Pertanika J. Sci. & Technol. 30 (2): 1197 - 1218 (2022)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Identification of Blood-Based Multi-Omics Biomarkers for Alzheimer's Disease Using Firth's Logistic Regression

Mohammad Nasir Abdullah[1]*, Yap Bee Wah[2,3], Abu Bakar Abdul Majeed[4], Yuslina Zakaria[5] and Norshahida Shaadan[4]

[1]Department of Statistics, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Perak, Kampus Tapah, 35400 UiTM, Tapah, Perak, Malaysia
[2]Institute for Big Data Analytics and Artificial Intelligence, Kompleks Al-Khawarizmi, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Selangor, Malaysia
[3]Centre of Statistical and Decision Science Studies, Universiti Teknologi MARA, 40450 UiTM, Shah Alam, Selangor, Malaysia
[4]Brain Research Laboratory, Faculty of Pharmacy, Universiti Teknologi MARA, Cawangan Selangor, Kampus Puncak Alam, 42300 UiTM, Puncak Alam, Selangor, Malaysia
[5]Faculty of Pharmacy, Universiti Teknologi MARA, Cawangan Selangor, Kampus Puncak Alam, 42300 UiTM, Puncak Alam, Selangor, Malaysia

ABSTRACT

Alzheimer's disease (AD) is a progressive and relentless debilitating neurodegenerative disease. A post-mortem microscopic neuropathological examination of the brain revealed the existence of extracellular β-amyloid plaques and intracellular neurofibrillary tangles. An accurate early diagnosis of AD is difficult because various disorders share the initial symptoms of the disease. Based on system biology, the multi-omics approach captures and integrates information from genomics, transcriptomics, proteomics, cytokinomics, and metabolomics. This study developed an AD prediction model based on the integrated blood-based multi-omics dataset involving 32 AD patients and 15 non-AD subjects. The integrated multi-omics dataset consists of 16 transcript genes, 14 metabolites, and nine cytokines. Due to the complete separation and multicollinearity issues, Firth's logistic regression model was then developed to predict AD using the principal components. The model revealed 18 potential biomarkers of AD, consisting of seven metabolites, two transcriptomes, and nine cytokines. These potential biomarkers show an upregulated

ARTICLE INFO

Article history:
Received: 18 August 2021
Accepted: 12 November 2021
Published: 14 March 2022

DOI: https://doi.org/10.47836/pjst.30.2.19

E-mail addresses:
nasir916@uitm.edu.my (Mohammad Nasir Abdullah)
yapbeewah@uitm.edu.my (Yap Bee Wah)
abubakar@uitm.edu.my (Abu Bakar Abdul Majeed)
yuslina@uitm.edu.my (Yuslina Zakaria)
norshahida588@uitm.edu.my (Norshahida Shaadan)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

risk in the AD group compared to the non-AD subjects. The possibility of using these biomarkers as early predictors of AD is discussed.

Keywords: Alzheimer's disease, biomarkers, complete separation, Firth's logistic regression, multi-omics

INTRODUCTION

Alzheimer's disease (AD) is a progressive and debilitating disorder. Rare autosomal dominant mutations seem to cause the early onset of AD (EOAD) (Bertram et al., 2007; Cummings & Jeste, 1999; Gross et al., 2012). There is currently no treatment available to cure AD (Cayton et al., 2008; Ibáñez et al., 2013; Maskery et al., 2020; Von Schulze et al., 2020). An accurate early diagnosis of AD is also difficult because initial symptoms of the disease are shared with a variety of disorders, which reflect common neuropathological features (Humpel, 2011; Minter et al., 2016). Genetic studies provide an opportunity to elucidate the cause of a disease for early detection or cure (Marioni et al., 2018; Tanzi, 2012; Waring & Rosenberg, 2008). The genetic study of AD has advanced over the last decade, where more than twenty independent loci or locations of genes on the chromosome are known to be associated with the disease, besides the well-established gene, APOE (Marioni et al., 2018). Biomarkers can serve as predictors of health and disease. They can be used to indicate normal biological processes, abnormal pathogenic conditions, or pharmacological responses to therapeutic drugs (Gomez-Ramirez & Wu, 2014; Humpel, 2011; Zhang, 2011). In the past decade, omics approaches and technologies have contributed to studying the metabolome, lipidome, and proteome in a complex disease such as AD (Clark et al., 2021; Hasin et al., 2017).

