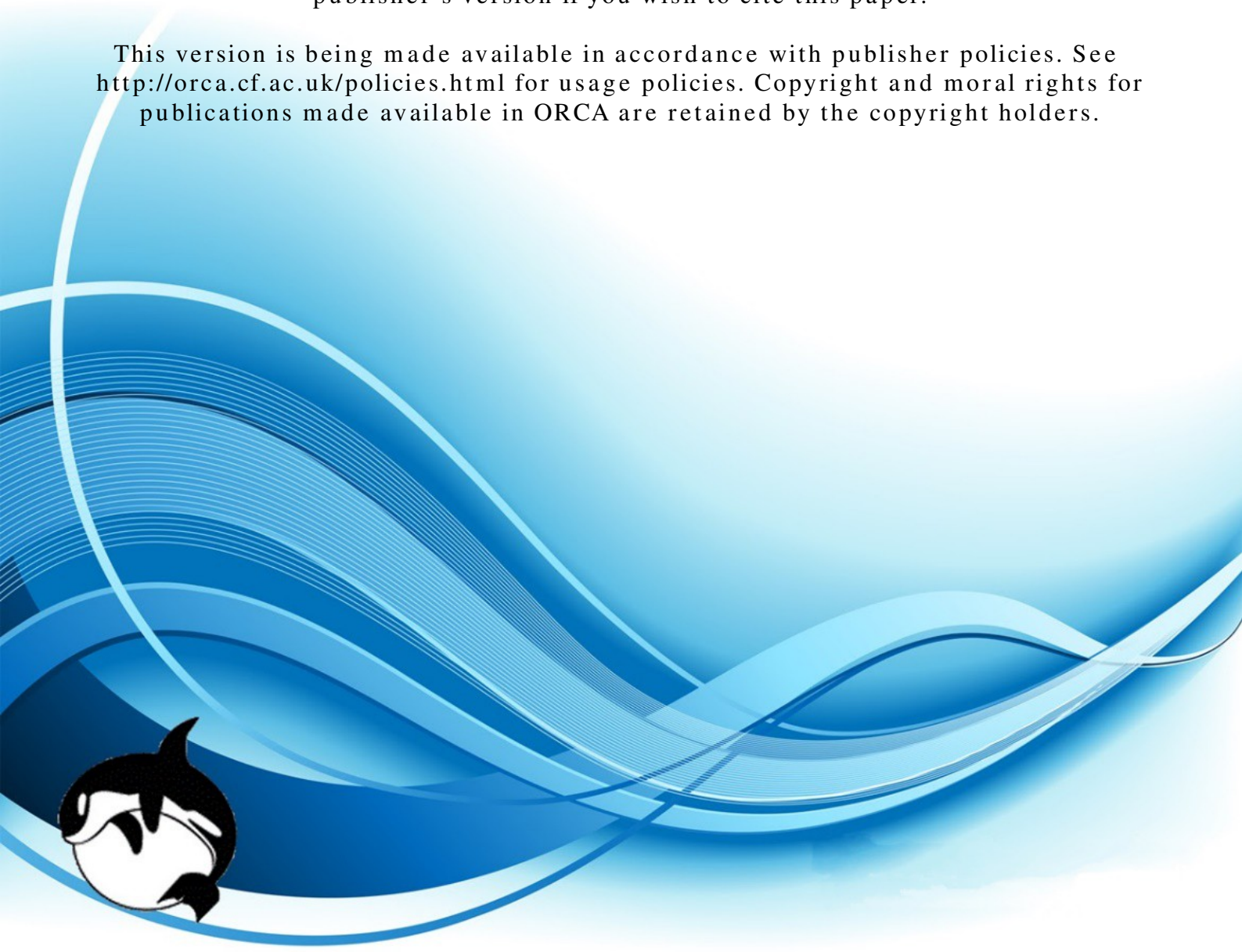

Polygenic Risk Score Analysis of Pathologically Confirmed Alzheimer's Disease

Valentina Escott-Price PhD[1], Amanda J. 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

Title 111 characters
Running title 47 characters
Abstract 76 words
Introduction 257 words
Discussion 311 words
Word count 1373

Abstract

Previous estimates of the utility of polygenic risk score analysis for the prediction of Alzheimer's disease have given Area Under the Curve estimates of <80%. However, these have been based on the genetic analysis of clinical case control series. Here we apply the same analytic approaches to a pathological case control series and show a predictive AUC of 84%. We suggest that this analysis has clinical utility and that there is limited room for further improvement using genetic data.

