Title: Genetic analysis suggests high misassignment rates in clinical Alzheimer's cases and controls

Article Type: Brief Communication

Section/Category: Alzheimer's Disease & Other Dementias

Keywords: Alzheimer; Genetics; Diagnosis; Biomarkers

Corresponding Author: Dr. John Hardy,

Corresponding Author's Institution: Institute of Neurology

First Author: Valentina Escott-Price, PhD

Order of Authors: Valentina Escott-Price, PhD; Emily Baker, MSc; Maryam Shoia, PhD; Amanda Myers, PhD; Matt Huentelman, PhD; John Hardy

Manuscript Region of Origin: UNITED KINGDOM

Abstract: Genetic case-control association studies are often based upon clinically ascertained cases and population or convenience controls. It is known that some of the controls will contain cases, as they are usually not screened for the disease of interest. However, even clinically assessed cases and controls can be missassigned. For Alzheimer disease (AD) it is important to know the accuracy of the clinical aassignment. The predictive accuracy of Alzheimer's disease risk by polygenic risk score analysis has been reported in both clinical and pathologically confirmed cohorts. The genetic risk prediction can provide additional insights to inform classification of subjects to case and control sets at a preclinical stage. In this study we take a mathematical approach and aim to assess the importance of genetic component for assignment of subjects to AD positive and negative groups, and provide an estimate of misassignment rates in AD case/control cohorts accounting for genetic prediction modelling results. We estimate misassignment rates of ~ 30% in both cases and in controls.

Genetic analysis suggests high misassignment rates in clinical Alzheimer's cases and controls

- Through comparisons of genetics analyses of clinical case control studies with pathological case control studies of Alzheimer's disease, we show that at typical ages fr case control studies (70-80 years), about 30% on clinically assigned cases are likely to be in the early stages of disease.
- Biomarker studies need to take this into account when carrying out case control comparisons

Genetic analysis suggests high misassignment rates in clinical Alzheimer's cases and controls

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

1. Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, UK.

2. Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, Institute of Neurology, London, UK.

3. 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

4. Neurogenomics Division, The Translational Genomics Research Institute (TGen), Phoenix, AZ 85004

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

1

Abstract

Genetic case-control association studies are often based upon clinically ascertained cases and population or convenience controls. It is known that some of the controls will contain cases, as they are usually not screened for the disease of interest. However, even clinically assessed cases and controls can be missassigned. For Alzheimer disease (AD) it is important to know the accuracy of the clinical aassignment. The predictive accuracy of Alzheimer's disease risk by polygenic risk score analysis has been reported in both clinical and pathologically confirmed cohorts. The genetic risk prediction can provide additional insights to inform classification of subjects to case and control sets at a preclinical stage. In this study we take a mathematical approach and aim to assess the importance of genetic component for assignment of subjects to AD positive and negative groups, and provide an estimate of misassignment rates in AD case/control cohorts accounting for genetic prediction modelling results. We estimate misassignment rates of ~ 30% in both cases and in controls.

Introduction

Genetic case-control association studies are often based upon clinical assessment of cases and population or convenience controls. It is clearly the case that some of the controls can potentially contain patients in the early stage of disease, as they are not typically screened for the disease. It is assumed that the number of controls, who are actually cases, is relatively small and can be estimated by the prevalence of the disease in the population (e.g. ~3% lifetime prevalence of AD).

Polygenic risk score (PRS) analysis enhances the predictability of the diagnosis of AD [Escott-Price at al 2015]. The largest contributors to AD risk analysis, the E4 allele (risk) and the E2 allele (protective) gave AUC of 0.68 (E4 alone) and 0.69 (E4+E2) as compared to overall PRS AUC=0.75 in clinical cohorts [*ibid*]. In a recent PRS analysis, we showed that the area under the curve (AUC) in a pathologically confirmed case/control series was 0.84 [Escott-Price et al. 2017]. In addition, in a case/control sample of pathologically confirmed individuals who carry neither the E4 or E2 allele (i.e. E3 homozygotes) the PRS gave AUC ~0.83 [95% CI: 0.80-0.86]) [Escott-Price et al submitted]. When this was tested in clinical series the AUC was reduced from 0.75 in the whole dataset to 0.65 in E3 homozygotes [*ibid*]. This reduction in PRS in the clinical but not pathological series is indicative of a substantial misassignment rate in the former.