Integrative omics is a new biological research field that studies system biology, capturing information from genomics, transcriptomics, proteomics, metabolomics, and cytokinomics. For instance, transcriptomics is the full complement of messenger ribonucleic acid (mRNA) in a cell or tissue at any given moment. It is a form of protein synthesis which results in a corresponding protein complement to the proteome. It has been used to describe the global mRNA expression of a particular tissue, yielding information about the transcriptional differences between two or more states (Romero et al., 2006). In contrast, metabolomics aims to identify and quantify the global composition of "metabolites" of a biological fluid, tissue, or organism. Metabolites are small molecules (non-polymeric compounds) that participate in general metabolic reactions and are required for the maintenance, growth, and normal function of cells (Kussmann et al., 2006; Romero et al., 2006; Zhou et al., 2014). Metabolomics is an in-depth study since the metabolic network is downstream from gene expression and protein synthesis, where it reflects more closely cell activity at a functional level (Romero et al., 2006). Cytokinomics is a large-scale study of small proteins commonly known as cytokines or glycoproteins produced by several

cell types in biological systems (Clerici, 2010). They are a group of proteins concealed by cells of the immune system that act as chemical messengers. Inflammation was found to initiate or cause the deterioration of AD neurodegeneration. Several different cytokines have been reported to be higher in AD patients (Park et al., 2020; Swardfager et al., 2010; Zheng et al., 2016).

Multi-omics integration is important as more information is needed on the inter-individual variations and complex biomarkers' interrelations on AD identification and disease progression. There are some issues when dealing with multi-omics data, which are the correlated features (genes). Genes usually work in a group, are connected to other genes, and form a network to operate well in the human body. Thus, this would be a challenge in predicting the biomarkers of the disease because classical statistical methods usually do not tolerate correlated features (multicollinearity). For most multi-omics data that focus on certain disease measures, there would be a risk of a complete separation issue. It usually happens when the sample size of the dataset is lower than the number of variables. In certain diseases under study, some of the biomarkers in the multi-omics dataset would show a tremendous gap between cases and control groups. This condition would interfere with the analysis of finding other biomarkers. These conditions might complicate the process of data analysis since most of the standard analytical procedures do not focus on monotone likelihood estimation. The monotone likelihood was the effect of a complete separation dataset.

Furthermore, the analysis of multi-omics data might be complicated due to existing conditions such as multicollinearity and complete separation issues. There are currently no analytical methods that can simultaneously address this condition.

Past studies have explored many methods to handle data with multicollinearity, such as ridge regression, partial least square regression, and principal component analysis (Adnan et al., 2006; Rahayu et al., 2017). According to Rougoor et al. (2000), when the number of observations is large, the difference in performance across the methods is often minimal. No one approach dominates the others.

When a complete separation issue occurs, the options to (i) increase the sample size, (ii) combine the category with the separation issue with a similar one (for more than two categories), (iii) remove the class (for more than two categories) can be considered. However, increasing the sample size in a clinical trial is not always feasible, and combining categories is not always practicable, particularly when there are only two categories, and each category is meant to be mutually exclusive. Finally, omitting the category may be too risky, as the category may be crucial to the study. The separation problem can be solved using Firth's (1993) penalised MLE method and the exact logistic regression method. This study aims to develop an AD prediction model using Firth's (1993) logistic regression and identify potential biomarkers for AD classification using an integrated blood-based multi-omics dataset involving Malaysian patients.

METHODOLOGY

This study utilised transcriptomics, metabolomics, and cytokinomics datasets. The datasets were obtained from the "Towards Useful Ageing (TUA): Neuroprotective model for healthy longevity among the Malaysian elderly" research programme funded by the Long-term Research Grant Scheme (LRGS) of the Ministry of Higher Education, Malaysia. The study design for the data collection was matched case-control. The study population was elderly patients with AD enrolled at the Memory and Geriatric Clinic of the University of Malaya Medical Centre (UMMC), University of Malaya, Malaysia. The control group consisted of the elderly without AD.

The inclusion criteria for the multi-omics dataset were as follows; for the AD group, the age of patients was 65 years or more, who fulfilled the criteria of probable AD based on the Revised National Institute of Neurological and Communication Disorders—Alzheimer's disease and Related Disorder Association, and a neurologist or geriatrician made the diagnosis. The additional requirement for the AD group was that the mini-mental state examination (MMSE) score of the subject was less than or equal to 26 (Dayana et al., 2014; Hasni et al., 2016). Importantly too, patients' and/or caregivers' consent should be obtained.

As a start, the dataset was checked for outliers using Rosner's method (Rosner, 1975). Rosner's approach was chosen because it can detect several outliers in a sample dataset while also reducing false positives. Then, the multi-omics dataset is assessed with the separation test using linear programming developed by Konis (2007). Konis (2007) provided a solution to measure the separation between successes and failures in the binary response framework. A complete separation occurs when the parameter estimate of β diverges to ±∞ (Heinze & Schemper, 2002). Boxplot is used to illustrate the separation issues in the multi-omics dataset and validate it using the linear programming method (Appendix - Figure S2).