Introduction

Polygenic risk score (PRS) analysis enhances the predictability of the diagnosis of Alzheimer's disease (AD) over the use of just the apolipoprotein E locus (1). In a recent PRS analysis, we showed that the area under the curve (AUC) in the recent genome wide association study (GWAS), was 0.79 (1). However, the study samples in these cohorts were largely comprised of clinical cases of AD, and the diagnostic accuracy of these is not perfect as recent clinical trial failures have highlighted (2). In addition, the majority of controls which are used in GWAS, are sampled from a general population and are often underaged to develop AD. This diagnostic uncertainty has also been demonstrated by the observation of c9orf72 expansions (a locus causing frontotemporal dementia) within some of the clinical AD cohorts used in the generation of the GWAS and AD sequencing data (3).

Having a better understanding of the diagnostic utility of PRS is of importance for two reasons: first, because it enables the accurate assessment of how much risk for disease there is still left to be found and this is important in setting research goals, and second, because this type of analysis could be used in the refinement of inclusion criteria for clinical trials and eventually, in clinical health care recommendations.

We have previously reported a GWAS in clinically characterized and neuropathologically confirmed samples of AD and matched controls (4): in this analysis, we apply PRS to these pathological data to determine whether some of the "missing heritability" of AD is due to clinical misdiagnosis.

Methods

The sample characteristics of the dataset used in this study were the same as in our original analysis. This project was declared IRB exempt (Medstar Project #.2003-118) under the Code of Federal Regulations, 45 CFR, 46. Eight cases and eight controls had corrupted data files and were omitted (4). This left 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. Association analysis was performed for each SNP using logistic regression analysis as implemented in snptest (5).

We performed predictive modelling using polygenic score based upon SNPs with p-value cut-off $p=10^{-4}$, 10^{-3}, 0.01, 0,05, … 0.5 as in (7) as predictor variables. These sets of SNPs are capturing APOE and index GWAS SNPs (7) either directly or via their proxies. For prediction modelling we converted imputed "dosage" genotypes in our data to "most probable genotype" with probability over 90%. The individual polygenic risk scores were generated as sum of the risk alleles weighted by effect sizes as in the International Genomics of Alzheimer's Project (IGAP) study (7), then were further adjusted for first 10 principal components and standardized. The models were fitted using the above mentioned individual polygenic risk scores and predicting AD/control status in our study. This is the most powerful way of testing the prediction ability of the strongest genetic predictors to date: however our study was part of the IGAP study (7) (~3% overlap) and therefore the results will be marginally overfitted. We accounted for this overfitting in our analysis as below.

Since summary statistics for the IGAP (7) data excluding our sample was not available to us, we estimated the effect of possible bias using simulations. For that we first simulated a sample of 17008 cases and 37154 controls, matching the IGAP stage-I study, for a typical SNP with minor allele frequencies=0.2 and effect size of odds ratio (OR)=1.05. This OR matches the average effect size for IGAP pruned SNPs with association p-value≤0.5, mean(B_{IGAP})=0.05, mean(SE_{IGAP})=0.035; OR_{IGAP}=exponential(0.05)=1.05. Then we randomly removed 1101 cases and 583 controls (matching our study size) and recalculated the association effect size 1000 times de novo. The "removal-based-simulated" effect sizes for a single typical SNP were

found to be normally distributed with mean B_{SIM}=0.05 (SD_{SIM}=0.004) (not shown). Assuming that the removed sample is a random subset of cases and controls, the expected distribution of the IGAP pruned SNPs effect sizes should have the same mean but slightly increased standard error:

$$SE_{IGAP\text{-}ADJ}=SE_{IGAP}*sqrt(N)/sqrt(N-N_o),$$

where N is the IGAP sample size, N_o is the overlap sample size. In particular, we can roughly expect the mean$(SE_{IGAP})_{ADJ}$=0.035/sqrt(0.97)=0.0355, where 0.035 is the mean(SE_{IGAP}) of the effect size for IGAP pruned SNPs with association p-value≤0.5.

To adjust prediction modelling for overlapping samples, we ran further simulations where the effect sizes for each SNP in the IGAP study were simulated as b~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. The sd=0.12*SE_{IGAP} was chosen empirically to allow for both the variability due to IGAP B-coefficient estimate and due to random subsample removal. As a rough example, multiplying the mean$(SE_{IGAP})_{ADJ}$ by 0.12 results in a standard deviation, which is approximately matching the "removal-based-simulated" SD_{SIM}: 0.12*mean$(SE_{IGAP})_{ADJ}$ =0.12*0.0355 ≈ 0.004= SD_{SIM}. Thus, in each simulation step, each SNP in IGAP had a simulated effect size and p-value corresponding to this effect size. Then the SNPs were reselected, repruned and the polygenic scores recalculated. The prediction accuracy of the simulated PRS was calculated at each simulation (N simulations =1000) and mean of simulated AUC for SNPs with p≤0.5 was reported and is discussed below.

