Polygenic Risk Score Analysis of Alzheimer's Disease in APOE3 homozygotes.

Valentina Escott-Price PhD[1], Amanda Myers PhD[2], Matt Huentelman PhD[3] and John Hardy PhD[4]*.

1. Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, UK.
2. Department of Psychiatry & Behavioral Sciences, Programs in Neuroscience and Human Genetics and Genomics and Center on Aging, Miller School of Medicine, University of Miami, Miami, FL USA
3. Neurogenomics Division, The Translational Genomics Research Institute (TGen), Phoenix, AZ 85004
4. Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, Institute of Neurology, London, UK.

address for correspondence at j.hardy@ucl.ac.uk

Keywords
Alzheimer's disease, genetics, pathology, APOE

Running title: Prediction of Alzheimer's disease in APOE3 homozygotes

Word count: 1482

Abstract

We and others have previously shown that polygenic risk score analysis (PRS) has considerable predictive utility for identifying those at high risk of developing Alzheimer's disease (AD) with areas under the curve of >0.8. However, by far the greatest determinant of this risk is the apolipoprotein E locus with the E4 allele alone giving an AUC of ~0.68 and the inclusion of the protective E2 allele increasing this to ~0.69 in a clinical cohort. An important question is to determine how good PRS is at predicting risk in those who do not carry the E4 allele (E3 homozygotes, E3E2 and E2E2) and in those who carry neither the E4 or E2 allele (i.e. E3 homozygotes). We have tested this in a cohort of pathologically confirmed AD cases and controls by taking out of the analysis, those individuals carrying an E4 allele and those carrying either an E4 or an E2 allele. This had surprisingly little effect on the PRS (AUC ~ 0.83 [95% CI: 0.80-0.86]). We carried out an additional analyses taking out all SNPs within 1MB of the APOE locus and this had no further effect on the AUC. From a practical perspective this suggests that PRS analysis will have predictive utility even in E4 negative individuals.

Introduction

Polygenic risk score (PRS) analysis enhances the predictability of the diagnosis of AD (1). In a recent PRS analysis, we showed that the area under the curve (AUC) in a pathologically confirmed case/control series was 0.84 (2). However, by far the largest contribution to this risk analysis is the E4 allele (risk) and the E2 allele (protective) which gave AUC of 0.68 (E4 alone) and 0.69 (E4+E2) as compared to overall PRS AUC=0.75 in clinical sample (1). An important practical and theoretical consideration is to understand how good PRS is when the risk at the APOE locus is removed. When this was tested in clinical series (1)

the AUC was reduced from 0.75 in the whole dataset to 0.65 in E3 homozygotes. We determined to test this in our pathological series by removing from the analysis, first all E4 carriers and then, all E4 and E2 carriers from both the case and the control data sets. In a subsequent analysis we repeated these analyses in the same samples removing all SNPs from within 1Mb of the APOE locus to ensure we were not capturing other risk at this locus.

Methods

The sample characteristics of the original dataset used in this study were the same as in our previous analysis (3). This project was declared IRB exempt (MedstarProject #2003-118) under the Code of Federal Regulations, 45 CFR, 46. The primary data consisted of 1011 cases and 583 controls. The total number of imputed single nucleotide polymorphisms (SNPs) was 36,481,940. The number of SNPs with Info score above 0.8 was 11,016,052. From these, the number of SNPs with MAF>=0.01 was 7,868,100 and these were used in the analysis (see ref 4). Association analysis was performed for each SNP using logistic regression analysis as implemented in snptest (5). We eliminated first all those samples who had an E4 allele (leaving 354 cases and 454 controls) and then additionally those with both an E4 and an E2 allele (leaving 321 cases and 365 controls).

Results

The original AUC for the full pathologically confirmed dataset was 0.84, when adjusted for the overlap between the training set, used for SNPs selection and the pathologically confirmed dataset (2). The original unadjusted AUC was 0.866. In this paper, for the full pathologically confirmed dataset, we estimate AUC=0.73 [0.71-0.75] for E4 and AUC= 0.75 [0.73-0.77] for E4 and E2. Removing all individuals with an E4

allele only reduced the unadjusted AUC from 0.866 to 0.834 and then removing all E2 carriers (i.e. restricting analysis to E3 homozygotes) had a further small effect and reduced the AUC to 0.831. Taking out all the SNPs at the APOE locus (chr 19:44,400-46,500KB) had no further effect on the predictive value. Thus, in contrast to the results obtained with clinical series, the AUC is only marginally reduced by removing APOE4 and APOE2 carriers.