A study at National Institute on Aging Alzheimer Disease Centers [Beach et al. 2012] had reported measures of agreement between stratified levels for the clinical and neuropathologic diagnosis of AD in a sample of 919 subjects, who were classified based on their clinical categorization as "probable AD," "possible AD," or "not AD." The "not AD" group included non-AD dementias and subjects with no dementia were excluded. The highest sensitivity (87.3%) reported in [Beach et al. 2012], was when the clinical diagnosis was defined as clinically probable or possible AD, and neuropathologic AD definition was defined as "frequent neuritic plaque density score" and Braak neurofibrillary tangle stage V or VI. In practice, most of cases in clinical case/control samples are collected with "probable AD" diagnosis. For this combination of clinical and neuropathologic criteria, analysis of mismatched clinical and neuropathologic diagnoses provides sensitivity of 76.6% [Beach et al. 2012]. This means that

when the clinical diagnosis was defined as probable AD and the neuropathologic diagnosis as frequent neuritic plaques with Braak stage V-VI, 23.4% of people did not have frequent neuritic plaque density, despite their positive clinical diagnoses. Furthermore, more than third of APOE4 noncarriers with clinical diagnosis of mild-to-moderate Alzheimer's dementia, had minimal Alzheimer's disease plaque accumulation in cerebral cortex [Monsell et al 2015].

In this study we aim to estimate misassignment rate in controls based upon genetic prediction accuracy in clinical and neuropathology confirmed samples of AD cases and controls. We did this because when GWAS are presented, a very frequent question asked is what proportion of controls are actually early cases? In this analysis we seek to answer that question. We derive mathematical formulae to compare case/control classification by clinical diagnosis and true pathology status accounting for a hidden layer of genetic classification between diseased subjects and controls. These formulae were used to illustrate the potential misassignment rates in clinical data samples, using the reported values of prediction (by PRS) accuracy in AD pathology confirmed samples of cases and controls [Escott Price et al. 2017].

Methods

Misassignment rate estimates in a clinical sample.

To estimate misassignment rates in case/control samples based upon PRS prediction accuracy and a neuropathologic examination, we derive analytical formulae. We first constructed three 2x2 contingency tables (also known as confusion matrices in the prediction modelling field), describing: 1) clinical AD diagnosis (case/control) vs PRS prediction (yes/no) in a clinical sample, 2) pathologically confirmed AD status (yes/no) vs PRS prediction (yes/no), and 3) pathologically confirmed AD status vs clinical diagnosis. The latter table was expressed in terms of prediction accuracy measures (sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)), estimated from clinical and pathologically confirmed samples (see Appendix).

To estimate the misassignment rate in controls, the analytical formulae require us to fix the parameter of AD misdiagnosis rate in cases. Since most of cases in clinical case/control samples are collected with "probable AD" diagnosis and in

4

the pathology confirmed study [Escott-Price at al 2017] the neuropathologic criterion for cases was Braak stage V or VI, we used sensitivity of 76.6% for AD misdiagnosis rates as reported in [Beach et al 2012]. In addition according to [Escott-Price et al submitted], among APOE4 non-carriers with the clinical diagnosis of mild-to-moderate AD, 37% had minimal neuritic plaques, and we used this value as an approximation of the misdiagnosis rate in the E3 homozygous cases.

