Exploring Alzheimer's Disease Biomarkers with Linear Modelling

Annie G. Bryant, OLET5608 May 2022 | The University of Sydney

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative condition that is the primary cause of dementia globally[1]. One of the primary neuropathological hallmarks of AD is the aggregation of hyperphosphorylated tau protein (p-Tau) into neurofibrillary tangles inside neurons[2]. While this occurs within the brain in AD, such pathological changes can also be detected in the cerebrospinal fluid (CSF) that flows between the brain and spinal cord. Several landmark studies over the past decade have shown that elevated levels of p-Tau protein in the CSF correlate with the progression of tau pathology in the brain and can even predict onset of cognitive decline years in advance. Given the relevance of this exciting biomarker, I sought to leverage open-access AD biomarker data to better understand the cognitive and biological factors that relate to increased CSF p-Tau via linear modelling.

2. Dataset

Cross-sectional cognitive, neuroanatomical, and demographic data were acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[3], which is a multi-centre neuroimaging consortium that provides open-access AD biomarker data. After preliminary data cleaning, I retained a dataset with N=701 observations (i.e. participants), N=1 outcome variable (CSF p-Tau levels), and N=19 predictor variables. These predictor variables can be generally organized into six categories: Cognitive (ADAS-11[4], ADAS-13[5], MMSE[6], CDR-SB[7]), Demographic (age, gender, education, race, marital status), genetic (*APOE* gene E4 allele status), neuroimaging (fluorodeoxyglucose positron emission tomography; FDG), neuroanatomical volume (ventricles, hippocampus, whole brain, entorhinal, fusiform, middle temporal cortex, total intracranial region), and peripheral biomarker (CSF amyloid-beta protein levels). The first four values and a description of each of these features is provided in **Table 1** on the next page.

Of note, I manually encoded PTGENDER, PTRACCAT, PTMARRY, and APOE4 as factors since they are classes rather than continuously distributed numerical variables. I also discretised PTEDUCAT into a new variable (PTHIGHERED) two groups: (1) 16 or fewer years and (2) greater than 16 years, since this is not a truly continuous variable.

3. Analysis

3.1 Exploratory data analysis and visualization

Before beginning any modelling, I visualized the univariate distribution for the predictor and outcome variables in this dataset (**Figure 1**). Most of the quantitative (continuously-distributed) predictor terms are somewhat normally distributed, with the exception of ADAS11, ADAS13, CDRSB, and Ventricles which exhibit positive skews; additionally, the outcome feature (PTAU) also exhibits a positive skew.

Variable	First 4 Values	Description
RID	3, 5, 10, 14	Unique participant identifier ID
AGE	81.3, 73.7, 73.9, 78.5	Age in years
PTGENDER	Male, Male, Female, Female	Sex of the participant
PTEDUCAT	18, 16, 12, 12	Years of education
PTRACCAT	White, White, White, White	Ethnicity
PTMARRY	Married, Married, Married, Divorced	Marital status
APOE4	TRUE, FALSE, TRUE, FALSE	Presence of one or two APOE E4 alleles
FDG	1.08355, 1.29343, 1.11532, 1.25096	Average fluorodeoxyglucose (FDG) positron emission tomography (PET) metabolism of angular, temporal, and posterior cingulate cortices
ABETA	741.5, 547.3, 357.4, 1582.0	Amount of amyloid-beta protein in the CSF
PTAU	22.83, 33.43, 31.26, 16.68	Amount of phosphorylated tau protein in the CSF
MMSE	20, 29, 24, 29	Mini-Mental State Examination (MMSE) cognitive score
CDRSB	4.5, 0.0, 5.0, 0.0	Clinical Dementia Rating Sum of Box (CDR-SB) score
ADAS11	22.00, 8.67, 12.33, 4.33	Alzheimer's Disease Assessment Scale (ADAS) Cognitive Subscale test with eleven components of language and memory function
ADAS13	31.00, 14.67, 24.33, 8.33	ADAS11 test with additional delayed recall and digit cancellation tasks
Ventricles	84599, 34062, 26820, 46279	Volume (mm3) of the ventricles
Hippocampus	5319, 7075, 5485, 6730	Volume (mm3) of the hippocampus
WholeBrain	1129830, 1116630, 1033540, 861749	Volume (mm3) of the whole brain
Entorhinal	1791, 4433, 2676, 3581	Volume (mm3) of the entorhinal cortex
Fusiform	15506, 24788, 16761, 13779	Volume (mm3) of the fusiform gyrus
MidTemp	18422, 21614, 19741, 17798	Volume (mm3) of the middle temporal cortex
ICV	1920690, 1640770, 1471180, 1269540	Volume (mm3) of the whole brain (i.e. brain tissue plus ventricles combined)

Table 1. Description and first four values of all variables in the ADNI dataset.