Results

The primary results (QQ-plot and Manhattan plot) were consistent with our previous analysis of these data (4). There were no genome-wide significant hits apart from APOE locus.

We compared the results of our analysis with 21 index genome-wide significant SNPs identified in the IGAP (7) study (see Table 1). Sixty three percent of IGAP GWAS index SNPs (14 out of 22) show larger effects in our dataset compared to the original report (7), of which 5 have significantly larger effect sizes (see the last column of Table 1), including the two SNPs tagging the APOE status.

The results of predictive modelling are presented in Table 2. Training on the whole IGAP the prediction accuracy AUC reaches 86% (Figure 1) when all SNPs with p≤0.5 are included in the model, However, as discussed above there is an element of overfitting in this analysis as our data was part of the IGAP analysis. Accounting for this possible inflation using simulation (see Methods), the prediction accuracy is 84% (95%CI 82-86%).

Discussion

These data systematically confirm, in the context of genome wide data, our results examining the APOE locus (4): genetic prediction is better in the context of autopsy confirmed cases and controls. This has implications for our view of how much genetic variability remains to be found: in an earlier analysis, we estimated that the theoretical maximal genetic variance to be found would generate an AUC of 82% (95% confidence interval 78%-85%) (8). The figure now identified, based on the genome wide analysis of a pathological cohort is 84% (95% confidence interval 82-86%). Thus the theoretical and assessed the figures for risk prediction accuracy overlap and both are larger than the AUC of 0.75 assessed using clinical cohorts (1). There is thus further evidence that polygenic risk profiling captures the SNP-heritability very well with regards to common variation in AD, although of course, heritability estimates (8) were constructed on clinical diagnoses of AD so strict comparisons are hazardous. This does not imply that there are no genetic findings of very rare variants (f<0.1%) still to be made, although the increasing predictability of genetic findings (9-12) and the fact that most new findings relate to already identified pathways implies that research may be better focused on the targeted sequencing of established pathways, bioinformatic analyses of multi-omics data sets, and cell biology rather than on large scale genome wide sequencing projects in unrelated sporadic AD individuals. These data also illustrate that there is a degree of misdiagnosis in the clinical AD series (3), and even more so in population based controls.

A final implication of these data is that genome wide genotyping and PRS based analytic strategies are reasonably effective at predicting those who will develop disease. They also suggest that this predictive utility is unlikely to improve much more. This strategy may therefore now be useful for designing clinical trials and eventually in clinical practice.

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 from Innovate UK (UK Department of Business) and VEP was (2015-2016) a consultant for Cytox who are developing an Affymetrix based genetic testing array for Alzheimer's disease.

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) Karran E, Hardy J. A critique of the drug discovery and phase 3 clinical programs targeting the amyloid hypothesis for Alzheimer disease. Ann Neurol. 2014 Aug;76(2):185-205.

3) Majounie E, Abramzon Y, Renton AE, et al. Repeat expansion in C9ORF72 in Alzheimer's disease. N Engl J Med. 2012 Jan 19;366(3):283-4.

4) 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.

5) 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

6) Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009;4:1088-93.

7) Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45:1452-8.

8) 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.

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

10) Jones L, Holmans PA, Hamshere ML, et al. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. PLoS One. 2010 Nov 15;5(11):e13950. doi: 10.1371

11) Matarin M, Salih DA, Yasvoina M, et al A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. Cell Rep. 2015 Feb 3;10(4):633-44.

12) Huang K-L, Marcora E, Pimenova A, et al. A common haplotype lowers SPI1 (PU.1) expression in myeloid cells and delays age at onset for Alzheimer's disease. bioRxiv. doi: https://doi.org/10.1101/110957. Posted March 21, 2017.