Results

Estimation of misdiagnosis rates in a clinical sample

Assume that in a sample of N clinically screened subjects ($N_{cas}^{(c)}$ cases and $N_{con}^{(c)}$ controls), $N_{cas}^{(p)}$ and $N_{con}^{(p)}$ are the numbers of true cases and controls, that will be pathology confirmed (we use superscripts "(c)" and "(p)" to distinguish between the numbers of clinically and pathology based classifications, respectively). In this settings, the range for the numbers of subjects who were clinically and neuropathologically confirmed as AD, are between $\max\left\{0, N_{cas}^{(p)} - N_{con}^{(c)}\right\}$ and $\min\left\{N_{cas}^{(c)}, N_{cas}^{(p)}\right\}$. This means that in the worst case scenario, all clinical cases are in fact unaffected (zero overlap), and in the best case scenario all clinical cases were given the correct diagnosis and will be confirmed neuropathologically. Similarly, the range for the numbers of controls who were also neuropathologically confirmed as "no AD" is between $\max\left\{0, N_{con}^{(c)} - N_{cas}^{(p)}\right\}$ and $\min\left\{N_{con}^{(c)}, N_{con}^{(p)}\right\}$. In reality, these numbers are somewhere in between. To calculate what are these numbers in real data, we use values of prediction/classification accuracy reported in actual case/control studies.

For a clinical sample the best PRS prediction accuracy (Area Under the Curve) was reported as AUC=0.75 with sensitivity and specificity $Se^{(c)} = Sp^{(c)} = 0.69$ [1]. The PRS prediction accuracy values in pathologically confirmed sample of cases and controls were published in [Escott Price et al. 2015] and are $Se^{(p)} = Sp^{(p)} = 0.79$, and $NPV^{(p)} = 0.69$. (The latter numbers however might be marginally overestimated, due to 3% overlap of the discovery and test samples used in [Escott Price at al. 2017].) Using these prediction accuracy values, we

construct the confusion matrices (tables A1 and A2 in Appendix 1) in the clinical sample [1] of the total of $N = 4603$ (3049 Alzheimer's disease cases and 1554 controls) individuals, as:

Table 1.
Clinical diagnosis

Table 2. Pathologically confirmed status

Genetic test

	Yes	No
Yes	$a = 2096$	$b = 485$
No	$c = 953$	$d = 1069$
Total	$N_{cas}^{(c)} = 3049$	$N_{con}^{(c)} = 1554$

	Yes	No
Yes	$A = 2285$	$B = 359$
No	$C = 607$	$D = 1352$
	$N_{cas}^{(p)} = 2892$	$N_{con}^{(p)} = 1711$

From these two tables we cannot simply imply that out of 3,049 clinical cases, 2,892 cases will be pathologically confirmed, as some subjects, who are unaffected according to the clinical assessment, may actually have AD. Using sensitivity of 76.6% reported in [Beach et al 2012], we estimate the number of true cases (which were clinically diagnosed as AD and expected also be pathologically confirmed) 3,049*0.766 ≈2,336 (denoted as x in Appendix). Then the number of controls which expected to be pathologically confirmed is $N_{con}^{(c)} - N_{cas}^{(p)} + x$ =1,554 - 2,892 + 2,335 = 998 (denoted as y in the equation (1) in Appendix). Finally, in this sample we obtain mis-assignment rate (MAR) in controls MAR=557/1554=0.36 (see equation (2) in Appendix).

For E3 homozygous subjects in the clinical cohort [Escott Price et a; 2015], the genetic based prediction AUC was lower (AUC=0.65) with sensitivity and specificity $Se^{(c)} = Sp^{(c)} = 0.60$ (N cases=1090 and N controls=947). The values of the genetic prediction accuracy measures in pathologically confirmed sample [3] were $Se^{(p)} = Sp^{(p)} = 0.745$, and $NPV^{(p)} = 0.768$. Clinical AD misdiagnosis rates in non-carriers of the apolipoprotein E4 allele are higher for subjects who are unscreened for E4 alleles. Using 37% as the approximation to AD misdiagnosis rate for E3 homozygous individuals [Monsell et al 2015], gives the misassignment rate in controls of about 29% clinical samples [Escott-Price et al 2015]. That is, about 29% of persons assigned as controls in the clinical series at the sge of these series (late 70s), are in the early stages of disease.

