A pathway based approach for analyzing gene expression for Alzheimer's Disease diagnosis

Nicola Voyle [*1,2], Aoife Keohane[1], Stephen Newhouse[1,5], Katie Lunnon[3], Caroline Johnston[1,5], Hilkka Soininen[4], Iwona Kloszewska[6], Patrizia Mecocci[7], Magda Tsolaki[8], Bruno Vellas[9], Simon Lovestone[1,10], Angela Hodges[1], Steven Kiddle [†1,2], Richard JB Dobson [‡ †1,5], and the AddNeuroMed research group[11]

[1]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

[2]MRC Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UK

[3]University of Exeter Medical School, Exeter, UK

[4]Department of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland

[5]NIHR Biomedical Research Centre for Mental Health and Biomedical Research Unit for Dementia at South London and Maudsley NHS Foundation, London, UK

[6]Medical University of Lodz, Lodz, Poland

[7]Institute of Gerontology and Geriatrics, University of Perugia, Perugia, Italy

[8]3rd Department of Neurology, Aristotle University, Thessaloniki, Greece

[9]INSERM University of Toulouse, Toulouse, France

[10]Department of Pyschiatry, Oxford University, Oxford, UK

[11]www.innomed-addneuromed.com

[*]Corresponding author: nicola.voyle@kcl.ac.uk

[†]Joint last author

[‡]Corresponding author: richard.j.dobson@kcl.ac.uk

Abstract

Background Recent studies indicate that gene expression levels in blood may be able to differentiate subjects with Alzheimer's Disease (AD) from normal elderly controls and mild cognitively impaired (MCI) subjects. However, there is minimal replicability at the single marker level. A pathway-based interpretation of gene expression may prove more robust.

Objectives This study aimed to investigate whether a case/control classification model built on pathway level data was more robust than a gene level model and may consequently perform better in test data. The study used two batches of gene expression data from the AddNeuroMed (ANM) and Dementia Case Registry (DCR) cohorts.

Methods Our study used Illumina Human HT-12 Expression BeadChips to collect gene expression from blood samples. Random forest modeling with recursive feature elimination was used to predict case/control status. Age and APOE $\epsilon 4$ status were used as covariates for all analysis.

Results Gene and pathway level models performed similarly to each other and to a model based on demographic information only.

Conclusions Although the novel approach used here has contradicted the reviews suggesting that we may see greater concordance and hence predictive ability at the pathway level we have benchmarked pathways against genes in datasets that had been extensively harmonised. Further work should focus on the use of pathway level scores that incorporate pathway topology and the use of an endophenotype based approach.

Keywords Alzheimer's Disease, Gene Expression, Blood, Pathways

1 Introduction

The most common form of dementia is AD. It is predicted that by 2050, 1 in every 85 people will be living with the disease [1]. No disease modifying treatments are available for AD and existing treatments only provide short term symptomatic relief in a subset of patients [2]. Additionally, in the early stages (between 2 and 15 years prior to the development of clinical symptoms) the disease is difficult to diagnose. Villemagne et al. and Jack et al. hypothesise that characteristic AD pathology (the presence of amyloid-β (Aβ) plaques and hyperphosphorylated tau tangles in the brain) begins to develop up to 20 years prior to clinical diagnosis [3, 4]. This extended prodromal stage is an important window in which to target treatments that may be able to alter the course of the disease; provided people could be sensitively and accurately diagnosed. Aβ, tau and phosphorylated-tau levels are indicative of AD pathology in this prodromal period and can be measured in cerebrospinal fluid (CSF) and by positron emission tomography (PET) imaging [5]. The procedures involved in attaining these measurements can be invasive or expensive and require specialised administration, equipment and expertise. The development of a less invasive, potentially cheaper technique, such as a blood test, would offer significant advantages [6].

Recent studies indicate that gene expression levels in blood may be able to differentiate AD subjects from normal elderly controls and MCI subjects with prodromal disease [7, 8, 9, 10]. Han, Wang and Zend et al. provide an overview of studies of gene expression associated with AD-related phenotypes [11]. They state that the blood transcriptome is vital in the disease mechanism of AD and should therefore be investigated further in independent studies of a large sample size. A more general summary of gene expression data in neurodegenerative diseases is given by Cooper-Knock et al. [12]. This review emphasises the dysregulation in neuroinflammation and intracellular signalling pathways including calcium signalling in AD. The commonality between these reviews is that they both highlight minimal replicability at the single marker level. Furthermore, Han et al. report a greater concordance between differentially expressed genes at the pathway level. A pathway-based interpretation of gene expression may therefore prove more robust across different sample populations. Such an approach may also reduce noise and dimensionality.

Although previous gene expression studies in AD have retrospectively identified pathways altered in disease [9] this is the first study to use pathway scores for each individual to build predictive models across the population. This study used Gene Set Variation Analysis (GSVA) to estimate pathway variablility across samples in the population by calculating sample-wise pathway scores [13]. GSVA outperformed other single sample enrichment methods such as ZSCORE, Pathway Level Analysis of Gene Expression (PLAGE) and Single Sample Gene Set Enrichment Analysis (SSGSEA) in simulated data, when differentiating between

two types of leukemia and in survival analysis of ovarian carcinoma [13]. GSVA scores have been used in univariate t-testing for oncology survival analysis and to calculate scores for specific pathways of interest as well as in unsupervised clustering methods for disease sub-typing [14, 15, 16]. We combine, for the first time, GSVA scoring with a supervised machine learning approach to build an AD classifier.

This study used blood expression data from subjects participating in the ANM and DCR studies to develop models of clinical diagnosis. The performance of the models is compared with those generated using pathway level measures of expression.

