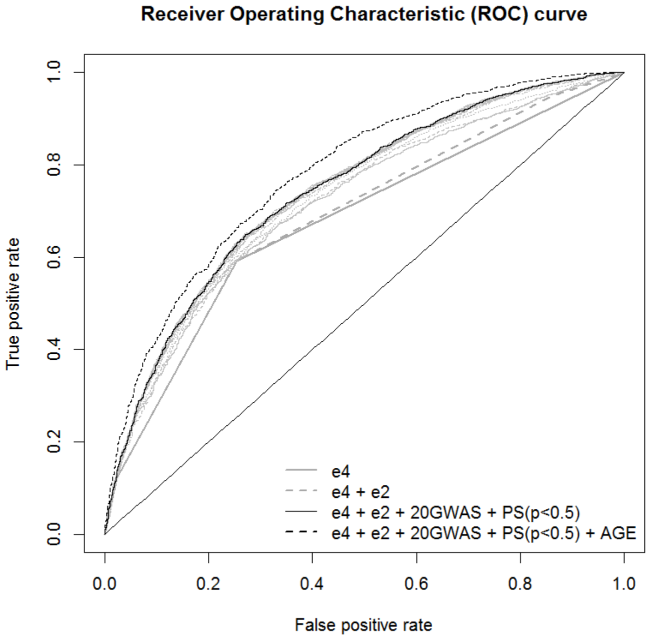
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.

Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506(7487):185-90.

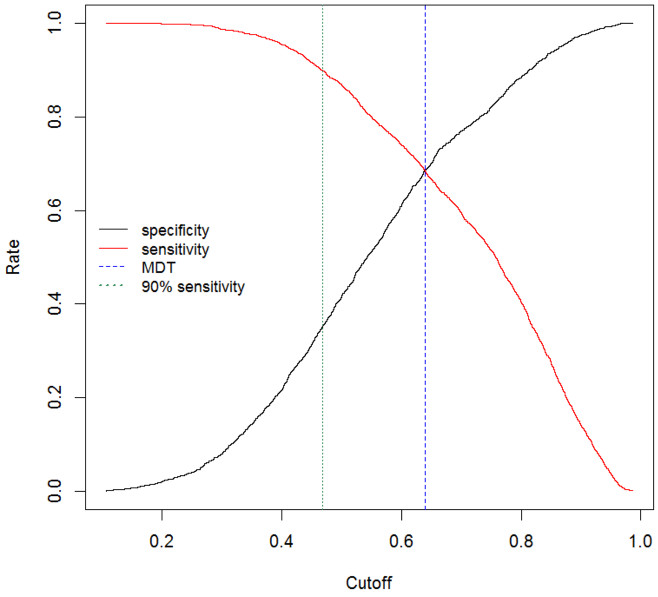
Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA : the journal of the American Medical Association. 2010;303(18):1832-40.

Stergiakouli E, Hamshere M, Holmans P, Langley K, Zaharieva I, de CG, et al. Investigating the contribution of common genetic variants to the risk and pathogenesis of ADHD. The American journal of psychiatry. 2012;169(2):186-94.

Supplemental Figure 1. ROC curves for predictive models with different predictors for risk of Alzheimer's disease.



Supplemental Figure 2. Sensitivity-Specificity plot for the best predictive model which includes e4, e2, the polygenic score variable based upon proxies for the 20 GWS SNPs and the PS for SNPs with AD association p-values<0.5.


Rate
specificity
sensitivity
MDT
90% sensitivity
Cutoff