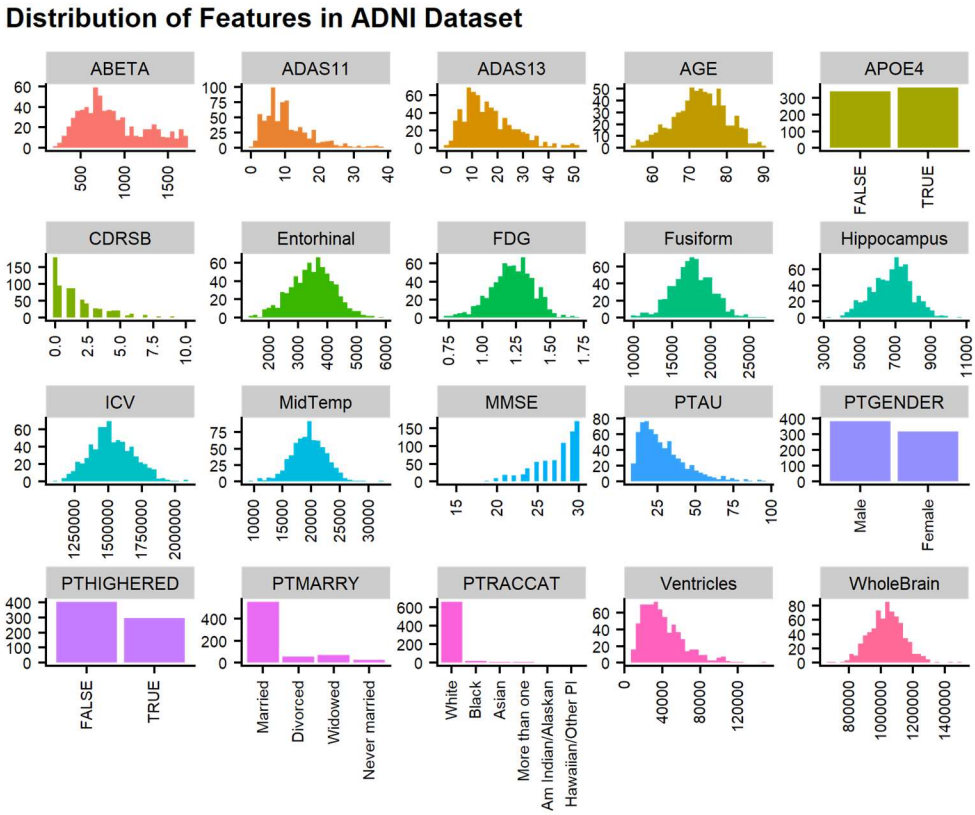

Figure 1. Distribution of all variables in the ADNI dataset.

3.2 Model fitting and diagnostics

There are four core assumptions of a linear regression model[8]:

1. A linear relationship between the predictor and outcome variable(s)
2. Normal distribution of residuals
3. Constant variation across residuals
4. Independent observations

For reference in subsequent sections, the five ordinary least squares (OLS) regression models evaluated are outlined in **Table 2**.

Model	Formula
Model 1	lm(PTAU ~ ., data = ADNI_data)
Model 2	lm(PTAU ~ . - ADAS11 - ICV, data = ADNI_data)
Model 3	lm(log(PTAU) ~ . - ADAS11 - ICV, data = ADNI_data)
Model 4	lmtest::coeftest(model3, vcov = vcovHC(model3, "HC3"), save=T)
Model 5	MASS::stepAIC(model3, direction="both", trace=FALSE)

Table 2. OLS models evaluated in this analysis and their corresponding R formula calls.

3.2.1 Linear relationship

Given the vast differences in scale between e.g. FDG and WholeBrain, I first applied z-score normalization before constructing any models, in which each feature is mean-centred and standardized. This technique puts each quantitative variable in the same scale, thus enabling direct comparison of coefficients derived from subsequent linear models. Once z-scored, the linear relationship between each quantitative predictor and CSF PTAU can be visually assessed with scatterplots (**Figure 2**). Of note, the y-axis in **Figure 2** shows log-transformed CSF PTAU levels, which is further discussed in Section 3.2.2. There is a visually positive linear association between the outcome (PTAU levels in CSF) and ADAS11, ADAS13, AGE, and a visually negative linear association with ABETA, Entorhinal, FDG, Fusiform, Hippocampus, ICV, MidTemp, MMSE, and Ventricles, WholeBrain.