2 Materials and Methods

2.1 Cohort

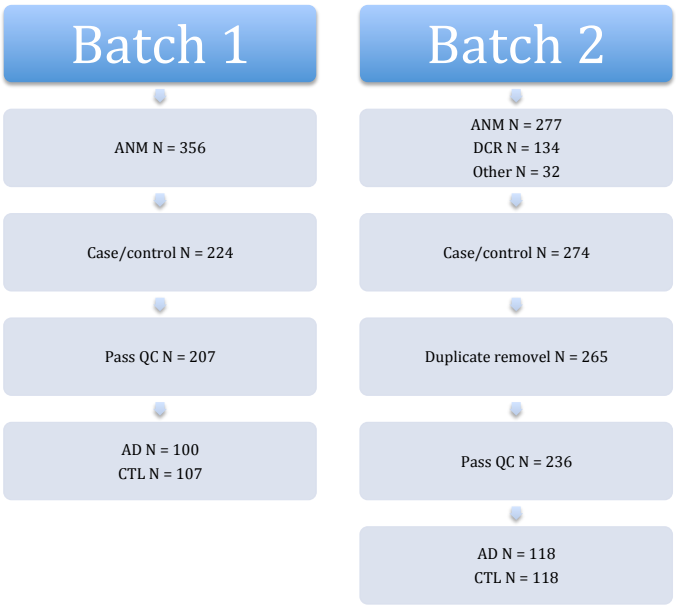
ANM is a European multi-center study aiming to develop biomarkers for AD [17]. Subjects with an AD diagnosis as well as those with MCI and healthy controls were recruited from centres based in Kuopio, Lodz, London, Perugia, Thessaloniki and Toulouse. Details of study design and enrolment are provided by Lunnon et al. [10]. Subjects for the DCR were recruited from the Maudsley and Kings Healthcare Partners which incorporates the Alzheimers Research UK (ARUK) cohort [18] from whom gene expression data has not previously been reported.

The present study used data from 748 subjects: 614 subjects from ANM and 134 subjects from DCR.

2.2 Gene expression

Whole blood samples (2.5ml) were collected after 2 hours of fasting into Paxgene Blood RNA tubes (BD) and extracted as in Lunnon et al. [9]. Illumina Human HT-12 Expression BeadChips were used to analyse the whole transcriptome according to the manufacturers protocol. The gene expression analysis was run in two batches at two different sites. Batch 1 contained samples from 356 ANM subjects run on version 3 of the BeadChip, as previously described [9, 10]. Batch 2 contained samples from 411 subjects: 134 from DCR and 277 from ANM run on version 4 of the BeadChip. Samples from 19 subjects were included in both batches. See Figure 1 for an overview of sample numbers.

Figure 1: Overview of sample numbers in batch 1 and 2 gene expression



2.3 Statistical Analysis

2.3.1 Data Pre-processing

The data pre-processing performed in this study is different to that used for the original analysis by Lunnon et al. [9, 10]. Raw gene expression data was exported from Illumina's Genome studio and processed in R (version 3.1.1 [19] using the lumi package [20] and custom in-house pre-processing scripts (GitHub http://bit.ly/1vjyKNo). Briefly, raw expression data was subject to a model based background correction for bead array [21]. This used negative bead expression levels to correct for background noise. The data was then log base 2 transformed and robust spline normalized in lumi [20]. Outlying samples were iteratively identified using fundamental network concepts and removed following the methods described by Oldham et al. [22]. To reduce any batch effects we adjusted for technical categorical variables using ComBat [23]. Continuous technical artefacts were accounted for by taking the first principal component across housekeeping and undetected probes and regressing this against technical variables. Variables significantly associated with the first principal component were then regressed against expression for each probe, and the mean adjusted residuals taken forward for all further analyses. Finally, the data was reduced to a subset of probes that could be reliably detected in 80% of samples in at least one diagnostic group. Finally, subjects were excluded where there were discrepancies between the recorded sex and sex determined by the XIST (ILMN_1764573), USP9Y (ILMN_2056795) and EIF1AY (ILMN_1755537 and ILMN_2228976) X- and Y- linked genes.

Demographic data for the ANM and DCR subjects was extracted using CohortExplorer [24].

2.3.2 Gene Set Variation Analysis (GSVA)

Gene level expression data were condensed to sample wise, pathway level scores using Gene Set Variation Analysis (GSVA) [13]. GSVA groups genes into pathways defined by the Broad Institute Collection of Curated Pathways [25] and outputs a score, per sample, for each of these sets. If a gene has multiple probes, all probes are included [c1]. We restricted GSVA to only include pathways with between 10 and 500 genes. The scores range from 1 to -1 indicating the extent to which a pathway is up or down regulated, respectively. The generation of GSVA scores is detailed in supplementary methods section 7.

[c1] *Nicola*: I am confirming this with the package author.

2.3.3 Data analysis

Clinical diagnosis (AD versus non-demented elderly control) classification models were built using batch 1 gene expression data. Variable selection was performed using recursive feature elimination (RFE) and the creation of a tolerance set using the 'pickSizeTolerance' function in R. This function finds a smaller set of variables while maintaining model accuracy [26]. Three Random Forest (RF) models were built, the first of which was a model based on demographic data alone (*demographic model*) [27]. The demographic variables included were those that were significant in the batch 1 population: sample collection site, age, years in full time education and *APOE* status (defined as the presence of any number of $\epsilon 4$ alleles)(Table 1). Two further models were built based on these demographic variables and gene level data (*gene model*) or GSVA scores (*pathway model*). The purpose of the *demographic model* is to provide a comparator for the gene and pathway models. If models that include blood expression information (as well as demographics) are no more informative than demographic variables alone there is no benefit in including this information. All model building was performed in the statistical software R (Version 3.1.1) using the Caret package [26].

Each model was used to predict the diagnostic status of subjects in batch 2. Model statistics including accuracy, sensitivity and specificity were generated and compared between the *demographic model*, *gene model* and *pathway model*. Receiver Operator Curve (ROC) analysis was also performed in batch 2 data using R packages ROCR and pROC [28, 29].

Full details of model building are provided in supplementary methods section 8.

Additionally, variable importance (determined as the change in Gini index) was examined in the *pathway model*. Permutation tests were run to assess the size of the observed variable importance scores relative to those expected under the null hypothesis of no association. To achieve this, 1000 permutations of the demographic variables (including diagnosis) were performed and each time a RF model built. The importance measures of each pathway were then compared to that of the original model to generate an empirical p-value. A p-value of less than 0.05 was considered significant. The validity of the pathways selected in the *pathway model* was also investigated; a random set of pathways (of the same size as the final *pathway model*) were selected, and used to build a RF model. This process was repeated 1000 times and the accuracies across all models compared to create an empirical p-value.