Discussion

It has been reported that Alzheimer's disease misclassification rates range between 14%-37% depending on the exact clinical and neuropathologic criteria used and whether the individuals were screened for APOE E4 alleles [Beach et al 2012, Monsell et al. 2015]. In addition, recent clinical trials show that 20% of all patients (and more than 33% of those who were noncarriers of the apolipoprotein E4 allele) with mild-to-moderate Alzheimer's dementia did not show an elevation in amyloid on positron emission tomography (PET) imaging [Salloway et al. 2014, Doody et al. 2014].

To conduct an actual autopsy based study on unaffected individuals aiming to identify of AD cases among them, is difficult to justify unless it is a part of a large population screening study. Here to use the genetic prediction findings and mathematically to estimate misassignment in controls. Our earlier 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 [Escott Price et al. 2017 and under review], indicating that APOE is an independent risk factor for the disease. Therefore we argue that it is not sufficient just to screen for APOE to classify subjects for example, in AD clinical trials.

Our results show that the misassignment rates in controls in clinical case-control studies is likely to be high (>30%). It would be expected to see increased number of actual controls among E3 homozygous subjects as those individuals do not carry the strongest AD predictor. Indeed, the negative predictive value, or the percentage of correctly predicted controls, in the pathology confirmed sample is higher than in clinical cohort (NPV=0.77 and 0.57 in pathology confirmed and clinical samples, respectively). However, the misdiagnosis rate in of cases in E3 homozygotes is high (37%), which implies reduced but still relatively high rates of missagnments, as compared to unscreened for APOE sample (29% vs 36%, respectively).

These levels of misassignment rates in both cases and controls reduce not only the power of statistical analyses in case/control series but also the PRS prediction accuracy in clinical samples. In biomarker studies of Alzheimer's

disease they suggest that no biomarker will be able to give clean separations between those diagnosed with disease and those designated as controls since considerable proportions of both categories will be misclassified. As CSF and blood biomarkers of disease are assessed in clinical series, this inevitable misclassification, with ~30% of both cases and ~30% of controls being categorised in the wrong group.

References

Beach TG, Monsell SE, Phillips LE, Kukull W. (2012) Accuracy of the Clinical Diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. J Neuropathol Exp Neurol 71(4):266-273

Corneveaux JJ, Myers AJ, Allen AN, et al. Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and

neuropathologically verified individuals. Hum Mol Genet. 2010 Aug 15;19(16):3295-301

Doody RS, Thomas RG, Farlow M, et al; Alzheimer's Disease Cooperative Study Steering Committee; Solanezumab Study Group. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. N Engl J Med. 2014;370(4):311-321.

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.

Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of Alzheimer's Disease in cases without APOE4 or APOE2 alleles. JAMA Network Open (under review)

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

Monsell SE, Kukull WA, Roher AE, Maarouf CL, Serrano G, Beach TG, Caselli RJ, Montine TJ, Reiman EM. (2015) Characterizing Apolipoprotein E ε4 Carriers and Noncarriers With the Clinical Diagnosis of Mild to Moderate Alzheimer Dementia and Minimal β-Amyloid Peptide Plaques JAMA Neurol. 72(10):1124-31

Salloway S, Sperling R, Fox NC, et al; Bapineuzumab 301 and 302 Clinical Trial Investigators. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *N Engl J Med*. 2014;370(4):322-333.

Appendix.