We tested three possible explanations for this finding: first, the people who get AD without an E4 allele have more of the AD risk alleles, i.e. allele at other loci have a bigger effects in the absence of APOE4;second, the effects of APOE and other risk SNPs are independent and third, the results are driven by inflation due to the overlap between the discovery (IGAP) and test (E3E3 pathologically confirmed AD cases and controls) datasets.

First we ran GWAS analysis with snptest software only for E3E3 cases and controls. The majority of the top (IGAP) SNPs did not show statistically significant association in this small sample and their effect sizes were not higher than the effect sizes in the whole data set (data not shown). This strongly suggests that the E3 homozygotes with disease do not have a greater excess of other AD risk alleles.
Next we counted the numbers of risk and alternative alleles for sets of SNPs at different significance thresholds (reported by the IGAP study) for each subject in the E3 homozygotes and in the rest of the sample. We compared the average numbers (per person) of risk and alternative alleles by means of chi-square test for 2x2 table: (Risk Allele - Alternative Allele) x (E3 homozygotes – other genotypes). This analysis was performed in cases and controls separately as cases in general may have more risk alleles than controls. The results are summarized in Table 1. The were no significant differences in the mean numbers of

risk and alternative alleles per person among E3 homozygotes versus the other genotypes in either pathologically confirmed AD cases nor pathologically confirmed controls.

We also compared the prediction accuracy of the best model (PRS for SNPs with p-values≤0.5) with and without APOE locus in three subsamples, namely E4 carriers (644 cases and 115 controls), E4 and E2 carriers (677 cases and 204 controls) and E3 homozygotes (321 cases and 365 controls). Table 2 shows the estimated AUC in those sub-samples for the PRS models with more significant SNPs (p≤0.001 in IGAP study) and best predictive PRS model (1), combining all available independent SNPs with p-values ≤0.5, when APOE region is included and excluded. The results clearly show that the PRS prediction accuracy is almost the same in any sub-sample, when APOE is excluded. Note that the full sample (shown in the second column of Table 3), has the largest overlap with IGAP, and therefore the AUC estimate there has the most (~2%) inflation (see (2) for details).

Finally we adjusted our main result (AUC=0.831 in E3 homozygotes) for the overlap with the discovery IGAP dataset using simulation approach as described in (2). The overlap between IGAP dataset and our E3 homozygote dataset was 1.3% (686 individuals in E3E3 dataset out of 54,162 individuals in the IGAP data). We simulated 1000 times effect sizes of the pruned set of SNPs with mean $b \sim N(B_{IGAP}, sd=0.12*SE_{IGAP})$, where B_{IGAP} is the beta-coefficient and SE_{IGAP} is the standard error for that SNP in the IGAP study, and the coefficient 0.12 was estimated empirically (see (2) for details). The adjusted AUC and the confidence intervals were calculated as average AUC and CI over the 1000 simulations, AUC_{ADJ} = 0.83 [95% CI: 0.80-0.86].

Figure 1 shows distribution of standardized PRS of E3E3 cases and E3E3 controls. In the group of negative polygenic extreme (PRS smaller than -2), there were 17 controls and 0 cases. In the positive extreme group (PRS greater than 2), there were 11 cases and 1 control. Looking at the extremes (PRS < -1.5) and (PRS > 1.5), there were 1 case and 49 controls and 41 cases and 4 controls, respectively.

Discussion

Our results show that the prediction accuracy of PRS in the pathologically confirmed sample of E3 homozygotes carriers is high and equivalent to the prediction accuracy in the samples of in the whole dataset. This finding indicates APOE is an independent risk factor for the disease. This result is in contrast to the PRS observed in clinical cohorts where restricting analyses to E3 homozygotes resulted in a large reduction in the PRS. We believe this is likely to be because of poor diagnostic accuracy among those labeled as AD in the absence of an E4 allele: this interpretation consistent with post mortem follow up of Alzheimer clinical trials which suggested a diagnostic inaccuracy of up to 25%. In this context, all results derived from the analysis of clinically defined E4 negative cases should be interpreted with caution. From a mechanistic perspective, this result suggests that the genetic architecture of AD is similar in E3 homozygotes to that in other genotypes since a similar proportion of risk is captured by PRS in all genotypes. This result offers little support to the belief that E3 homozygotes with Alzheimer's disease have more predisposing variants at other loci.

Acknowledgements

This manuscript is dedicated to the memory of our colleagues who worked on generating these data:- Christopher B. Heward and Jason

J. Corneveaux. We thank the patients and their families for their selfless donations. The data generation for this project was supported by funding from Kronos Science. Additional funding was from the National Institutes of Health as well as NIH EUREKA grant R01-AG-034504 to AJM and AG041232 (NIA) to AJM and MH as well as Intramural funds NIH (JH and AJM). Analytical work was supported the MRC JPND PERADES grant MR/L501517/1 (JH and VEP).