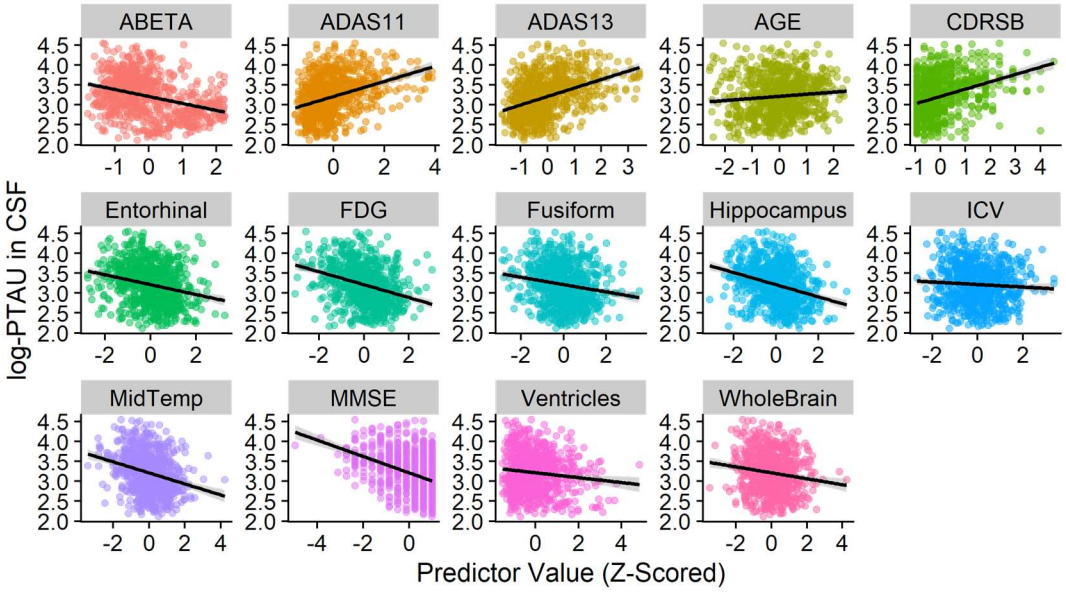

Figure 2. Linear relationships between each quantitative (z-scored) predictor variable and the log-transformed outcome variable (PTAU in CSF).

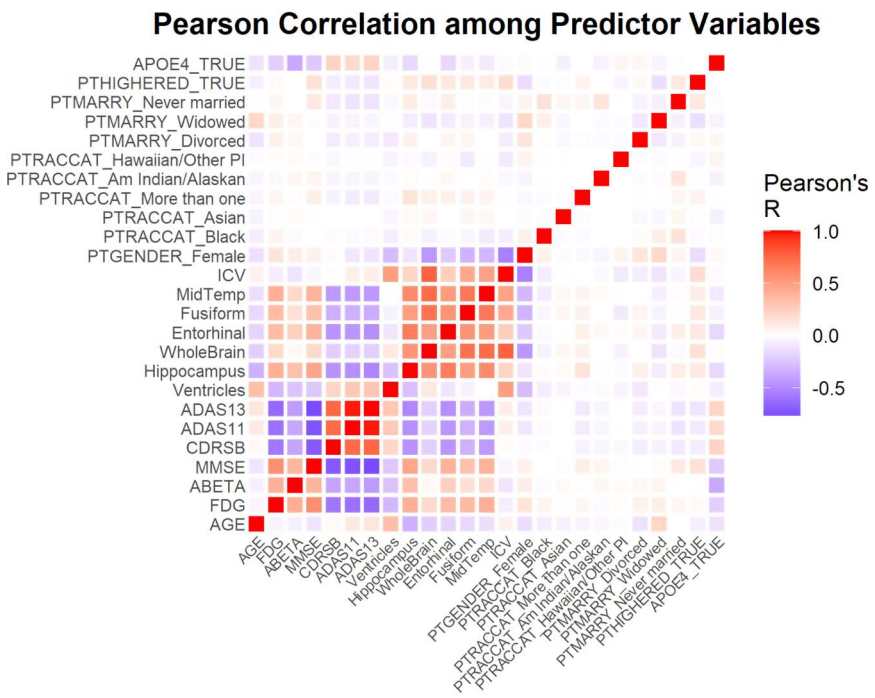

Figure 3. Pearson correlation (R) among all predictor terms in ADNI dataset.