Assume that in a sample of N total subjects, the proportion of clinical cases is known (f). Then the numbers of "clinical cases" and "clinical controls" are $N_{cas}^{(c)} = fN$ and $N_{con}^{(c)} = (1-f)N$, respectively. We further assume that a genetic test, e.g. PRS, divides the subjects into two groups called "predicted clinical cases" and "predicted clinical controls" with sensitivity $Se^{(c)}$ and specificity $Sp^{(c)}$. Then all entries of the "clinical" classification table (Table A1) can explicitly be calculated.

Table A1. Classification table comparing genetic test outcome with clinical diagnosis.

		Clinical diagnosis		
		Yes	No	Total
Genetic test	Yes	$a = NfSe^{(c)}$	$b = N(1-f)\left(1-Sp^{(c)}\right)$	$a+b$
	No	$c = Nf(1-Se^{(c)})$	$d = N(1-f)Sp^{(c)}$	$c+d$
	Total	$a+c = Nf = N_{cas}^{(c)}$	$b+d = N(1-f) = N_{con}^{(c)}$	N

Table A2 is the classification table for pathologically confirmed cases and controls in the same hypothetical sample of N subjects, where A, B, C and D values are the numbers of the true positive, false positive, false negative and true negative predictions by genetic information, respectively. These values are unknown, however, the prediction accuracy estimates which compare pathologically confirmed disease status with genetic prediction, can be obtained from published studies (e.g. for AD, Escott-Price et al (2017)). Let $Se^{(p)}$, $Sp^{(p)}$, $PPV^{(p)}$ and $NPV^{(p)}$ (sensitivity, specificity, positive and negative predictive values, respectively) be known from an external study.

Table A2. Classification table comparing genetic test outcome with true pathologically confirmed status.

		Pathologically confirmed status		
		Yes	No	Total
Genetic test	Yes	A	B	$A+B$
	No	C	D	$C+D$
	Total	$A+C = N_{cas}^{(p)}$	$B+D = N_{con}^{(p)}$	N

Genetic Analysis of Misdiagnosis of Alzheimer's

The sensitivity, specificity and negative predictive values are defined as

$Se^{(p)} = \frac{A}{A+C}$, $Sp^{(p)} = \frac{D}{B+D}$, $NPV^{(p)} = \frac{D}{C+D}$. Together with the expression for the

total number of subjects, $N = A + B + C + D$, the entries of the Table A2 can be
calculated as

$$A = D\frac{\beta}{\gamma}, B = \alpha D, C = \beta D, D = \frac{N}{1+\alpha+\beta+\frac{\beta}{\gamma}},$$

where $\alpha = \frac{1-Sp^{(p)}}{Sp^{(p)}}$, $\beta = \frac{1-NPV^{(p)}}{NPV^{(p)}}$, and $\gamma = \frac{1-Se^{(p)}}{Se^{(p)}}$.

Finally, to identify how many controls are likely to be pre-cases in the clinical
sample and vice versa, we construct Table A3, which compares clinical diagnosis
with pathologically confirmed status. In Table A3, x is the number of subjects
whose clinical diagnosis is correct (i.e. will be pathologically confirmed as having
AD), and y is the number of healthy controls who will die without AD.

**Table A3. Classification table comparing clinical diagnosis with true
pathology status.**

		Pathologically confirmed status		
		Yes	No	Total
Clinical diagnosis	Yes	x	$a + c - x$	$a + c = N_{cas}^{(c)}$
	No	$= \begin{array}{c} b + d - y \\ A + C - x \end{array}$	y	$b + d = N_{con}^{(c)}$
	Total	$A + C = N_{cas}^{(p)}$	$B + D = N_{con}^{(p)}$	N

The numbers of correctly assessed controls are

$$y = N_{con}^{(c)} - N_{cas}^{(p)} + x, \qquad (1)$$

and the mis-assessment rate (MAR) in controls is

$$MAR = \left(N_{cas}^{(p)} - x\right)/N_{con}^{(c)}. \qquad (2)$$

Note for both equations (1) and (2), the number of true positive cases x needs to
be defined.