If there is a complete separation issue among the datasets, the common binary logistic regression is not able to fit the data because of the existence of monotone likelihood estimates. Firth's (1993) logistic regression (penalized ML estimation for logistic regression) was suggested as a solution for this situation (Firth, 1993; Heinze & Schemper, 2002; Kosmidis & Firth, 2010). Fundamentally, Firth's (1993) penalized method is used to extract a regular probability function with a bias term that is receptive to small sample size and rare targets (Rahman & Sultana, 2017). Firth (1993) used the penalty term in the ML-based score function (Equation 1) to remove first-order bias.

$$\frac{1}{2} trace\left[l(\beta)^{-1} \partial I\,(\beta)/\partial\,\beta_j\right] \qquad [1]$$

Firth's (1993) penalized likelihood is defined as Equation 2

$$l_{Firth}(\beta) = l(\beta) \times \left|I(\beta)\right|^{0.5} \quad [2]$$

where the log of the likelihood is defined as Equation 3

$$\log l_{Firth}(\beta) = \log\big(l(\beta)\big) + 0.5 * \log(|I(\beta)|) \quad [3]$$

where $I(\beta)$ denotes the Fisher information matrix. Next, Firth's (1993) penalized score function can be interpreted as Equation 4

$$U_{Firth}(\beta) = \frac{\partial}{\partial\beta}\left[\log\big(l(\beta)\big) + 0.5 * \log(|I(\beta)|)\right] \quad [4]$$

$$= \sum([(y_i - \pi_i + h_i \times (0.5 - \pi_i)] \times x_i)$$

where, h_i is the diagonal elements in Firth's (1993) likelihood structure of the predicted matrix *H*. The predicted matrix is defined as Equation 5

$$H = W^{0.5} \times X \times (X' \times W \times X)^{-1} \times X' \times W^{0.5} \quad [5]$$

where *W* is the diagonal matrix of $[\pi_i \times (1 - \pi_i)]$ and *X*, the regular design matrix. In Firth's (1993) penalized approach, proper estimating equations are defined to lead the estimator to become unbiased. This approach is useful with separated data where the first-order bias is removed and a small sample size. Besides, the approach promises the point estimates to be finite even in the monotone likelihood situation when ordinary ML estimation does not exist (Siino et al., 2018).

In the modelling phase, the univariable Firth's (1993) logistic regression was fitted on the multi-omics dataset to determine the individual biomarkers. This study has selected biomarkers with a p-value of less than 0.25 to fit into the multivariable Firth's (1993) logistic regression model (Hosmer Jr et al., 2013).

Variable selection using forward, backward elimination and stepwise selection were applied to get the best AD prediction model. If each method selected different biomarkers, this study would suspect ill-conditioning (or multicollinearity) among the biomarkers.

In the presence of multicollinearity, Principal Component Analysis (PCA) was performed to cluster the correlated biomarkers. Bartlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy would be calculated to determine if PCA is appropriate for the data. Then, PCA with varimax rotation was performed to obtain the principal components (PCs) from the multi-omics dataset. The rotation was made to facilitate the interpretation so that each biomarker is associated with a small block of observed biomarkers (Acal et al., 2020). The PCs were extracted based on the

eigenvalue of more than 1. Based on Jackson (1993), the eigenvalue of more than one would cluster or group the biomarkers into (PCs). Retaining 95% of the total variance, on the other hand, would not yield a promising result because there is a risk that many of the components retained will be noise or trivial components (Jackson, 1993). The scree plot was used to visualise the number of PCs extracted from the PCA model. The linear programming method was then employed to determine if there were complete separation issues for the PCs.

The multivariable Firth's (1993) logistic regression was fitted using the PCs extracted from the PCA with varimax rotation. The likelihood ratio test is performed to compare the individual PCs from the full model fitted using all principal components in the PCA with varimax rotation.

Before establishing the final model, diagnostic and influential statistic tests were carried out to check for outliers. The model goodness of fit tests was measured using the Hosmer and Lemeshow test. The classification table and area under the receiver operating characteristic (ROC) curve were used to evaluate the performance of the model. The model fitted the data when the p-value > 0.05 for the Hosmer-Lemeshow test, the classification rate of more than 70%, and the area under the ROC curve exceeded 0.80 (Hosmer Jr et al., 2013). On top of that, the influential statistics were checked using the Delta Chi-Square ($\Delta X^2{}_j$), Delta Deviance (ΔD_j), and Pregibon Delta Beta $\left(\Delta\hat{\beta}_j\right)$. The $\Delta X^2{}_j$ and ΔD_j were based on the ninety-fifth (95th) percentile of the distribution where under *m*-asymptotic, these quantities would be distributed approximately as $X^2_{(1)}$ with $X^2_{0.95(1)}$=3.84. The cut-off points to identify $\Delta X^2{}_j$ and ΔD_j were four. Whereas, for $\left(\Delta\hat{\beta}_j\right)$ larger than one, the case is considered an outlier. The flowchart of all the steps taken is in Appendix–Figure S1.

