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

Valentina Escott-Price PhD[1,2], Emily Baker[2], Maryam Shoai[3], Ganna Leonenko[1], Amanda J. Myers PhD[4], Matt Huentelman PhD[5] and John Hardy PhD[3]*.

Accepted to Neurobiology of Aging.

1. Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, UK.
2. Dementia Research Institute, Cardiff University, UK
3. Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, Institute of Neurology, London, UK.
4. 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
5. Neurogenomics Division, The Translational Genomics Research Institute (TGen), Phoenix, AZ 85004

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

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 misassigned. For Alzheimer's disease (AD) it is important to know the accuracy of the clinical assignment. 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 a genetic component for the 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. The derived formulae provide a tool to estimate misassignment rates in any sample. This approach can also provide an estimate of the maximal and minimal misassignment rates and therefore could be useful for statistical power estimation at the study design stage. We illustrate this approach in two independent clinical cohorts and estimate misdiagnosis rate up to 36% in controls unscreened for the *APOE* genotype, and up to 29% when E3 homozygous subjects are used as controls in clinical studies.

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 et 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. (2018)]. 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 the 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 the 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 a third of *APOE* E4 non-carriers 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. This is necessitated by the frequently asked question "what proportion of controls are actually early cases", when dealing with GWAS results? 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

Derivation of misassignment rate estimates in a clinical sample.

Misassignment rate was calculated using derived analytical formulae based on sensitivity and specificity. 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 1). The numerical results that we provide in this paper are entirely derived from estimates made in previous publications. The estimates can be entirely populated using the clinical case/control numbers of the study.

Samples used for illustration of misassignment rates estimation

We applied the derived formulae and estimated misassignment rates in two independent clinical cohorts. The first is the Genetic and Environmental Risk in Alzheimer's Disease (GERAD) consortium data [Harold et al. (2009)]. This is the first account where the prediction utility of AD PRS was reported. The best prediction accuracy using PRS was achieved when SNPs were pruned for linkage disequilibrium with parameters r2=0.1 and a window of 1000kb, and the most strongly associated AD SNPs with p-values≤0.5 were included in the individual PRS. The other independent dataset was The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. This is a publically available database (http://www.loni.ucla.edu/ADNI/) which started in 2004 and contains genetic, imaging and biomarker data for about 900 individuals between the age of 55 and 90. Clinical diagnosis and genetic information were available for 770 individuals, who were either already diagnosed with AD (N=47) or had mild cognitive impairment (MCI) (N=459) or healthy controls (N=262) at baseline. We generated PRS for the ADNI participants in the same way as for GERAD data, using IGAP stage 1 [Lambert et al. (2013)] summary statistics to inform AD PRS.

For prediction accuracy estimates in a pathologically confirmed sample, we used sensitivity/specificity/PPV/NPV estimates reported in [Escott-Price et al. (2017)] for a pathologically confirmed sample of 1,011 cases and 583 controls. This series was obtained from 21 National Alzheimer's Coordinating Center (NACC) brain banks and from the Miami Brain Bank as previously described [Corneveaux et al., 2010; Myers et al., 2007; Petyuk et al., 2018; Webster et al., 2009]. Our criteria for inclusion were as follows: self-defined ethnicity of European descent (in an attempt to control for the known allele frequency differences between ethnic groups), neuropathologically confirmed Alzheimer's disease or no neuropathology present, and age of death greater than or equal to 65. Neuropathological diagnosis was defined by board-certified neuropathologists according to the standard NACC protocols [Beekly et al., 2004]. Samples derived from subjects with a clinical history of stroke, cerebrovascular disease, Lewy bodies, or comorbidity with any other known

neurological disease were excluded. Alzheimer's disease or control neuropathology was confirmed by plaque and tangle assessment with 45% of the entire series undergoing Braak staging [Braak and Braak, 1995]. Samples were de-identified before receipt, and the study met human studies institutional review board and HIPPA regulations. This work is declared not human-subjects research and is IRB exempt under regulation 45 CFR 46.