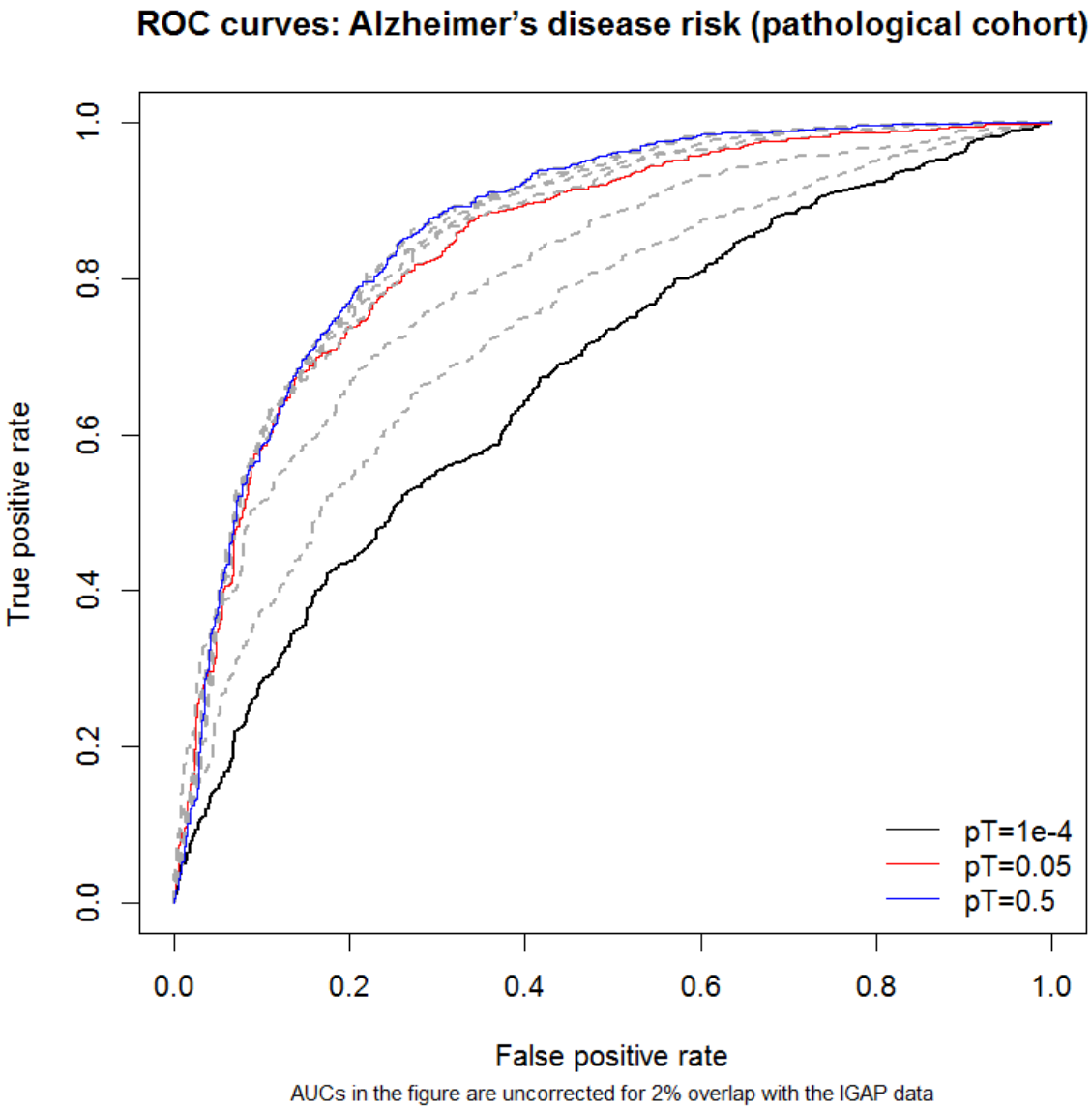
ROC curves: Alzheimer's disease risk (pathological cohort)
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
pT=1e-4
pT=0.05
pT=0.5
AUCs in the figure are uncorrected for 2% overlap with the IGAP data

Table 1. Comparison of neuropathologically confirmed data analysis with IGAP genome-wide significant index SNPs (Lambert et al 2013)