3 Results

3.1 Data pre-processing

As a result of pre-processing 12 samples in batch 1 and 49 samples in batch 2 failed quality control (QC) and were removed. The majority of these samples failed QC as they were identified as outliers. Additionally, some samples were removed because the sex of the individual recorded in the clinical database did not match the biological sample (2 samples in batch 1 and 7 in batch 2).

Samples from 19 subjects were present in both batch 1 and batch 2. Samples from 14 of these individuals passed QC in both batches; only data from batch 1 was used and the other was discarded. Correlation between the two batches was at least 0.9 for all individuals (Supplementary Figure S3). Batch 2 gene expression data contains subjects from the DCR whereas batch 1 does not. This study used the same protocols, staff and facilities as the London sample collection site within ANM.

Principal components analysis (PCA) was performed across the batch 2 gene expression data from DCR and ANM subjects from London. The first three principal components (accounting for >40% of variation) were linearly regressed against the study the individual was enrolled in (DCR or London ANM) and found to be non-significant. Therefore, it was deemed appropriate to group DCR subjects with London ANM, allowing the model trained in batch 1 data to be simply applied to batch 2 data.

After data processing only subjects with either an AD diagnosis at all visits or control status at all visits were analysed further: 207 subjects in batch 1 and 236 in batch 2.

Only gene probes that mapped between the version 3 and version 4 chips used to generate batches 1 and 2, respectively, were used for analysis (5212 probes). The Broad Institute Collection of Curated Pathways matched these probes to 834 pathways [25].

3.2 Cohort demographics

Table 1: Population demographics

	AD	Control	P-value
Batch 1			
N	100	107	
Sex (% female)	69	58.9	0.149
APOE status (% of *APOE* ϵ4 positive)	57	32.7	< 0.001
APOE ϵ4 load (% with loads 0; 1; 2)	43; 40; 17	67.3; 29; 3.7	< 0.001
Median age [IQR] (years)	76 [10]	73 [9]	< 0.001
Median MMSE score [IQR]	22 [7.25]	29 [1]	< 0.001
Median years in fulltime education [IQR]	7 [5]	11 [8]	< 0.001
Sample collection site (% from KPO; LDZ; LND; PRG; THS; TLS)	32; 15; 7; 26; 12; 8	21.5; 13.1; 21.5; 21.5; 6.5; 15.9	0.011
Batch 2			
N	118	118	
Sex (% female)	63.6	61.9	0.893
APOE status (% of *APOE* ϵ4 positive)	52.5	24.6	< 0.001
APOE ϵ4 load (% with loads 0; 1; 2)	47.5; 39.8; 12.7	75.4; 20.3; 4.2	< 0.001
Median age [IQR] (years)	78 [9]	74 [8]	0.001
Median MMSE score [IQR]	21 [8]	29 [2]	< 0.001
Median years in fulltime education [IQR]	9 [7]	11 [5]	0.001
Sample collection site (% from KPO; LDZ; LND; PRG; THS; TLS)	10.2; 18.6; 35.6; 19.5; 10.2; 5.9	17.8; 7.6; 51.7; 17.8; 3.4; 1.7	0.002

Individuals were positive for *APOE* ϵ4 if at least one *APOE* ϵ4 allele was seen in their genotype.
APOE ϵ4 load was the number of alleles seen in a subjects genotype.
Kruskal Wallis Chi-Squared was used to test between cases and controls for continuous data.
Fishers exact was used to test between cases and controls for categorical data.
KPO = Kuopio; LDZ = Lodz; LND = London; PRG = Perugia; THS = Thessaloniki; TLS = Toulouse

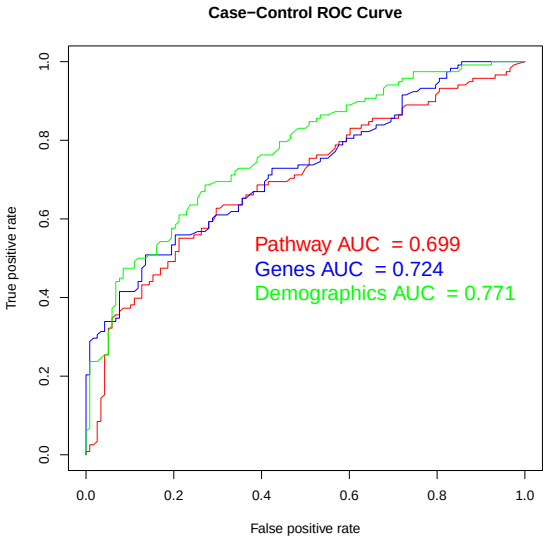
3.3 Data analysis

Table 2: Random forest model results in independent test data

Model	Accuracy [95% CI]	Sensitivity	Specificity	AUC ROC
Demographic model	0.686 [0.623; 0.745]	0.534	0.839	0.771
Demographic model (no sample collection site)	0.674 [0.610; 0.733]	0.678	0.669	0.761
Pathway model	0.657 [0.592; 0.717]	0.610	0.703	0.699
Gene model	0.657 [0.592; 0.717]	0.568	0.746	0.724

CI = Confidence interval; AUC ROC = Area under the receiver operating curve.

Figure 2: ROC curves for Random Forest models in independent test data



3.3.1 Demographic model

The following demographic variables were associated with case/control status in our cohorts (Table 1): age, sex, *APOE* status, years in full time education and sample collection site. These variables were therefore used in multivariate modelling using RFE. The optimal cross-validated accuracy was found when including all variables; calculation of a tolerance set excluded the variable representing the Lodz sample collection site. Variable importance scores showed age as the most important covariate followed by years in full time education and then *APOE* status and sample collection site.

In batch 2 test data the model achieved an accuracy of 0.69, sensitivity of 0.53 and specificity of 0.84. The area under the ROC curve was 0.77 (See Table 2 and Figure 2).