To estimate the misassignment rate in controls, the analytical formulae require us to fix the parameter of AD misdiagnosis rate in cases. Since most cases in clinical case/control samples are collected with clinically "probable AD" or "probable or possible AD" diagnosis, and in the pathology confirmed study [Escott-Price et al. (2017)] the neuropathologic criterion for cases was Braak stage V or VI, we used sensitivity of 76.6% and 87.3% for AD misdiagnosis rates as reported in [Beach et al. (2012)]. In addition, according to [Escott-Price et al. (2018)], among *APOE* E4 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 **c**linically and **p**athology based classifications, respectively). The range for the number of subjects who were clinically and neuropathologically confirmed as having 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 number 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, the number will lie somewhere in this range. To calculate 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 $Se^{(c)} = Sp^{(c)} = 0.69$ [Escott-Price et al. (2015)]. The PRS prediction accuracy values in a pathologically confirmed sample of cases and controls were published in [Escott Price et al. (2017)] as $Se^{(p)} = Sp^{(p)} = 0.79$, and $NPV^{(p)} = 0.69$. (The latter numbers however might be marginally overestimated, due to the 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 [Escott-Price et al. (2015)] of the total of $N = 4{,}603$ (3,049 Alzheimer's disease cases and 1,554 controls) individuals, as:

Table 1.
Clinical diagnosis (GERAD)

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

Table 2. Pathologically confirmed status (derived estimates)

Yes	No
A =2285	B =359
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 the misassignment rate (MAR) in controls MAR=557/1554=0.36 (see equation (2) in Appendix 1).

For E3 homozygous subjects in the clinical cohort [Escott Price et al. (2015)], the genetic based prediction AUC was lower (AUC=0.65) with sensitivity and

specificity $Se^{(c)} = Sp^{(c)} = 0.60$ (N cases=1,090 and N controls=947). The values of the genetic prediction accuracy measures in pathologically confirmed sample [Escott-Price et al. (2017)] 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 age of these series (late 70s), are in the early stages of disease.

In an attempt to replicate our result in an independent sample we used the ADNI data. The ADNI cohort is older than GERAD; the mean age in the GERAD sample was 73.8 [SD=8.6] and 71.4 [SD=11.1], and the mean age in the ADNI sample at the last point of assessment was 78.4 [SD=7.1] and 78.9 [SD=7.6], in cases and controls respectively. Similarly to Tables 1 and 2, Tables 3 and 4 present the results for ADNI data. In this dataset we estimated the PRS for each individual as described in [Escott-Price et al. (2015)] and calculated $Se^{(c)} = Sp^{(c)} = 0.678$, PPV=0.621, NPV=0.731, AUC=0.747). The values of the genetic prediction accuracy measures in pathologically confirmed sample [Escott-Price et al. (2017)] were $Se^{(p)} = Sp^{(p)} = 0.79$, and $NPV^{(p)} = 0.686$. To estimate the misassignment rate in controls with our analytical approach, we used sensitivity value 87.3%, which corresponds to the oldest group of people (83.2 years) with "clinically probable or possible" AD in the [Beach et al. (2012)] paper. Our analytical approach gives the misassignment rate in controls of 44.6%. The R code detailing these analyses is presented in Appendix 2.

Table 3.
Clinical diagnosis (ADNI)

Genetic test	Yes	No
Yes	a =118	b = 72
No	c =56	d =152

Table 4. Pathologically confirmed status (derived estimates)

Yes	No
A = 199	B = 31
C =53	D =116

Total	$N_{cas}^{(c)}$=174	$N_{con}^{(c)}$=224	$N_{cas}^{(p)}$=252	$N_{con}^{(p)}$=147

In these data we have also attempted to directly calculate misassignment rates in controls. There were 262 controls available at baseline assessment. On average within 4.7 years, 15 people have progressed to AD, 47 people have developed mild cognitive impairment (MCI), and 200 individuals did not change their diagnosis. This suggests the current misassignment rate is in between 5.7-21.7%. The mean age of the progressors was 75.2 [4.0], and for those who did not progress the average age was 74.1 [SD=5.7] years at the baseline of assessment. However, since AD is age dependent, it is expected that more controls will progress to AD when they reach age 85+. The incidence rate of AD increases almost exponentially with increasing age until 85 years of age. It is still debated whether the incidence will further increase at more advanced ages or will reach a plateau at a certain age [Qiu et al., 2009]. Since there were only 5 individuals of age 85+ in the ADNI data at the baseline, we were unable to estimate incidence rates directly. Here we used incident rates estimates (~55 persons per 1000-years at age 85+) reported by [Qiu et al., 2009]. Thus we can expect an additional 55% of the sample to develop AD after 10 years, which is slightly above of the analytical estimate (44.6%).

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 non-carriers 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)].

Conducting an actual autopsy based study on unaffected individuals, aiming to identify AD cases among them, is difficult to justify unless it is a part of a large

population screening study. Here we use the genetic prediction findings to mathematically estimate the 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 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.