RESULTS AND DISCUSSION

The multi-omics data were obtained from a study of 32 AD patients and 16 non-AD subjects, and there were 16 transcript genes, 14 metabolites, and nine cytokines. Data cleaning for the dataset was performed by detecting and rectifying outliers and influencing data points using the Rosner method (Rosner, 1975). Then, the separation issue was resolved using the linear programming method developed by Konis (2007). Table 1 confirmed that ten biomarkers in the dataset had a complete separation issue since the intercept and coefficient were infinite. For metabolomics, the biomarkers that had separation issues were tryptophan, N-(2-hydroxyethyl) icosanamide, phytosphingosine, N-(2-hydroxyethyl) palmitamide, and methacholine, while for cytokinomics, they were interleukin-1β or IL-1β, IL-6, IL-10, IL-13, and human interferon-inducible protein 10 or IP-10. Fortunately, there were no biomarkers with complete separation issues from the transcriptomics group.

A univariable Firth's (1993) logistic regression was done to select potential biomarkers included in the model. Out of 39 biomarkers, only nine biomarkers, namely dihydroceramide

Table 1
Biomarkers with complete separation detected using the linear programming method (p=39)

Variable	Intercept	Coefficient	Variable	Intercept	Coefficient
Dihydroceramide C2	0	0	IFITM3	0	0
(Z)-N-(2-hydroxyethyl)icos-11-enamide	0	0	LY6G6D	0	0
Cholest-5-ene	0	0	MC1R	0	0
Tryptophan	$-\infty$	∞	MRPL18	0	0
N-(2-hydroxyethyl)icosanamide	$-\infty$	∞	SPOCD1	0	0
11,12-dihydroxy arachidic acid	0	0	ST14	0	0
3-hydroxylidocaine	0	0	TOR1AIP2	0	0
Phytosphingosine	$-\infty$	∞	TRIM16L	0	0
N-(2-hydroxyethyl)palmitamide	$-\infty$	∞	UBXN7	0	0
1-hexadecanoyl-sn-glycerol	0	0	VEGFB	0	0
20 alpha-dihydroprogesterone glucuronide	0	0	IL-1β	$-\infty$	∞
Methacholine	∞	$-\infty$	IL-6	$-\infty$	∞
2-oxo-docosanoic acid	0	0	IL-12	0	0
cis-11-hexadecenal	0	0	IFN-γ	0	0
ANKRD28	0	0	IL-10	∞	$-\infty$
CCDC92	0	0	IL-13	∞	$-\infty$
DEFA3	0	0	IP-10	$-\infty$	∞
FBXO32	0	0	MCP-1	0	0
GRIA4	0	0	MIP-1α	0	0
HDAC7	0	0		0	0

C2, 3-hydroxylidocaine, 20-alpha-dihydroprogesterone glucuronides, muscle atrophy F-box gene (FBXO32), histone deacetylase 7 (HDAC7), interferon-induced transmembrane protein 3 (IFITM3), melanocortin 1 receptor (MC1R), torsin 1A interacting protein 2 (TOR1AIP2), and vascular endothelial growth factor B (VEGFB) were not significant in the univariable model. The full univariable Firth's (1993) logistic regression is presented in Appendix - Table S1.

In the next stage of analysis, the variable selection procedure was applied to choose the significant transcriptomics, metabolomics, and cytokinomics biomarkers. Not all non-significant biomarkers were excluded in the subsequent analysis. Only biomarkers with a p-value > 0.25 were excluded from the variable selection procedure. The excluded biomarkers were 20 alpha-dihydroprogesterone glucuronides, HDAC7, IFITM3, MC1R, and TOR1AIP2. The selected biomarkers are presented in Table 2.

Firth's (1993) logistic regression with variable selection procedure (forward selection, backward elimination, and stepwise selection) selected tryptophan as the significant biomarker. The correlation among the biomarkers is then investigated, and the results

revealed a high correlation (multicollinearity) among the biomarkers within and between the three omics groups. The variance inflation factor (VIF) shows 17 biomarkers having multicollinearity issues since the VIF is more than ten and the tolerance value is less than 0.2. The biomarkers that have multicollinearity issues are presented in Table 3.

Thus, PCA with varimax rotation was carried out, and Bartlett's test was significant, indicating that the correlation matrix is not an identity matrix [Chi-square (df): 1940.19 (56)], p-value < 0.05. The KMO measure of sampling adequacy was 0.78 (greater than the threshold of 0.6) for PCA.

Table 2
List of 34 selected biomarkers for further analysis (n_{AD} = 32, $n_{non\text{-}AD}$ = 16)

List of selected biomarkers				
Dihydroceramide C2	11,12-dihydroxy arachidic acid	CCDC92	ST14	IL-12
Methacholine	N-(2-hydroxyethyl)palmitamide	DEFA3	TRIM16L	IFN-γ
Cholest-5-ene	N-(2-hydroxyethyl)icosanamide	FBXO32	UBXN7	IL-10
Tryptophan	1-hexadecanoyl-sn-glycerol	GRIA4	VEGFB	IL-13
cis-11-hexadecenal	2-oxo-docosanoic acid	LY6G6D	IL-1β	IP-10
Phytosphingosine	(Z)-N-(2-hydroxyethyl)icos-11-enamide	MRPL18	IL-6	MCP-1
3-hydroxylidocaine	ANKRD28	SPOCD1		MIP-1α