Additionally, a model that did not contain the sample collection site was built. The aim was to create a model based on demographics that would be available to clinicians. This model had a slightly decreased accuracy in comparison to the *demographic model* but outperformed the *pathway model* and *gene model* in accuracy, sensitivity and area under the ROC curve at 0.67, 0.68 and 0.76 respectively. Interestingly, the specificity of the model was lower than all others at 0.67 (Table 2.

3.3.2 Gene model

The top 5% of variables from the bootstrapped variable importance calculations (261 variables) were carried forward to the RFE model building. The optimal cross-validated accuracy from RFE in the *gene model* was found for all of the 261 variables; calculation of a tolerance set reduced this set to only 13, excluding all demographic variables. For a list of genes see Table 3.

In batch 2 test data, the *gene model* accuracy was lower than that of the *demographic model* and equal to the *pathway model*. The sensitivity, specificity and area under the ROC curve of the *gene model* lay between the *demographic* and *pathway* models at 0.59, 0.75 and 0.72 respectively. (See Table 2 and Figure 2). Note that the *pathway model* showed higher sensitivity while specificity and AUC ROC were higher in the *demographic model*.

Table 3: Genes in *gene model* with variable importance scores

Gene (Illumina ID)	Variable importance	Gene symbol	Entrez ID	Gene name
ILMN_2189936	11.9	RPL36AL	6166	Ribosomal protein L36a-like
ILMN_2189933	10.8	RPL36AL	6166	Ribosomal protein L36a-like
ILMN_2097421	10.5	MRPL51	51258	Mitochondrial ribosomal protein L51
ILMN_2237746	10.4	ING3	54556	Inhibitor of growth family, member 3
ILMN_1695645	9.2	CETN2	1069	Centrin, EF-hand protein, 2
ILMN_1784286	7.9	NDUFA1	4694	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex
ILMN_1652073	7.0	LOC653658	653658	Ribosomal protein S23 pseudogene 8
ILMN_1716053	7.0	AK2	204	Adenylate kinase 2
ILMN_1732328	6.5	LOC646200	646200	
ILMN_1776104	5.9	NDUFS5	4725	NADH dehydrogenase (ubiquinone) Fe-S protein 5
ILMN_1753892	5.8	LOC654121	654121	
ILMN_1745343	5.4	ZMAT2	153527	Zinc finger, matrin-type 2
ILMN_2048326	4.7	RPS27A	6233	Ribosomal protein S27a

3.3.3 Pathway model

The top 5% of variables from the bootstrapped variable importance calculations (42 variables) were carried forward to the RFE model building. The optimal cross-validated accuracy from RFE in the *pathway model* was found for all of the 42 variables; calculation of a tolerance set reduced this set to only 6 pathways (Table 4), excluding all demographic variables.

Permutation tests of variable importance were performed to assess the size of effect relative to that observed under the null hypothesis of no association. Of the 6 pathways, 2 achieved nominal significance with a p-value < 0.05 and are indicated with a * in Table 4. Additionally, we compared the model accuracy of 1000 models comprising 6 random pathways. This yielded a p-value of 0.007 indicating that, statistically, the final model performs significantly better than a model of random pathways.

In batch 2 test data the model accuracy was lower than that of the *demographic model* at 0.66 however, the sensitivity was higher at 0.61. Both specificity and area under the ROC curve were lower than the *demographic model* at 0.70. (See Table 2 and Figure 2).

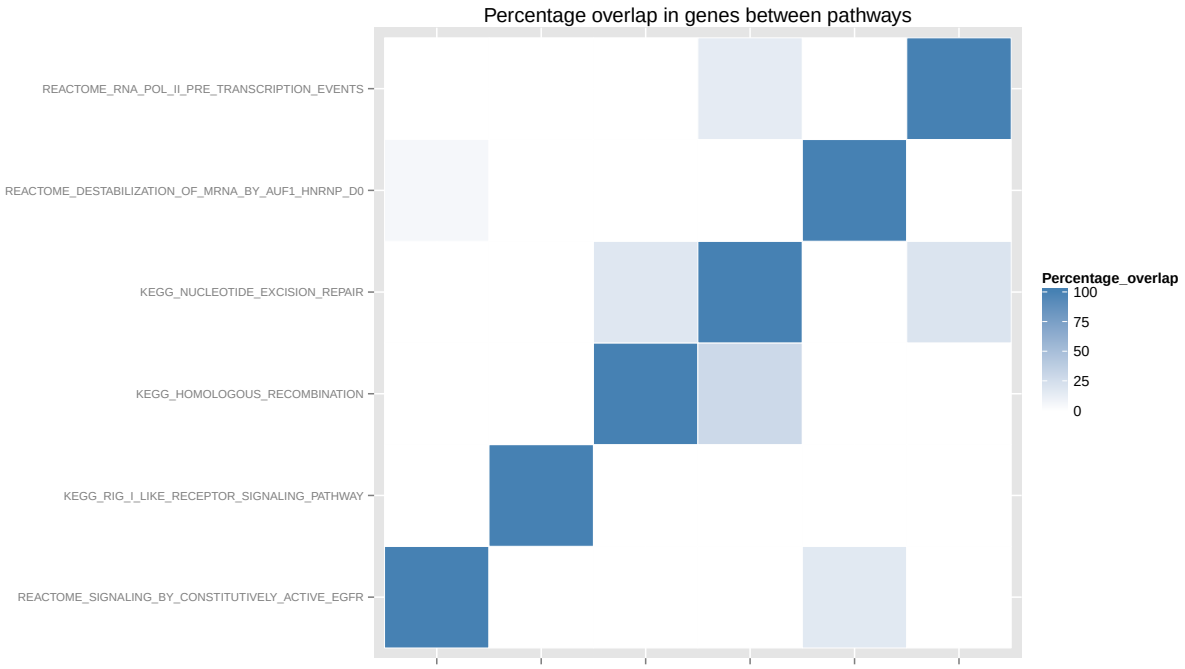
There is minimal overlap in genes between the different pathways included in the final *pathway model*. This is illustrated by the sparse percentage overlap map shown in Figure 3 and supports the idea that each pathway is contributing an independent signal to the model. Of the 13 genes included in the *gene model* only two of them (ILMN_1695645 and ILMN_2048326) appear in any of the pathways in the *pathway model*.