In this study we derive analytical formulae to estimate misassignment rates in clinical studies. These formulae are based upon sensitivity, specificity, PPV and NPV estimated from clinical and pathologically confirmed studies. However, the PPV and NPV estimates must be adjusted according to the case/control ratio of the clinical study for which misassignment rates are being estimated (unless the ratios are equal), and here we show how to calculate sample prevalence adjusted PPV and NPV. To demonstrate how these equations can be used in practice, we calculate misalignment rates in two independent clinical cohorts. Our headline figures are of course dependent on the quantities reported in previous studies. However, the approach is generalisable to other studies, and the misassignment rates can be easily recalculated.

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 an 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 of cases in E3 homozygotes is high (37%), which implies reduced but still relatively high rates of misassignments, as compared to the sample not screened for *APOE* (29% vs 36%, respectively).

In the ADNI data, there were 262 controls available, of them 15 progressed to AD, 47 developed MCI and 200 did not. This suggests a misassignment rate between 8.0-23.7%, however, as AD is age dependent, it is expected that more

controls will progress to AD when they reach age 85+ (prevalence of AD is 18% and 33% in 70-85 and 85+, respectively). Projecting the latter prevalence to this data, the misassignment rate expected is about 40%, which is similar to our estimates.

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.

Author contributions

VEP, EB and GL carried out the data analysis. VEP and JH designed the study and wrote the original draft. MS carried out quality control analyses of the genetic data. AM, MH and JH were responsible for sample collection and data generation.

Potential Conflict of Interest

JH and VEP are co-grantees of Cytox from Innovate UK (UK Department of Business). MS has previously been employed on this grant.

Acknowledgements

JH's genetic work is supported by an Anonymous Donor and the Dolby Foundation. JH and VEP are partly supported by the UKDRI. 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. Data collection was funded by Kronos. Data and biomaterials were collected from several National Institute on Aging (NIA) and National Alzheimer's Coordinating Center (NACC, grant #U01 AG016976) funded sites. Marcelle Morrison-Bogorad, PhD., Tony Phelps, PhD and Walter Kukull PhD are thanked for helping to co-ordinate this collection.

Additional non-NIA funded sites and full acknowledgments for the sample collection is given in Corneveaux et al. 2010.

References

Beach, T. G., et al. (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.

Beekly, D.L., Ramos, E.M., van Belle, G., Deitrich, W., Clark, A.D., Jacka, M.E., and Kukull, W.A. (2004). The National Alzheimer's Coordinating Center (NACC) Database: an Alzheimer disease database. Alzheimer Dis Assoc Disord 18, 270-277.

Braak, H., and Braak, E. (1995). Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol Aging 16, 271-278; discussion 278-284.

Corneveaux, J. J., 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(16): 3295-3301.

Doody, R. S., et al. (2014). Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. N Engl J Med 370(4): 311-321.

Escott-Price, V., et al. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain 138(12): 3673-3684.

Escott-Price, V., et al. (2017). Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Ann Neurol 82(2): 311-314.

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)

Hebert, L. E., et al. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology 80(19): 1778-1783.

Monsell, S. E., et al. (2015). Characterizing Apolipoprotein E epsilon4 Carriers and Noncarriers With the Clinical Diagnosis of Mild to Moderate Alzheimer Dementia and Minimal beta-Amyloid Peptide Plaques. JAMA Neurol 72(10): 1124-1131.

Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., et al. (2007). A survey of genetic human cortical gene expression. Nat Genet 39, 1494-1499.

Petyuk, V.A., Chang, R., Ramirez-Restrepo, M., Beckmann, N.D., Henrion, M.Y.R., Piehowski, P.D., Zhu, K., Wang, S., Clarke, J., Huentelman, M.J.*, et al.* (2018). The human brainome: network analysis identifies HSPA2 as a novel Alzheimer's disease target. Brain 141(9):2721-2739

Salloway, S., et al. (2014). Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. N Engl J Med 370(4): 322-333.

Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., et al. (2009). Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet 84, 445-458.

Yesavage, J. A., et al. (2002). Modelling the prevalence and incidence of Alzheimer's disease and mild cognitive impairment. Journal of Psychiatric Research 36(5): 281-286.

Appendix 1.

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.

Genetic test \ Clinical diagnosis	Yes	No	Total
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 number of 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.