Table 3
List of biomarkers that have multicollinearity issues

Biomarker	VIF	Tolerance
Dihydroceramide C2	15.49	0.06
(Z)-N-(2-hydroxyethyl)icos-11-enamide	14.90	0.07
Tryptophan	517.23	< 0.01
N-(2-hydroxyethyl)icosanamide	968.76	< 0.01
11,12-dihydroxy arachidic acid	18.57	0.05
Phytosphingosine	169.25	0.01
N-(2-hydroxyethyl)palmitamide	232.26	< 0.01
1-hexadecanoyl-sn-glycerol	10.78	0.09
Methacholine	42.12	0.02
CCDC92	13.95	0.07
GRIA4	10.98	0.09
MRPL18	11.26	0.09
IL-1β	32.11	0.03
IL-6	13.66	0.07
IL-10	18.69	0.05
IL-13	23.33	0.04
MCP-1	13.34	0.07

The covariates (biomarkers) with a factor loading of 0.4 and higher (indicating satisfactory loading) were valid and significant contributors to the component. Based on the Scree plot in Figure 1, seven components had an eigenvalue > 1. It would mean that only seven PCs were extracted from the varimax rotation method, and the total variance explained by these components was 79.29%.

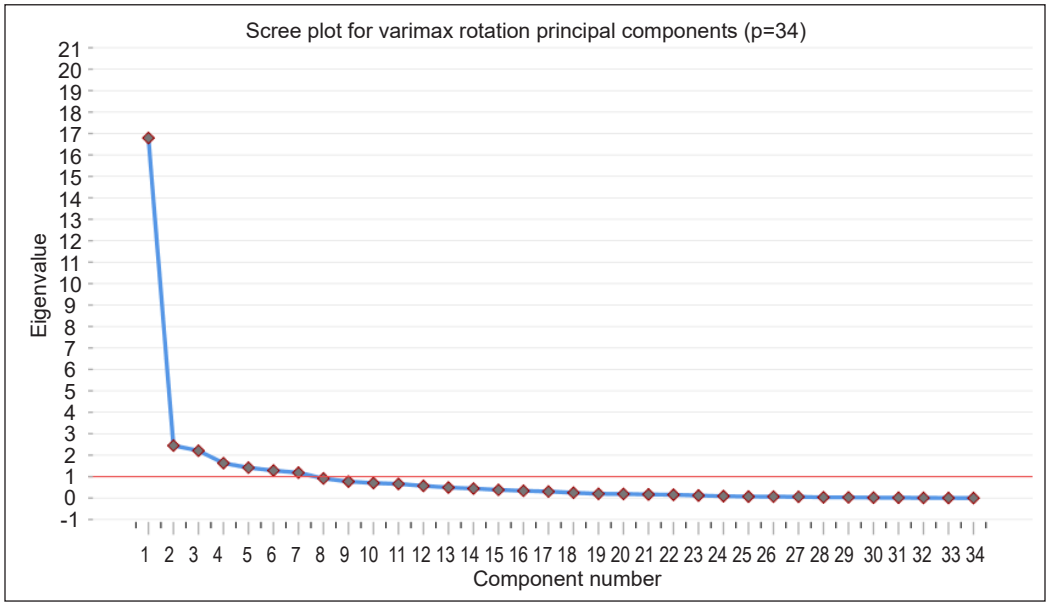


Figure 1. Scree plot for varimax rotation principal components

The first PC (PC1 with 36.89% variance explained) represents 18 biomarkers from metabolomics, transcriptomics, and cytokinomics. The biomarkers from PC1 were N-(2-hydroxyethyl) icosanamide, tryptophan, methacholine, IL-13, N-(2-hydroxyethyl) palmitamide, IL-1β, IL-6, IL-10, IP-10, phytosphingosine, 1-hexadecanoyl-sn-glycerol, monocyte chemoattractant protein 1 (MCP-1), suppressor of tumorigenicity-14 (ST14), IL-12, Macrophage inflammatory protein-1 alpha (MIP-1α), cis-11-hexadecenal, Interferon-gamma (IFN-γ), and Tripartite Motif Containing 16 Like (TRIM16L). There were four biomarkers in the second PC or PC2 (11.17% of variance explained), namely glutamate ionotropic receptor AMPA type subunit 4 (GRIA4), coiled-coil domain containing 92 (CCDC92), mitochondrial ribosomal protein L18 (MRPL18), and ultrabithorax domain-containing protein 7 (UBXN7).

The third PC (PC3 with 10.72% of variance explained) consisted of four biomarkers: VEGFB, cholest-5-ene, lymphocyte antigen six family member G6D (LY6G6D) and dihydroceramide C2. The two biomarkers for PC4 with 5.53% variance explained were Spen Paralogue and Orthologue C-terminal (SPOC) domain containing 1 (SPOCD1) and defensin alpha 3 (DEFA3).