Table 4: Pathways in *pathway model* with variable importance scores

Pathway	**Number of genes in pathway**	**Variable importance**
REACTOME SIGNALING BY CONSTITUTIVELY ACTIVE EGFR	18	21.5 *
KEGG HOMOLOGOUS RECOMBINATION	28	19.9 *
KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY	71	19.3
KEGG NUCLEOTIDE EXCISION REPAIR	44	15.6
REACTOME DESTABILIZATION OF MRNA BY AUF1 HNRNP D0	53	13.7
REACTOME RNA POL II PRE TRANSCRIPTION EVENTS	61	12.9

* = Nominally significant in permutation testing ($p < 0.05$)

Figure 3: Percentage overlap of genes belonging to pathways selected for the Random Forest *pathway model*



3.3.4 Misclassification

We discovered that 22% of controls used in the training data had reported memory complaints deemed not serious enough to reflect a change in diagnosis. By studying misclassification rates split by AD subjects, control subjects and control subjects with memory complaints we see that the most well classified group in both the *gene model* and *pathway model* is those subjects with memory complaints (See Supplementary figure S1. We also demonstrated that time since disease onset is not related to misclassification of AD subjects and controls subjects with memory complaints in the test data (Supplementary figure S2).

4 Discussion

In this study we investigated whether AD cases could be differentiated from control subjects using gene expression data analysed at the pathway level. We were particularly interested in confirming whether pathway level information created a more robust predictor of case/control status than expression data at the gene level as recent reviews have suggested [11]. Our results, using subjects from the ANM and DCR cohorts, show similar model performance in a pathway model compared to a gene and demographic only model. In this study, we do not find improved prediction using pathway level information. However, the robustness of the pathway based approach should be tested in other gene expression data from different populations and platforms.

We expect that the main benefit of the pathway based approach will be robustness when testing models across data generated with different platforms. However, it is useful to initially benchmark the pathway based approach against traditional gene models in data generated on similar platforms. We were able to do this, albeit in different versions of the same platform where we had to limit to probes on both versions, and found that the pathway models achieved comparable performance.

The six pathways included in the final *pathway model* focused around DNA repair, immune response to viral pathogens and ubiquitination. These are pathways similar to those identified by Lunnon et al. who studied overall pathway differences using an identical raw dataset that was processed differently [9]. As we would expect, 12 out of 13 of the genes in the final *gene model* were present in the genes used for modelling by Lunnon et al. The data had been processed slightly differently emphasising that these signals are robust to alterations in processing and modelling methods.

RF models are commonly used in biomarker studies [9, 30, 31]. However, it has been shown that they

exhibit variable selection bias being more likely to select continuous variables or those with many categories [32]. Additionally, the presence of correlated predictors (as is common in gene expression studies) can add further bias [33]. Strobl et al. aimed to address these issues with an ensemble-learning algorithm based on conditional inference trees; Conditional RF (CRF) models [34, 35]. We attempted to use this methodology in the present study. We hypothesised that the creation of an unbiased predictor may highlight different pathways and genes to those previously discovered, potentially allowing greater predictive ability. However, the process of creating a CRF model was computationally expensive even when using high performance computing resources. Model building considering the 834 pathways and 5,212 genes was consequently infeasible. Work to improve the efficiency of this method would be computationally beneficial and would allow the use of alternative variable importance measures. Measures such as mean decrease in accuracy and conditional mean decrease in accuracy would be an improvement over biased variable importance measures such as the Gini index, which was used in this study.

This study used the Broad institute collection of curated pathways to generate the *pathway model* and excluded less well curated gene sets. This method was chosen due to ease of application through the GSVA R package. It may be beneficial, although potentialy computationally costly, to create pathway level scores that also reflect pathway topology and thus add further detail to the model. Such methods have been created by Pyatnitskiy et al. [16] building on the work of others [36, 37]. The method detailed by Pyatnitskiy et al. does not depend on predefined gene sets as used in this analysis. However, it is also unable to control the number of genes in a pathway; a potential benefit of using GSVA.

The creation of a demographic model that excluded sample collection site led to a drop in accuracy. Although RNA extraction and analysis were performed at one site the blood collection may vary by location. We aimed to correct for batch effects occurring in extraction and analysis in the pre-processing. This highlights that although sample collection sites within multi-centre studies are following the same protocols major technical differences can still arise and remain after QC steps including batch correction. As much as possible, these differences should be quantified during extraction. Standardization for future biomarker development will aid this. It is possible that the sample collection site effect we see is driven by genetic differences between sites for some genes (expression quantitative trait loci). For a biomarker to have clinical utility it should be robust to such differences. However, in early exploratory work we are more likely to find results of interest if technical data artefacts are not creating a barrier.

The models created in this study all achieved an accuracy of approximately 70% with the *pathway model* having test sensitivity and specificity results of greater than 60%. The *pathway model* and *gene model* did

not outperform a model of demographics alone. Although the novel approach used here has contradicted the reviews suggesting that we may see greater concordance and hence predictive ability at the pathway level we have benchmarked pathways against genes in datasets that had been extensively harmonised. It is reassuring to see that pathways perform similarly to genes and further work is now needed to see if pathway concordance is more easily detected using other methodological approaches and in data generated by independent groups and platforms. Furthermore, we found that the heterogeneity of control subjects may be leading to reduced predictive accuracy and suggest that the use of an endophenotype may be beneficial in future work.

5 Conclusions

We have used subjects from the ANM and DCR studies to investigate case/control classification using gene and pathway level expression data. We hypothesised that a model built on pathway level data may be more robust than a gene level model and consequently perform better in test data. However, the models performed similarly to each other and to a model based on demographic information only. Further work should focus on the use of pathway level scores that incorporate pathway topology and the use of an endophenotype based approach.