I constructed a basic OLS linear regression model using all z-scored features in the dataset with PTAU as the outcome variable as a baseline; this model will be referred to as **Model 1**. To assess potential multicollinearity, I plotted the Pearson correlation between all predictor terms (**Figure 3**). Four features stand out as having particularly high correlation magnitudes: ADAS11, ADAS13, WholeBrain, and ICV. This was confirmed by calculating the variance inflation factor (VIF) across all predictor terms; all four of these terms had VIF values greater than 5, which is generally considered to be a threshold above which multicollinearity poses a problem[9]. Since ADAS13 encompasses allADAS11 test components with the addition of two new components, I dropped ADAS11 going forward. Additionally, WholeBrain captures just the brain tissue volume whereas ICV also contains the volume of CSF in the brain, so I opted to keep WholeBrain and drop ICV going forward. This OLS model in which ADAS11 and ICV are omitted will be referred to as **Model 2**.

3.2.2 Normal distribution of residuals

The second assumption of a linear model is that the residuals are normally distributed. This assumption can be assessed using a quantile-quantile (Q-Q) plot, which shows the quantiles of the standardized residuals (y axis) versus the quartiles of a normal distribution (x axis). **Figure 4** shows the Q-Q plot and histogram of residual values for **Model 2**, in which the residuals are clearly positively skewed. To mitigate this issue, I tried transforming the outcome variable (PTAU) with a Box-Cox power transformation (**Figure 5**)[10]. In this type of visualization, the centre dashed line around 0 represents the

optimal λ for the power transformation, and the outer two vertical lines are the 95% confidence interval for λ.

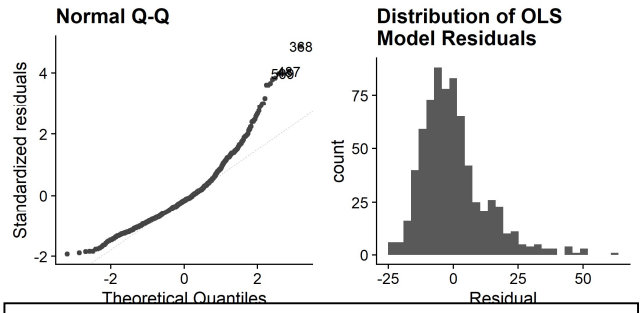

Figure 4. Distribution of residuals in **OLS Model 2**.

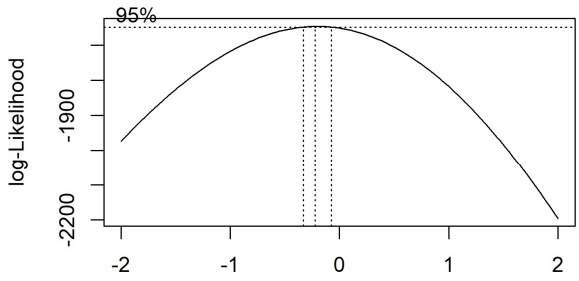

Figure 5. Box-Cox power transformation of PTAU.

While the 95% confidence interval does not precisely cover any λ value associated with a typical power transformation, we might still try using λ = 0 for a log transformation of the outcome variable (PTAU). I visually inspected the distribution after log-transforming PTAU versus square root-transforming or inverse-transforming it (**Figure 6**), which confirms that the log transform yields the closest to a normal distribution of PTAU. After adding log-transformed PTAU to **Model 2** (which is hereafter referred to as **Model 3**), I visualized the Q-Q plot and residual histogram again (**Figure 7**). The residuals do more closely follow a normal distribution now; given the large sample size (N=701 observations), it is safe to assume that this distribution is sufficiently normal based on the central limit theorem.