PC5 with 5.298% variance explained also had two biomarkers, which were FBXO32 and ankyrin repeat domain 28 (ANKRD28). Only one biomarker, 2-oxo-docosanoic acid for PC6, was explained with a 4.936% variance. The last component, PC7 (4.75% of variance explained), consisted of two biomarkers, which were 3-hydroxylidocaine and 11,12-dihydroxy arachidic acid. The total variance explained by the seven components was 79.30%.

Before fitting Firth's (1993) logistic regression model, using rotated PCs scores, a test for complete separation was carried out for each of the seven principal components. Table 4 shows the separation detection using the linear programming method, and only PC1 had a separation issue since the intercept and coefficient were infinity. Thus, it was acceptable to fit Firth's (1993) logistic regression using the seven principal components.

The results for multivariable Firth's (1993) logistic regression model using seven principal components in Table 5 show that only PC1 was statistically significant [Wald statistic: 3.879, p-value < 0.001]. The adjusted odds ratio for PC1 was 10.65, which was the largest adjusted odds ratio compared to other PCs.

To ensure that any important PCs were not left out, the likelihood ratio test was done to examine the importance of each PC in the model. The likelihood ratio test compared the full model with 7 PCs and the model without the specific PCs to identify the crucial principal components of the full model. As shown in Table 6, only PC1 has a p-value < 0.05, meaning that PC1 is the most important component in Firth's (1993) logistic regression model for predicting AD.

Table 4
Separation detection of factor loadings using linear programming (p=7)

Variable	Intercept	Coefficient
PC1	∞	∞
PC2	0	0
PC3	0	0
PC4	0	0
PC5	0	0
PC6	0	0
PC7	0	0

Table 5
Multivariable Firth's (1993) *logistic regression of 7 PCs for Alzheimer's disease subject (n_{AD}=32, $n_{non\text{-}AD}$=16)*

Rotated Scores	$\widehat{\beta}$[a]	s.e.[b]	Wald Statistic[c]	Adjusted OR[d] (95% CI)[e]	p-value
PC1	2.3652	0.6098	3.88	10.65 (3.22, 35.18)	<0.001
PC2	0.7450	0.5881	1.27	2.11 (0.67, 6.67)	0.2052
PC3	0.6055	0.5785	1.05	1.83 (0.59, 5.69)	0.2953
PC4	0.3596	0.5666	0.64	1.43 (0.47, 4.35)	0.5257
PC5	0.2649	0.5754	0.46	1.30 (0.42, 4.03)	0.6451
PC6	0.3577	0.5633	0.64	1.43 (0.47, 4.31)	0.5254
PC7	-0.0355	0.5426	-0.07	0.97 (0.33, 2.79)	0.9478

[a]Regression coefficient; [b]Standard error; [c]z-value; [d]Adjusted odds ratio; [e]95% confidence interval constant = 0.4663

Table 6
Likelihood ratio test of each principal component

Full model without PCs	Chi-Square	p-value	Notes
PC1	46.98	<0.001	PC1 is important
PC2	0.81	0.3676	PC2 is not important
PC3	0.47	0.4949	PC3 is not important
PC4	0.83	0.3617	PC4 is not important
PC5	0.83	0.3629	PC5 is not important
PC6	0.80	0.3698	PC6 is not important
PC7	0.83	0.3634	PC7 is not important

Thus, the final model with PC1 represents 18 correlated biomarkers (7 metabolomics, 2 transcriptomics, and 9 cytokinomics) related to AD. The PC1 biomarkers include: (1) TRIM16L and ST14 from the transcriptomics dataset, (2) seven important metabolites from the metabolomics dataset were N-(2-hydroxyethyl) icosanamide, tryptophan, methacholine, N-(2-hydroxyethyl) palmitamide, phytosphingosine, 1-hexadecanoyl-sn-glycerol, and cis-11-hexadecenal, (3) IL-13, IL-1β, IL-6, IL-10, IP-10, MCP-1, IL-12, MIP-1α, and IFN-γ were the nine cytokinomics biomarkers.

The odds ratio for PC1 in Table 7 was 189.88 due to the effect of complete separation between the AD and non-AD groups. Furthermore, the large odds ratio may be due to the small sample size and unbalanced data. The odds ratio indicates that exposure to AD is higher for patients with these 18 biomarkers.

Table 7
Final model of Firth's (1993) *logistic regression* (n_{AD} = 32, $n_{non\text{-}AD}$ = 16)

Rotated Scores	$\widehat{\beta}$[a]	s.e.[b]	Wald Statistic[c]	Crude OR[d] (95% CI)[e]	p-value
PC1	5.246	1.686	3.111	189.88 (1.94, 8.55)	0.00186

[a]Regression coefficient [b]Standard error [c]z-value [d]Adjusted odds ratio [e]95% confidence interval
constant = 0.4663

The Hosmer-Lemeshow test [Chi-Square (df): 0.7439 (8), p-value > 0.05] indicates that the model fits the data. Since PC1 had complete separation of non-AD and AD, the classification rate, sensitivity, and specificity were 100%. The area under the ROC curve also indicated the perfect score of 1.0.