6 Acknowledgements including sources of support

This work was supported by the Alzheimer's Society, InnoMed (Innovative Medicines in Europe), an integrated project funded by the European Union of the Sixth Framework program priority (FP6-2004-LIFESCIHEALTH-5); Alzheimers Research Trust UK; the John and Lucille van Geest Foundation (AH); and the NIHR Biomedical Research Centre for Mental Health and Biomedical Research Unit for Dementia at the South London, Maudsley NHS Foundation Trust and Kings College London, and a joint infrastructure grant from Guy's and St Thomas' Charity and the Maudsley Charity. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement number 115372, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. Kuopio University Hospital (HS) and funding from UEF- BRAIN (HS). Steven Kiddle is supported by an MRC Career Development Award in Biostatistics (MR/L011859/1).

References

[1] Brookmeyer, R., Johnson, E., K, Z.-G. and Arrighi, H. (2007) Forecasting the global burden of Alzheimer' disease. *Alzheimer's & Dementia* **3**, 186 – 191.

[2] Corbett, A. and Ballard, C. (2012) New and emerging treatments for Alzheimer's disease. *Expert Opinion on Emerging Drugs* **17**, 147–156.

[3] Villemagne, V., Pike, K., Chételat, G., Ellis, K., Mulligan, R., Bourgeat, P., Ackermann, U., Jones, G., Szoeke, C., Salvado, O., Martins, R., O'Keefe, G., Mathis, C., Klunk, W., Ames, D., Masters, C. and Rowe, C. (2011) Longitudinal assessment of Aβ and cognition in aging and Alzheimer disease. *Annals of neurology* **69**, 181–192.

[4] Jack, C., Knopman, D., Jagust, W., Petersen, R., Weiner, M., Aisen, P., Shaw, L., Vemuri, P., Wiste, H., Weigand, S., Lesnick, T., Pankratz, V., Donohue, M. and Trojanowski, J. (2013) Personal View-Tracking pathophysiological processes in Alzheimer's disease:an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology* **12**, 207–216.

[5] Cedazo-Minguez, A. and Winblad, B. (2010) Biomarkers for Alzheimer's disease and other forms of dementia: Clinical needs, limitations and future aspects. *Experimental Gerontology* **45**, 5–14.

[6] Bazenet, C. and Lovestone, S. (2012) Plasma biomarkers for Alzheimer's disease: much needed but tough to find. *Biomarkers in Medicine* **6**, 441–454.

[7] Booij, B., Lindahl, T., Wetterberg, P., Skaane, N., Sæbø, S., Feten, G., Rye, P., Kristiansen, L., Hagen, N., Jensen, M., Bårdsen, K., Winblad, B., Sharma, P. and Lönneborg, A. (2011) A Gene Expression Pattern in Blood for the Early Detection of Alzheimer's Disease. *Journal of Alzheimer's Disease : JAD* **23**, 101–119.

[8] Rye, P., Booij, B., Grave, G., Lindahl, T., Kristiansen, L., Anderson, H., Horndalsveen, P., Nygaard, H., Naik, M., Hoprekstad, D., Wetterberg, P., Nilsson, C., Aarsland, D., Sharma, P. and Lönneborg, A. (2011) A Novel Blood Test for the Early Detection of Alzheimer's Disease. *Journal of Alzheimer's Disease : JAD* **23**, 121–129.

[9] Lunnon, K., Ibrahim, Z., Proitsi, P. and Lourdusamy, A. (2012) Mitochondrial Dysfunction and Immune Activation are Detectable in Early Alzheimer's Disease Blood. *Journal of Alzheimer's Disease : JAD* **30**, 685–710.

[10] Lunnon, K., Sattlecker, M., Furney, S., Coppola, G., Simmons, A., Proitsi, P., Lupton, M., Lourdusamy, A., Johnston, C., Soininen, H., Kloszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Geschwind, D., Lovestone, S., Dobson, R., Hodges, A. and the AddNeuromed Consortium (2013) A Blood Gene Expression Marker of Early Alzheimer's Disease. *Journal of Alzheimer's Disease : JAD* **33**, 737–753.

[11] Han, G., Wang, J., Zeng, F., Feng, X., Yu, J., Cao, H. Y., Yi, X., Zhou, H., Jin, L. W., Duan, Y., Wang, Y. J. and Lei, H. (2013) Characteristic Transformation of Blood Transcriptome in Alzheimer's Disease. *Journal of Alzheimer's Disease : JAD* **35**, 373–386.

[12] Cooper-Knock, J., Kirby, J., Ferraiuolo, L., Heath, P. R., Rattray, M. and Shaw, P. J. (2012) Gene expression profiling in human neurodegenerative disease. *Nature Publishing Group* **8**, 518–530.

[13] Hänzelmann, S., Castelo, R. and Guinney, J. (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 1–15.

[14] Newhook, T., Blais, E., Lindberg, J., Adair, S., Xin, W., Lee, J., Papin, J., Parsons, J. and Bauer, T. (2014) A Thirteen-Gene Expression Signature Predicts Survival of Patients with Pancreatic Cancer and Identifies New Genes of Interest. *PLoS ONE* **9**, e105631.

[15] Chéry, L., Lam, H., Coleman, I., Lakely, B., Coleman, R., Larson, S., Aguirre-Ghiso, J., Xia, J., Gulati, R., Nelson, P., Montgomery, B., Lange, P., Snyder, L., Vessella, R. and Morrissey, C. (2014) Characterization of single disseminated prostate cancer cells reveals tumor cell heterogeneity and identifies dormancy associated pathways. *Oncotarget* **5**, 9939–9951.

[16] Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E. and Kotelnikova, E. (2014) Clustering Gene Expression Regulators: New Approach to Disease Subtyping. *PLoS ONE* **9**, e84955.

[17] Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., Spenger, C., Tsolaki, M., Vellas, B., Wahlund, L.-O., Ward, M. and on behalf of the AddNeuroMed Consortium (2009) AddNeuroMed-The European Collaboration for the Discovery of Novel Biomarkers for Alzheimer's Disease. *Annals of the New York Academy of Sciences* **1180**, 36–46.