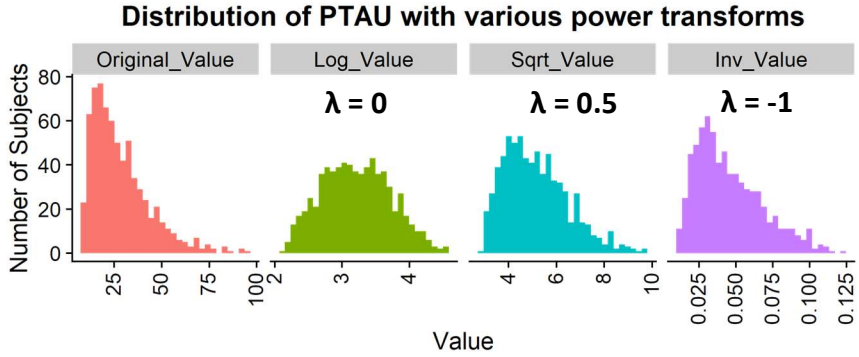

Figure 6. Distribution of CSF PTAU after applying various power transforms.

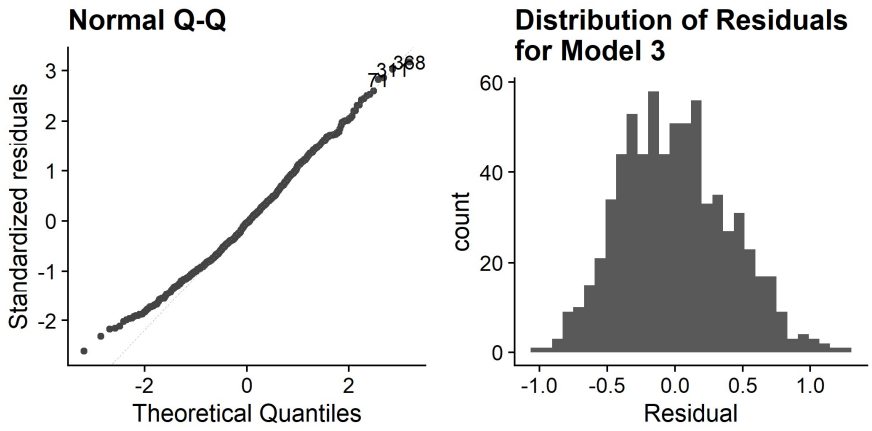

Figure 7. Distribution of residuals after log-transforming outcome variable PTAU (**Model 3**).

3.2.3 Constant variance across residuals

The third assumption of a linear model is that the residuals exhibit a consistent variance across all fitted values in the model -- in other words, that there is no discernible bias of the variance across residuals. We can evaluate this assumption by plotting either the raw residuals or the square root of the residuals versus the fitted values using the `ggfortify::autoplot` function, comparing **Model 2** with **Model 3** (**Figure 8**). For a linear model with constant variance across the fitted outcome variable values, we would expect to see a flat blue line in all four of these plots and no clear trend in residuals based on fitted values. However, the upper two plots (**Figure 8A-B**) show a funnelling effect in which the residuals fan out as the fitted values increase, indicating there is heteroscedasticity present in **Model 2**. I applied the Breusch-Pagan test to **Model 2**, which returned $p<0.0001$, supporting the conclusion that there is heteroscedasticity present in this model.

While heteroscedasticity does not appear to be an issue for **Model 3** (**Figure 8C-D**), the Breusch-Pagan test still returned $p<0.01$. As heteroscedasticity is an issue for linear models – particularly in estimating standard errors – I opted to calculate robust standard errors using a heteroscedasticity-robust estimator[11–13]. With this method, standard errors are calculated with the sandwich estimator of variance and are valid even in the presence of heteroscedasticity -- and are often larger than conventionally estimated standard errors. I implemented such an estimator using the "`sandwich`" package in R[14,15], and the resulting model will be referred to as **Model 4**.