Note that in this calculation we use PPV and NPV values derived directly from the clinical samples, and these values therefore are dependent on the case/control ratio. For example for Table A1, $PPV = \frac{a}{a+b}$, therefore $PPV = \frac{fSe^{(c)}}{fSe^{(c)} + (1-f)\left(1-Sp^{(c)}\right)}$, where f is the proportion of clinical cases in the sample (similar, $NPV = \frac{(1-f)Sp^{(c)}}{f(1-Se^{(c)}) + (1-f)Sp^{(c)}}$). If the known PPV and NPV values are prevalence-adjusted ($Prev$), then before others can use this approach to make

similar predictions from their own clinical studies, the PPV and NPV need to be adjusted accounting for the case/control ratio for their particular clinical study. The prevalence-adjusted $\overline{PPV}$ and $\overline{NPV}$ values are derived using *Bayes*' theorem: $\overline{PPV} = \frac{Se^{(c)}Prev}{Se^{(c)}Prev+(1-Sp^{(c)})(1-Prev)}$ and $\overline{NPV} = \frac{Sp^{(c)}(1-Prev)}{(1-Se^{(c)})Prev+Sp^{(c)}(1-Prev)}$. Then PPV and NPV can be expressed in terms of sensitivity, specificity, prevalence and the propotrtion of cases in the sample ratio as $PPV = \frac{\overline{PPV}f(1-Prev)}{(1-f)Prev-\overline{PPV}(Prev-f)}$ and $NPV = \frac{\overline{NPV}(1-f)Prev}{f(1-Prev)-\overline{NPV}(f-Prev)}$.

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

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

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.

Clinical diagnosis	Pathologically confirmed status: Yes	No	Total
Yes	x	$a + c - x$	$a + c = N_{cas}^{(c)}$
No	$b + d - y = A + C - x$	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 misassignment 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 number of people and thus are positive, the range of values for x is between $\max\left\{0, N_{cas}^{(p)} - N_{con}^{(c)}\right\}$ 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 misassignment rate in controls is also at its maximum: either $y = 0$, i.e. all controls (after a 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 4,603 subjects (3,049 cases and 1,554 controls, according to clinical assessment) [Escott-Price et al. (2015)]. 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 $Se^{(c)} = Sp^{(c)} = 0.69$ [Escott-Price et al. (2015)]. Prediction accuracy estimates which compare pathologically confirmed disease status with genetic prediction are $Se^{(p)} = Sp^{(p)} = 0.79$, and $NPV^{(p)} = 0.69$ [Escott-Price et al. (2017)]. Tables A1 and A2 then look as follows:

Clinical diagnosis (Table A1)

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

Pathologically confirmed status (Table A2)

Yes	No
A =2285	B =359
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 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 $\min(x)$ =1,338, 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 $\max(x)$ =2,892. 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

Appendix 2. Illustration of misassignment rates estimation in ADNI data with R script.

```
#TABLE A1
Ncas<-174
Ncon<-224
N<-Ncas+Ncon

mat<-matrix(c(118,56,72,152), ncol=2)
Sens<-mat[1,1]/sum(mat[,1]); Sens
Spec<-mat[2,2]/sum(mat[,2]); Spec
PPV<-mat[1,1]/sum(mat[1,]); PPV
NPV<-mat[2,2]/sum(mat[2,]); NPV
sum(mat)

alpha<-(1-Spec)/Spec
beta <-(1-NPV )/NPV
gamma<-(1-Sens)/Sens

d<-N/(1+alpha+beta+beta/gamma)
a<-d*beta/gamma
b<-d*alpha
c<-d*beta
clinical<-matrix(c(a,c,b,d), ncol=2)
clinical<-round(clinical)

#TABLE A2
Sens<-0.79
Spec<-0.79
PPV<-0.867
NPV<-0.686

alpha<-(1-Spec)/Spec
```

```
beta <-(1-NPV)/NPV
gamma<-(1-Sens)/Sens

D<-N/(1+alpha+beta+beta/gamma)
A<-D*beta/gamma
B<-D*alpha
C<-D*beta
path.conf<-matrix(c(A,C,B,D), ncol=2)
path.conf<-round(path.conf)

clinical
path.conf


#TABLE A3
f<-0.873 #sensitivity of 87.3% as reported in [Beach et al 2012]
t3<-matrix(0, ncol=2, nrow=2)
x<-round(f*(a+c))
t3[1,1]<-x
t3[1,2]<-a+c-x
t3[2,1]<-A+C-x
t3[2,2]<-x-A-C+b+d
t3<-abs(round(t3)); t3

#cases misdiag rate:
t3[1,2]/(t3[1,1]+t3[1,2])
#controls misdiag rate:
t3[2,1]/(t3[2,1]+t3[2,2])
```