[18] Hye, A., Lynham, S., Thambisetty, M., Causevic, M., Campbell, J., Byers, H., Hooper, C., Rijsdijk, F., Tabrizi, S., Banner, S., Shaw, C., Foy, C., Poppe, M., Archer, N., Hamilton, G., Powell, J., Brown, R., Sham, P., Ward, M. and Lovestone, S. (2006) Proteome-based plasma biomarkers for Alzheimer's disease. *Brain* **129**, 3042–3050.

[19] R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[20] Du, P., Kibbe, W. and Lin, S. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548.

[21] Xie, Y. (2010) *MBCB: MBCB (Model-based Background Correction for Beadarray)*. R package version 1.18.0.

[22] Oldham, M., Langfelder, P. and Horvath, S. (2012) Network methods for describing sample relationships in genomic datasets: application to Huntingdon's disease. *BMC Systems Biology* **6**, doi: 10.1186/1752–0509–6–63.

[23] Johnson, W. E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127.

[24] Dixit, A. and Dobson, R. (2014) CohortExplorer: A Generic Application Programming Interface for Entity Attribute Value Database Schemas. *JMIR Medical Informatics* **2**, e32.

[25] Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550.

[26] Kuhn, J., M. Contributions from Wing, Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z. and the R Core Team (2014) *caret: Classification and Regression Training*. R package version 6.0-35.

[27] Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News* **2**, 18–22.

[28] Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941.

[29] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.

[30] Burnham, S., Faux, N., Wilson, W., Laws, S., Ames, D., Bedo, J., Bush, A., Doecke, J., Ellis, K., Head, R., Jones, G., Kiiveri, H., Martins, R., Rembach, A., Rowe, C., Salvado, O., Macaulay, S., Masters, C. and Villemagne, V. (2014) A blood-based predictor for neocortical Aβ burden in Alzheimer's disease: results from the AIBL study. *Molecular Psychiatry* **19**, 519–526.

[31] Sattlecker, M., Kiddle, S., Newhouse, S., Proitsi, P., Nelson, S., Williams, S., Johnston, C., Killick, R., Simmons, A., Westman, E., Hodges, A., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S., Dobson, R. and Consortium, t. A. (2014) Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimer's & Dementia* **10**, 724–734.

[32] Strobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, doi:10.1186/1471–2105–8–25.

[33] Meng, Y., Yu, Y., Cupples, L., Farrer, L. and Lunetta, K. (2009) Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* **10**, doi:10.1186/1471–2105–10–78.

[34] Hothorn, T., Hornik, K. and Zeileis, A. (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* **15**, 651 –674.

[35] Strobl, C., Malley, J. and Tutz, G. (2009) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods* **14**, 323 –348.

[36] Tarca, A., Draghici, S., Khatri, P., Hassan, S., Mittal, P., Kim, J., Kim, C., Kusanovic, J. and Romero, R. (2009) A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82.

[37] Hung, J., Whitfield, T., Yang, T., Hu, Z., Weng, Z. and DeLisi, C. (2010) Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biology* **11**, doi:10.1186/gb–2010–11–2–r23.

Supplementary Information

7 GSVA in detail [13]

1. The expression profile of each gene is bought to a common scale using a non-parametric kernel estimation of the cumulative density function (CDF). For microarray data a Gaussian kernel is used.

2. Genes are ranked, per subject, by highest expression level.

3. The influence of outliers is reduced by working in terms of normalised gene rank. Let $z_{(i)j}$ be the gene rank for the i^{th} subject and the j^{th} gene. Define r_{ij} as the normalised gene rank and p as the number of genes.

$$r_{ij} = |\tfrac{p}{2} - z_{(i)j}|$$

4. A per subject distribution over the genes is produced to assess if the genes in the gene set are more likely to be in either tail of the rank distribution. This should create a p x *pathway* matrix for each subject. Let $l = 1, ..., p$, I_{in} be an indicator function for genes belonging to a pathway and I_{out} be an indicator function for genes not belonging to a pathway.

$$\textit{Gene level score} = \frac{\sum_{i=1}^{l} |r_{ij}|^{\tau} I_{in}}{\sum_{i=1}^{p} |r_{ij}|^{\tau} I_{in}} - \frac{\sum_{i=1}^{l} I_{out}}{p - |Pathway|}$$

5. A pathway level enrichment score is created using the classical method of maximum deviation. Alternatively, Hänzelmann et al. propose a new method to create a unimodal statistic under the null hypothesis that there is no change in pathway activity throughout the sample population. Here, $GeneLevelScore_{+}$ are all positive scores and $GeneLevelScore_{-}$ are all negative scores.

$$\textit{Enrichment Score} = MAX(0, GeneLevelScore_{+}) - |MIN(0, GeneLevelScore_{-})|$$

8 Data analysis in detail

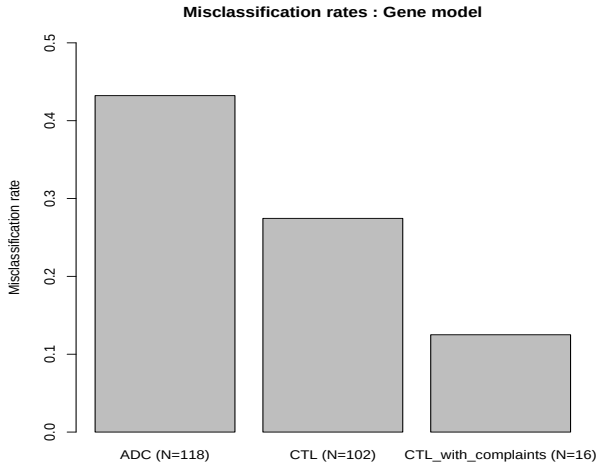
To build the *gene model* and *pathway model* the batch 1 data was bootstrapped 100 times. The bootstrap samples were of the same size as the batch 1 data and sampled with replacement. For each bootstrap sample a RF model was built using 5 fold cross validation to tune the model parameter *mtry* [27]. *mtry* is the number of variables randomly selected for consideration at each split in the decision tree. The model parameter that

determined the number of trees created within each model *ntree* was set to 501 throughout; an odd *ntree* was used to account for any ties. The change in Gini index was used to create variable importance scores. These were ranked across all variables per model and then summed across all bootstrap samples. Variables were ordered by this metric and plotted. The top 5% of variables were taken forward to the next stage.