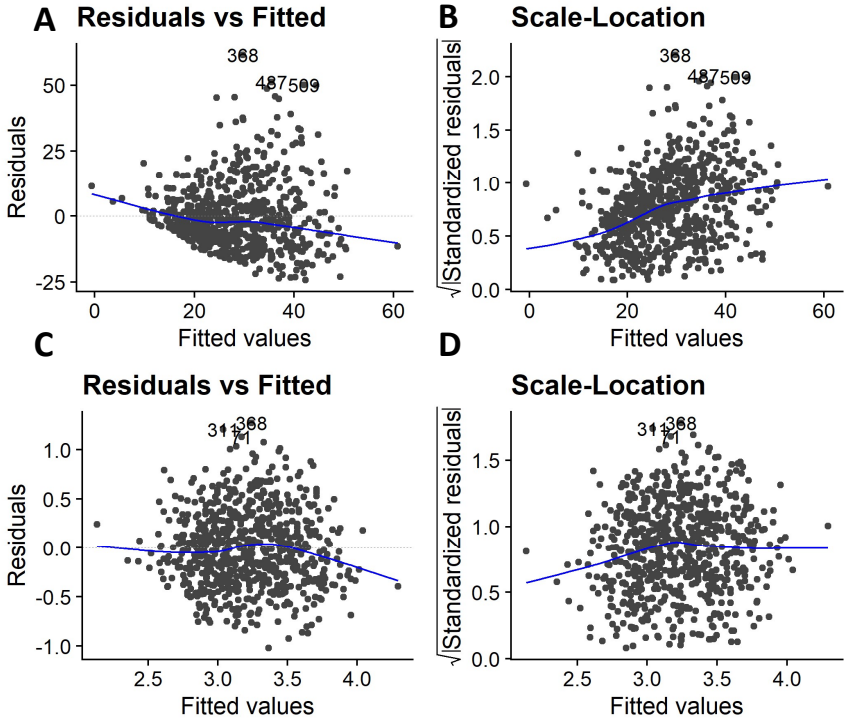

Figure 8. Residuals versus fitted values from `ggfortify::autoplot`.

8A-B are for **Model 2**, **8C-D** are for **Model 3**.

3.2.4 Independence of observations

The final assumption of a linear model is that the observations are independent. The ADNI dataset examined in this analysis contains N=701 distinct and biologically independent individuals, and as the

dataset is not time-series data, there is no concern for temporal autocorrelation. As such, it is safe to assume that all observations in this dataset are independent, thus fulfilling this final assumption.

4. Results

Term	Model 3	Model 4	Model 5
APOE4TRUE	**0.21*** (0.035)**	**0.21*** (0.037)**	**0.205* (0.035)**
ADAS13	**0.103** (0.03)**	**0.103* (0.031)**	**0.102* (0.027)**
AGE	**0.097*** (0.019)**	**0.097*** (0.02)**	**0.095* (0.018)**
PTGENDERFemale	**0.096* (0.04)**	**0.096* (0.039)**	**0.078* (0.037)**
WholeBrain	**0.063* (0.029)**	**0.063* (0.029)**	**0.053* (0.022)**
Entorhinal	0.028 (0.023)	0.028 (0.024)	
Fusiform	0.01 (0.025)	0.01 (0.024)	
PTHIGHEREDTRUE	0.005 (0.032)	0.005 (0.033)	
PTMARRYNever married	0.001 (0.087)	0.001 (0.092)	
CDRSB	-0.004 (0.026)	-0.004 (0.03)	
PTMARRYWidowed	-0.034 (0.055)	-0.034 (0.058)	
MidTemp	-0.035 (0.027)	-0.035 (0.027)	
FDG	**-0.045* (0.022)**	**-0.045* (0.022)**	**-0.05* (0.021)**
Hippocampus	-0.045 (0.026)	-0.045 (0.027)	**-0.039* (0.023)**
ABETA	**-0.074*** (0.019)**	**-0.074*** (0.017)**	**-0.075* (0.019)**
MMSE	**-0.075* (0.026)**	**-0.075* (0.028)**	**-0.07* (0.024)**
PTMARRYDivorced	-0.082 (0.06)	-0.082 (0.059)	
PTRACCATAsian	-0.135 (0.146)	-0.135 (0.231)	-0.135 (0.145)
PTRACCATBlack	-0.146 (0.093)	-0.146 (0.101)	-0.142 (0.091)
Ventricles	**-0.174*** (0.019)**	**-0.174*** (0.02)**	**-0.176* (0.019)**
PTRACCATMore than one	**-0.274* (0.138)**	**-0.274* (0.1)**	-0.275 (0.137)
PTRACCATAm Indian/Alaskan	-0.374 (0.292)	-0.374 (0.218)	-0.376 (0.289)
PTRACCATHawaiian/Other PI	-0.418 (0.292)	**-0.418*** (0.059)**	-0.406 (0.288)
R^2	0.3870	0.3870	0.3820
Adjusted R^2	0.3662	0.3662	0.3685

Table 3. Comparison of predictor term coefficients across OLS Models 3, 4, and 5. Values are presented as coefficient (SE). Predictor terms in bold had statistically significant coefficient estimates in at least one of the three models. Blank cells in the Model 5 column correspond to terms which were not retained in the stepwise regression model. ***p<0.001, **p<0.01, *p<0.05.