The diagnostic and influential statistics were conducted to examine the whole set of covariate patterns in the final model. Based on Hosmer Jr et al. (2013), the crude approximation to identify the outlier for delta Chi-Square ($\Delta X^2{}_j$) and delta deviance ΔD_j was based on the ninety-fifth (95th) percentile of the distribution; as under *m*-asymptotic, these quantities would be distributed approximately as $X^2_{(1)}$ with $X^2_{0.95(1)}$=3.84. Thus, the cut-off points to identify the outliers for delta chi-square and delta deviance was four.

Moreover, the influential diagnostic Pregibon delta beta $(\Delta\hat{\beta}_j)$ larger than one for an individual covariate pattern highlights that it is considered as an outlier. There was no influential statistic in PC1 since the values of $\Delta X^2{}_j$ versus $\hat{\pi}_j$ and ΔD_j versus $\hat{\pi}_j$ were lower than 4. Furthermore, the value of Pregibon delta beta $\Delta\hat{\beta}_j$ versus $\hat{\pi}_j$ was less than 1.0. Based on these values, there were no influential statistics in the model. Thus, the final Firth's (1993) logistic model was valid and appropriate for the data with complete separation issues.

Firth's (1993) logistic regression with variable selection procedures (forward selection, backward elimination, and stepwise selection) selected only one significant biomarker due to the presence of multicollinearity among biomarkers, and ten biomarkers were found to have complete separation issues. PCA was used to overcome the multicollinearity issue, while Firth's (1993) logistic regression was used as it is an appropriate model when complete separation occurs. PCA revealed seven principal components, and Firth's (1993) logistic regression revealed PC1 as the dominant component. PC1 also had a complete separation issue (indicating that it separated the AD and non-AD groups perfectly). A total of 18 important biomarkers were identified from the multi-omics dataset using Firth's (1993) logistic regression and PCA. There were seven metabolomics, two transcriptomics, and nine cytokinomics biomarkers in PC1. PC1 shows an upregulated risk in AD patients compared to non-AD subjects with these 18 biomarkers. To confirm the fitness of the model, a diagnostic and influential statistic was implemented. These potential multi-omics biomarkers are summarised in Table 8.

The potential biomarkers from transcriptomics were ST14 (Suppression of tumorigenicity) and TRIM16L (tripartite motif-containing 16 like). ST14 was found to be upregulated in this study. A similar finding indicating ST14 as an important upregulated transcriptomics biomarker of AD was found in many studies (Nazarian et al., 2020; Rousseaux et al., 2012; Yin et al., 2017).

Table 8
Summary of potential multi-omics biomarkers of AD

Transcriptomics	Metabolomics	Cytokinomics
ST14	n-(2-hydroxyethyl)icosanamide	MCP-1
TRIM16L	tryptophan	IL-1β
	methacholine	IL-13
	n-(2-hydroxyethyl)palmitamide	IL-6
	phytosphingosine	IL-10
	cis-11-hexadecenal	IP-10
	1-hexadecanoyl-sn-glycerol	IL-12
		MIP-1α
		IFN-γ

The metabolomics biomarker, N-(2-hydroxyethyl) icosanamide, was found to be upregulated in this study. Brand et al. (2015) reported that the N-(2-hydroxyethyl) icosanamide protects from neuronal death and is also involved with the inflammatory immune response. When the intensity of n-(2-hydroxyethyl) icosanamide increases, it would be an indicator for a person to develop AD potentially.

Methacholine was also an important biomarker of AD and had an upregulated effect on AD in this study. In contrast, previous studies reported methacholine as a downregulated gene, where every increment of methacholine intensity would decrease the risk of getting AD (Bavarsad et al., 2020; Jang et al., 2020). More data need to be obtained to verify these contradictory findings on methacholine.

As for N-(2-hydroxyethyl) palmitamide, D'Agostino et al. (2012) and Kuehl et al. (1957) reported similar findings that the metabolite is a potential biomarker of AD. Phytosphingosine is an upregulated metabolite biomarker of AD (Li et al., 2018; Sun et al., 2018; Li et al., 2010). The cis-11-hexadecenal metabolite was also reported by Berdyshev (2011) and was related to lipidomic disease (Kocak, 2020). Finally, the metabolomics biomarker, 1-hexadecanoyl-sn-glycerol was found to be upregulated for AD in this study. Currently, no studies have found 1-hexadecanoyl-sn-glycerol metabolite as a metabolomics biomarker for AD.