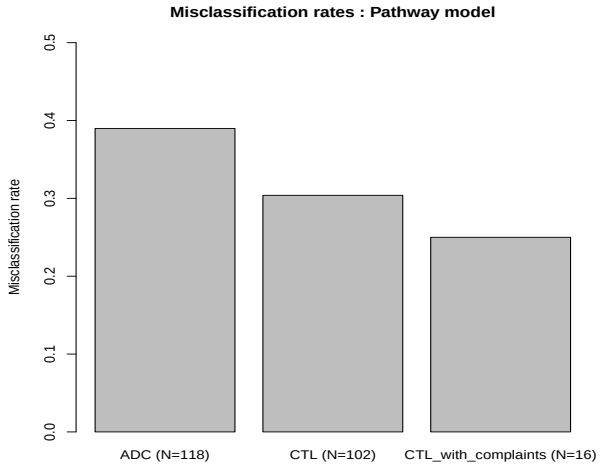
Recursive feature elimination (RFE) was performed on this subset of variables in the original batch 1 data for the *pathway model* and *gene model*. Building of the *demographic model* began at this stage using all variables. The feature elimination was again based on a RF model with *ntree* = 501 and investigated subsets of variables of all possible sizes. Caret's 'pickSizeTolerance' function was used (tolerance = 5%) to identify a further subset of variables. This function finds a smaller set of variables while maintaining model accuracy [26]. If this subset matched the RFE optimal set, the RFE model was taken forward. Otherwise, the optimum variables were selected using the 'selectVar' function in caret and a final RF model was built.

RF models were used throughout for their non-parametric, non-linear properties. Further, the use of bootstrapping in RF modeling and random selection of variables at each decision point decreases the dependence of these models on noise.

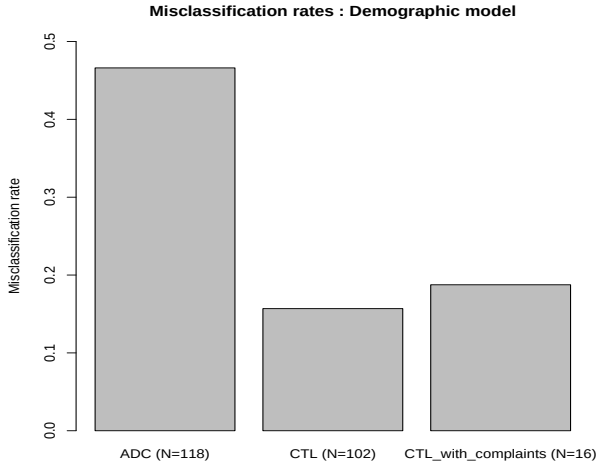
9 Misclassification investigation



(a) *Gene model*

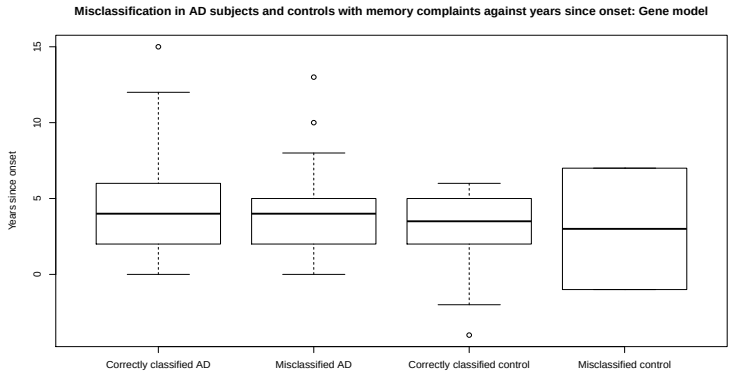


(b) *Pathway model*

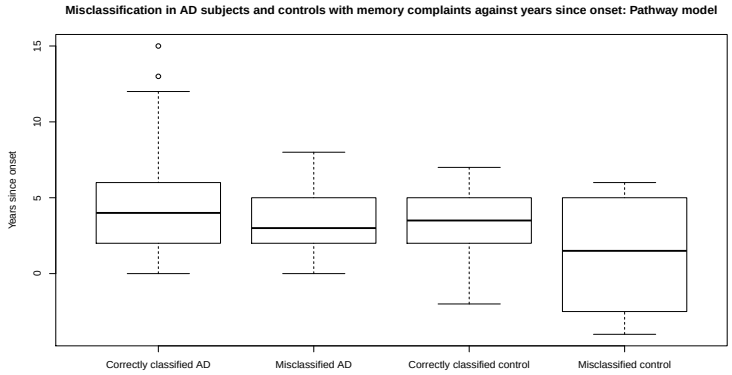


(c) *Demographic model*

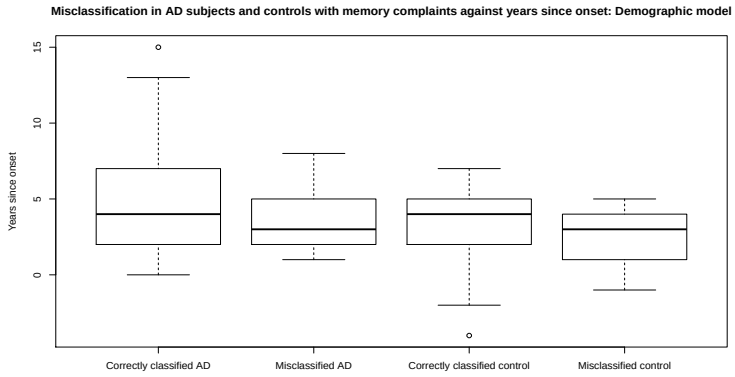
Figure S1: Misclassification rates by disease status in an independent test set



(a) *Gene model*



(b) *Pathway model*



(c) *Demographic model*

Figure S2: Years since onset by diagnosis and misclassification status in an independent test set

Figure S3: Concordance in gene expression between batches 1 and 2