					IGAP				Corneveaux				
SNP	CHR	BP	A1	A2	BETA	SE	OR	P	BETA	SE	OR	P	P_test.DIFF
rs6656401	1	207692049	A	G	0.167	0.017	1.181	5.69E-24	0.251	0.096	1.285	0.0093	0.805
rs4663105	2	127891427	A	C	-0.184	0.017	0.832	1.00E-26	0.026	0.085	1.026	0.763	0.992
rs6733839	2	127892810	T	C	0.197	0.014	1.217	6.94E-44	NA	NA	NA	NA	NA
rs35349669	2	234068476	T	C	0.076	0.014	1.078	3.17E-08	0.076	0.080	0.927	0.3445	0.501
rs190982	5	88223420	G	A	-0.076	0.014	0.927	3.23E-08	-0.119	0.082	0.888	0.1456	0.302
rs10948363	6	47487762	G	A	0.095	0.015	1.100	5.20E-11	0.121	0.088	0.886	0.1664	0.614
rs2718058	7	37841534	G	A	-0.077	0.013	0.926	4.76E-09	0.025	0.079	0.975	0.7499	0.900
rs1476679	7	100004446	C	T	-0.089	0.014	0.915	5.58E-10	-0.188	0.082	0.829	0.0224	0.119
rs11771145	7	143110762	A	G	-0.102	0.014	0.903	1.12E-13	-0.171	0.087	1.186	0.0503	0.218
rs28834970	8	27195121	C	T	0.100	0.013	1.105	7.37E-14	0.009	0.080	0.991	0.9079	0.133
rs9331896	8	27467686	C	T	-0.146	0.014	0.864	2.77E-25	-0.128	0.079	0.880	0.1057	0.586
rs10838725	11	47557871	C	T	0.079	0.014	1.082	1.12E-08	0.018	0.084	0.982	0.8273	0.239
rs983392	**11**	**59923508**	**G**	**A**	**-0.108**	**0.013**	**0.898**	**6.14E-16**	**-0.290**	**0.078**	**1.336**	**0.0002**	**0.011**
rs10792832	11	85867875	A	G	-0.140	0.013	0.869	9.32E-26	-0.173	0.079	0.842	0.0288	0.342
rs11218343	11	121435587	C	T	-0.262	0.034	0.770	9.73E-15	-0.461	0.187	1.586	0.0136	0.147
rs17125944	14	53400629	C	T	0.132	0.023	1.141	7.95E-09	-0.012	0.130	1.012	0.9282	0.137
rs10498633	**14**	**92926952**	**T**	**G**	**-0.095**	**0.016**	**0.910**	**5.54E-09**	**-0.271**	**0.088**	**1.311**	**0.0022**	**0.025**
rs8093731	18	29088958	T	C	-0.316	0.081	0.729	0.000105	0.229	0.409	0.795	0.5753	0.904
rs4147929	19	1063443	A	G	0.143	0.018	1.154	1.06E-15	0.112	0.097	1.119	0.2489	0.378
rs429358 (e4)	**19**	**45411941**	**T**	**C**	**-1.350**	**0.027**	**0.259**	**0**	**-1.748**	**0.115**	**0.174**	**8.2×10^{-52}**	**0.0004**
rs7412 (e4)	**19**	**45412079**	**T**	**C**	**-0.387**	**0.040**	**0.679**	**1.23E-22**	**-1.031**	**0.154**	**2.804**	**1.9×10^{-11}**	**2.46E-05**
rs3865444	**19**	**51727962**	**A**	**C**	**-0.067**	**0.014**	**0.935**	**2.97E-06**	**-0.223**	**0.083**	**1.250**	**0.0073**	**0.032**
rs7274581	20	55018260	C	T	-0.132	0.024	0.876	2.46E-08	-0.235	0.140	1.264	0.0934	0.235

Table 2. Predictive accuracy for 1101 clinically characterized and neuropathologically confirmed samples of AD and 583 controls. The PRS' were constructed using independent SNPs associated with AD in IGAP at different significance levels (MODEL column). Numbers of SNPs participating in the predictive model are given in column N SNPs.

MODEL	Effect	SE	p	NSNPs	Sensitivity	Specificity	**AUC**	AUC.L95	AUC.U95
1.00E-04	0.666	0.060	5.21E-29	299	0.617	0.617	**0.676**	0.649	0.703
0.001	0.981	0.067	1.23E-48	1184	0.686	0.686	**0.741**	0.716	0.766
0.01	1.385	0.078	1.14E-69	7030	0.734	0.734	**0.807**	0.786	0.829
0.05	1.740	0.092	2.37E-79	29017	0.770	0.770	**0.847**	0.827	0.867
0.1	1.813	0.094	2.59E-82	53329	0.770	0.770	**0.853**	0.834	0.873
0.2	1.861	0.096	8.24E-84	96791	0.775	0.775	**0.858**	0.839	0.878
0.3	1.899	0.097	3.23E-85	135642	0.789	0.789	**0.863**	0.843	0.882
0.4	1.931	0.098	1.32E-85	171672	0.785	0.786	**0.865**	0.846	0.884
0.5	1.943	0.099	8.22E-86	205068	0.790	0.791	**0.866**	0.847	0.886