Since all entries of this table represent the numbers of people and thus are
positive, the range of values for x is between $\max\{0, N_{cas}^{(p)} - N_{con}^{(c)}\}$ and

$\min\left\{N_{cas}^{(c)}, N_{cas}^{(p)}\right\}$, and the range of values for y is between $\max\left\{0, N_{con}^{(c)} - N_{cas}^{(p)}\right\}$ and $\min\left\{N_{con}^{(c)}, N_{con}^{(p)}\right\}$.

When the misdiagnosis rate in cases is at its maximum (i.e. value of $x=0$ or $N_{cas}^{(p)} - N_{con}^{(c)}$, if the number of pathologically confirmed cases is greater than the number of clinically assessed controls), then the miss-assessment rate in controls is also at its maximum: either $y = 0$, i.e. all controls (after pathology check) have initially been incorrectly diagnosed as cases, or $y = N_{con}^{(c)} - N_{cas}^{(p)}$, i.e. *all* pathologically confirmed cases were considered as controls in the clinical sample. The best case scenario is when x is at its maximum, i.e. all clinical diagnoses of cases were correct. Then y is at its maximum too, i.e. all controls in the clinical sample were pathology confirmed as clear of AD, or all subjects confirmed as "clear" were correctly assigned to the control group.

Table A4 demonstrates these two scenarios for a real sample of 4603 subjects (3049 cases and 1554 controls, according to clinical assessment) [8]. The proportion of cases is $f = 0.66$. In this sample the best AUC (Area Under the Curve) was reported as 0.75, the sensitivity and specificity $Se^{(c)} = Sp^{(c)} = 0.69$ [1]. Prediction accuracy estimates which compare pathologically confirmed disease status with genetic prediction are $Se^{(p)} = Sp^{(p)} = 0.79$, and $NPV^{(p)} = 0.69$ [2]. Tables A1 and A2 then look as follows:

		Clinical diagnosis (Table A1)		Pathologically confirmed status (Table A2)	
		Yes	No	Yes	No
Genetic test	Yes	$a = NfSe^{(c)}=2096$	$b =485$	$A =2285$	$B =359$
	No	$c =953$	$d =1069$	$C =607$	$D =1352$
	Total	$N_{cas}^{(c)}=3049$	$N_{con}^{(c)}=1554$	$N_{cas}^{(p)}=2892$	$N_{con}^{(p)}=1711$

From these two tables we cannot simply imply that out of 3049 clinical cases, 2892 cases were pathologically confirmed, as some subjects, which are unaffected according to the clinical assessment, may actually be pathologically confirmed AD cases.

When x (Table A3), is at its minimum, i.e. the misdiagnosis rate in cases is at maximum, then $y = 0$, i.e. all pathologically confirmed controls have been incorrectly clinically diagnosed as cases. In our real example min(x) =1338, which corresponds to the worst case scenario, the highest possible misdiagnosis rates 56% and 100% in cases and controls, respectively (see left section of Table A4).

The best case scenario is when x is at its maximum (right section of Table A4). In our example max(x) =2892. Then the misdiagnosis rate in cases is only 5%, and all subjects, clinically seen as controls, were pathologically confirmed as controls (misdiagnosis rate in controls is 0%).

Table A4. Hypothetical best and worst scenarios of misclassification of clinical and neuropathologic diagnoses of AD.

Pathologically confirmed status

		Worst scenario		Best scenario		Total
		Yes	No	Yes	No	
Clinical diagnosis	Yes	1338	1711	2892	157	$N_{cas}^{(c)}$=3049
	No	1554	0	0	1554	$N_{con}^{(c)}$=1554
	Total	$N_{cas}^{(p)}$=2892	$N_{con}^{(p)}$=1711	$N_{cas}^{(p)}$=2892	$N_{con}^{(p)}$=1711	$N = 4603$