Many data and biomaterials were collected from several National Institute on Aging (NIA) and National Alzheimer's Coordinating Center (NACC, grant #U01 AG016976).. A full listing off collection sites is given in ref. 4.

Author contributions

VEP carried out the PRS analysis. AM and MH generated the original data and quality controlled it for this analysis. JH designed the study and wrote the original draft. All authors obtained funds for the study and analysis and reviewed the drafts.

Potential Conflict of Interest

JH is a co-grantee of Cytox Genetics from Innovate UK (UK Department of Business).

References

1) Escott-Price, V., Sims, R., Bannister, C et al 2015. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain 138, 3673e3684.

2) Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic Risk Score Analysis of Pathologically Confirmed Alzheimer's Disease. Ann Neurol. 2017 Jul 20. doi: 10.1002/ana.24999. [Epub ahead of print]

3) Corneveaux, J.J., Myers, A.J., Allen, A.N. et al, 2010. Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. Hum. Mol. Genet. 19, 3295e3301.

4) Marchini J, Howie B, Myers S, McVean G and Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39 : 906-913

5) Lee SH, Yang J, Chen GB, et al. Estimation of SNP heritability from dense genotype data. Am J Hum Genet. 2013 Dec 5;93(6):1151-5.

6) Escott-Price V, Shoai M, Pither R, Williams J, Hardy J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. NeurobiolAging. 2017 Jan;49:214.e7-214.e11

Table 1. Comparison of the mean numbers of risk and alternative alleles per person in E3 homozygotes vs. other AD in cases and controls. APOE region is excluded.

Cases

SNP selection threshold	E3 homozygotes		Other genotypes		OR	P
	Mean No of risk alleles	Mean No of alternative alleles	Mean No of risk alleles	Mean No of alternative alleles		
0.0001	92.5	63.3	91.0	63.6	1.016	1
0.001	306.6	282.2	299.9	283.8	1.022	0.859
0.01	1777.9	1576.4	1766.1	1580.1	1.007	0.874
0.05	7348.8	6333.3	7315.1	6346.3	1.005	0.794
0.1	13329.7	11633.7	13300.3	11661.5	1.002	0.805
0.2	23935.0	21032.8	23868.4	21084.1	1.003	0.701
0.3	33149.4	29575.3	33081.5	29657.7	1.002	0.673
0.4	41612.3	37349.4	41561.4	37436.3	1.001	0.729
0.5	49267.5	44544.9	49197.0	44643.4	1.001	0.697

Controls

SNP selection threshold	E3 homozygotes		Other genotypes		OR	P
	Mean No of risk alleles	Mean No of alternative alleles	Mean No of risk alleles	Mean No of alternative alleles		
0.0001	87.6	66.8	88.4	67.7	0.991	1.00
0.001	293.8	293.2	292.8	295.4	1.003	0.973
0.01	1727.4	1621.3	1723.0	1623.0	1.003	0.961
0.05	7181.6	6461.3	7153.8	6471.3	1.004	0.833
0.1	13070.4	11836.9	13037.7	11860.3	1.003	0.810
0.2	23540.4	21344.5	23466.5	21388.0	1.003	0.703
0.3	32649.2	29969.2	32573.0	30039.3	1.002	0.684
0.4	41028.5	37805.2	40983.2	37869.4	1.001	0.785
0.5	48618.4	45048.2	48564.7	45118.6	1.001	0.776

Table 2. AUC for PRS models with IGAP-based p-value SNP selection thresholds 0.001 and 0.5. These results are not unadjusted for IGAP overlap (see text).

PRS model	AUC and 95% Confidence intervals in "[]"			
	whole sample	E4 carriers	E4E2 carriers	E3 homozygotes
PRS with SNPs p≤0.001	0.741 [0.72-0.78]	0.616 [0.56-0.67]	0.743 [0.70-0.78]	0.632 [0.59-0.67]
PRS with SNPs p≤0.5	0.866* [0.85-0.89]	0.831 [0.78-0.88]	0.868 [0.84-0.90]	0.831 [0.80-0.86]
PRS with SNPs p≤0.001 APOE region excluded	0.637 [0.61-0.67]	0.565 [0.51-0.62]	0.625 [0.58-0.67]	0.646 [0.61-0.69]
PRS with SNPs p≤0.5 APOE region excluded	0.840 [0.82-0.86]	0.821 [0.77-0.87]	0.837 [0.80-0.87]	0.834 [0.80-0.86]

* This AUC is reported in (2).

Figure 1. Distribution of standardised and PCA adjusted PRS in E3E3 cases and controls.