The resulting coefficients from OLS **Models 3 and 4** are shown in **Table 3**. Values are presented as the estimated coefficient (standard error). In the case of **Model 4**, the standard error (SE) values are those derived from the heteroscedasticity-robust standard estimator (see section 3.2.3). As these models both contain all predictor terms (with the exception of ADAS11 and ICV as discussed in section 3.2.1), I opted to explore feature selection with bidirectional stepwise regression using the `stepAIC()` function from the `MASS` package[16]. In this method, predictor terms are iteratively added based on their improvement to the model's Akaike Information Criterion (AIC), and terms are subsequently pruned if they do not contribute any benefit to model performance (as measured via AIC). The resulting model from stepwise

regression is referred to as **Model 5**, and the coefficients and SEs from retained predictor terms are also shown in **Table 3**.

Of note, the predictor terms with statistically significant coefficients are generally very consistent across the three models compared in **Table 3**. There are three key exceptions: (1) Hippocampus volume is only significant in **Model 5**; PTRACCATMore than one (i.e. being of more than one race) is only significant in **Models 3 and 4**; and PRACCATHawaiian/Other PI (i.e. being of Hawaiian/Pacific Islander descent) is only significant in **Model 4**. Otherwise, the statistically significant predictors with a positive association with PTAU CSF are APOE4TRUE (presence of one or two E4 alleles in the *APOE* gene), ADAS13, AGE, PTGENDERFemale, and WholeBrain, while the significant predictors with a negative association with PTAU CSF are FDG-PET metabolism, ABETA levels in CSF, MMSE cognitive scores, Ventricle volume, PTRACCATMore than one, and PTRACCATHawaiian/Other PI.

The three models compared in **Table 3** exhibit very similar adjusted R^2 values: **Model 3** = 0.3662, **Model 4** = 0.3662, and **Model 5** = 0.3685. The marginal improvement in R^2 in **Model 5** is likely due to the decreased number of predictor terms following stepwise regression, as fewer terms incur a smaller penalty for the adjusted R^2 value. The lack of appreciable difference in variance explained between **Models 3/4** and **Model 5** indicates that the variables omitted from **Model 5** explain minimal variance in CSF PTAU levels.

5. Discussion and conclusions

This analysis sought to elucidate the linear relationships between a diverse set of predictor terms and the amount of PTAU in the CSF of a large multi-centre AD biomarker consortium with statistically robust methodology. This analysis, while not causal in nature, identified several key clinical and neurobiological factors associated with CSF PTAU levels. The presence of one or more E4 alleles in the *APOE* gene had the largest coefficient that remained statistically significant across all models examined; this is congruent with a rich literature documenting the association between presence of the E4 allele and AD severity[17]. A higher score on the ADAS13 test indicates greater severity in cognitive decline, which helps explain the positive association between ADAS13 scores and CSF PTAU. Previous work has characterized a positive linear relationship between age and CSF PTAU levels even in healthy adults[18], which is echoed in the results presented here. The significant positive coefficient for female sex is in line with findings that being female comes with higher risk of developing AD[19]; however, a transgenic mouse model of AD showed that female sex was associated with heightened amyloid-beta pathology but not tau pathology[20]. The positive association between whole-brain volume and CSF PTAU is surprising given that elevated CSF PTAU is typically associated with neurodegeneration and reduced brain volume[21]

Decreased FDG-PET indicates reduced synaptic activity and is directly related to neurodegeneration in AD; the negative coefficient for FDG-PET metabolism found in this analysis corresponds with previous findings[22]. The negative coefficient for CSF ABETA is surprising, given that previous studies reported a positive correlation between CSF PTAU and ABETA levels[23]. The negative association between MMSE scores and CSF PTAU is in line with previous findings and may be partially mediated by amyloid-beta pathology[24]. The negative coefficient for ventricular volume is consistent with a previous study that found a similar relationship, with APOE allele status exerting a modulatory effect[25]. The negative coefficients for two racial backgrounds (“More than one” and “Hawaiian/Other PI”) are to be interpreted with caution given the very small number of participants in these groups (N=9 and N=2, respectively). However, as the

relationship between race and CSF biomarkers in AD has yet to be comprehensively analysed, these potential relationships warrant further study in a more racially diverse study population.

In addition to the aforementioned limitation due to limited non-Caucasian participants in this study, another limitation is that of feature selection. The dataset analysed in this project was manually cleaned and filtered prior to modelling based on data availability and domain knowledge, although some of the omitted variables may comprise the roughly 65% of unexplained variance in CSF PTAU levels. For example, a minority of participants had phosphorylated tau PET imaging data available, which is known to be correlated with CSF PTAU levels[26], but this variable was omitted due to insufficient observations with these measurements. Another limitation is that of model selection; while multiple OLS-based models were evaluated using power transformations and/or stepwise regression, it is possible that another type of linear model may be better suited for the dataset at hand. Future research should examine alternative models such as a generalised linear model.

6. References

1. Nichols, E. et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 18, 88–106 (2019).
2. Zhou, Y. et al. Relevance of Phosphorylation and Truncation of Tau to the Etiopathogenesis of Alzheimer's Disease. Front. Aging Neurosci. 10, 27 (2018).
3. Weiner, M. W. et al. The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. Alzheimers Dement. 13, 561–571 (2017).
4. Rosen, W. G., Mohs, R. C. & Davis, K. L. A new rating scale for Alzheimer's disease. Am. J. Psychiatry 141, 1356–64 (1984).
5. Mohs, R. C. et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. Alzheimer Dis. Assoc. Disord. 11, 13–21 (1997).
6. Folstein, M. F., Folstein, S. E. & McHugh, P. R. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198 (1975).
7. Hughes, C. P., Berg, L., Danziger, W., Coben, L. A. & Martin, R. L. A New Clinical Scale for the Staging of Dementia. Br. J. Psychiatry 140, 566–572 (1982).
8. Marill, K. A. Advanced statistics: linear regression, part I: simple linear regression. Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med. 11, 87–93 (2004).
9. James, G., Witten, D., Hastie, T. & Tibshirani, R. An Introduction to Statistical Learning: with Applications in R. vol. 112 (Springer US, 2013).
10. Box, G. E. P. & Cox, D. R. An Analysis of Transformations. J. R. Stat. Soc. Ser. B Methodol. 26, 211–252 (1964).
11. White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica 48, 817–838 (1980).
12. Eicker, F. Limit theorems for regressions with unequal and dependent errors. in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967).
13. Huber, P. J. The behavior of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Stat. Probab. Vol. 1 Stat. 5.1, 221–234 (1967).
14. Zeileis, A. Econometric Computing with HC and HAC Covariance Matrix Estimators. J. Stat. Softw. 11, 1–17 (2004).

15. Zeileis, A., Köll, S. & Graham, N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. J. Stat. Softw. 95, 1–36 (2020).
16. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S-PLUS. (Springer Science & Business Media, 2013).
17. Liu, C.-C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat. Rev. Neurol. 9, 106–118 (2013).
18. Sjögren, M. et al. Tau and Aβ42 in Cerebrospinal Fluid from Healthy Adults 21–93 Years of Age: Establishment of Reference Values. Clin. Chem. 47, 1776–1781 (2001).
19. Scheyer, O. et al. Female Sex and Alzheimer's Risk: The Menopause Connection. J. Prev. Alzheimers Dis. 5, 225–230 (2018).
20. Hirata-Fukae, C. et al. Females exhibit more extensive amyloid, but not tau, pathology in an Alzheimer transgenic model. Brain Res. 1216, 92–103 (2008).
21. Ingber, A. P. et al. Cerebrospinal Fluid Biomarkers and Reserve Variables as Predictors of Future "Non-Cognitive" Outcomes of Alzheimer's Disease. J. Alzheimers Dis. JAD 52, 1055–1064 (2016).
22. Petrie, E. C. et al. Preclinical Evidence of Alzheimer Changes: Convergent Cerebrospinal Fluid Biomarker and Fluorodeoxyglucose Positron Emission Tomography Findings. Arch. Neurol. 66, 632–637 (2009).
23. Hansson, O. et al. CSF biomarkers of Alzheimer's disease concord with amyloid-β PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts. Alzheimers Dement. 14, 1470–1481 (2018).
24. Blennow, K. et al. Predicting clinical decline and conversion to Alzheimer's disease or dementia using novel Elecsys Aβ(1–42), pTau and tTau CSF immunoassays. Sci. Rep. 9, 19024 (2019).
25. Ott, B. R. et al. Brain ventricular volume and cerebrospinal fluid biomarkers of Alzheimer's disease. J. Alzheimers Dis. JAD 20, 647–657 (2010).
26. Tissot, C. et al. Comparing tau status determined via plasma pTau181, pTau231 and [18F]MK6240 tau-PET. eBioMedicine 76, 103837 (2022).