In the cytokinomics group, IL-1β, IL-6, IL-13, and MIP-1α were identified as important cytokines biomarkers. Yin et al. (2016) reported that IL-1β and homozygous APOE4 combined were associated with an increased hazard of developing AD. Furthermore, IL-1β was also reported with six accompanying pathways that linked it to AD, which are tumour necrosis factor (TNF-α), TGF-β, c-June N-terminal kinase (JNK), extracellular-signal-regulated kinase (ERK), LPS, and nerve growth factor (NGF) (Xie et al., 2015). It was also reported that the levels of IL-6 and IFN-γ were significantly higher in altered T-lymphocytes of AD patients compared to the non-AD group (Azad et al., 2014). In this study, IFN-γ was found to have a significant relationship with AD. In addition, some studies have reported that the increment of IL-6 would influence the progression of the cognitive decline in AD (Licastro et al., 2003; Mrak & Griffin, 2005). IL-10 and IL-13 were said to be anti-inflammatory cytokines by their ability to suppress genes for pro-inflammatory cytokines (Rubio-Perez & Morillas-Ruiz, 2012). These results were in line with Dayana et al. (2014) and Hasni et al. (2016).

The interferon gamma-induced protein 10 (IP-10) or C-X-C motif chemokine 10 (CXCL-10) indicated an upregulated or elevated risk of AD. A study by Minter et al. (2016) supported that CXCL-10 was positively correlated with the severity of the cognitive decline in AD patients. Furthermore, in an animal study, CXCL-10 was implicated in the disease progression of APPSWE/PS1ΔE9 mice where deletion of the gene ameliorated amyloidosis and cognitive decline (Minter et al., 2016).

CONCLUSION

In the presence of complete separation, the maximum likelihood estimation method in logistic regression will provide an infinite estimate of the covariate coefficient. Firth's (1993) logistic regression uses a penalized likelihood estimation method and is the appropriate solution to the separation issue for logistic regression. The important biomarkers identified from the multi-omics dataset showed a strong correlation among transcripts, metabolites, and cytokine biomarkers. This study supported past findings that applied an integrative multi-omics approach to establish significant AD-associated biomarkers. Multi-omics studies may have an important role in developing the diagnosis and treatment of AD. Future research can explore machine learning approaches for the identification of biomarkers.

The novelty of the current work is developing a solution on how to deal with a dataset with multicollinearity among predictors and complete separation issues. The ensemble method of Principal Component and Firth Logistic Regression would ultimately contribute to the theory and practice when facing both situations simultaneously in the dataset. Until now, there has been no study published that deals with these two situations together.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Teknologi MARA (UiTM) Perak, Tapah Campus and UiTM Shah Alam for conducting this study. University research funding supported this research under the BESTARI Grant (600-IRMI/DANA 5/3/BESTARI (113/2018)). They also acknowledge the research team of the LRGS project: "Towards Useful Ageing (TUA): Neuroprotective model for healthy longevity among the Malaysian elderly"—Ministry of Higher Education (MOHE) (600-RMI/LRGS 5/3 [3/2012]) for their data and support.

REFERENCES

Acal, C., Aguilera, A. M., & Escabias, M. (2020). New modeling approaches based on varimax rotation of functional principal components. *Mathematics*, *8*(11), 1-15. https://doi.org/10.3390/math8112085

Adnan, N., Ahmad, M. H., & Adnan, R. (2006). A comparative study on some methods for handling multicollinearity problems. *Matematika*, *22*(2), 109-119.

Azad, F. J., Talaei, A., Rafatpanah, H., & Yousefzadeh, H. (2014). Association between cytokine production and disease severity in Alzheimer's disease. *Iranian Journal of Allergy, Asthma & Immunology*, *13*(6), 433-439.

Bavarsad, K., Saadat, S., Roshan, N. M., Hadjzadeh, M. A. R., & Boskabady, M. H. (2020). Effects of levothyroxine on lung inflammation, oxidative stress and pathology in a rat model of Alzheimer's disease. *Respiratory Physiology and Neurobiology*, *277*, Article 103437. https://doi.org/10.1016/j.resp.2020.103437

Berdyshev, E. V. (2011). Mass spectrometry of fatty aldehydes. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, *1811*(11), 680-693. https://doi.org/10.1016/j.bbalip.2011.08.018

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. *Nature Genetics*, *39*(1), 17-23. https://doi.org/10.1038/ng1934

Brand, B., Hadlich, F., Brandt, B., Schauer, N., Graunke, K. L., Langbein, J., Repsilber, D., Ponsuksili, S., & Schwerin, M. (2015). Temperament type specific metabolite profiles of the prefrontal cortex and serum in cattle. *PloS One*, *10*(4), Article e0125044. https://doi.org/10.1371/journal.pone.0125044

Cayton, H., Graham, N., & Warner, J. (2008). *Alzheimer's and other dementias*. Class Publishing.

Clark, C., Dayon, L., Masoodi, M., Bowman, G. L., & Popp, J. (2021). An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alzheimer's Research and Therapy*, *13*(1), 1-19. https://doi.org/10.1186/s13195-021-00814-7

Clerici, M. (2010). Beyond IL-17: New cytokines in the pathogenesis of HIV infection. *Current Opinion in HIV and AIDS*, *5*(2), 184-188. https://doi.org/10.1097/COH.0b013e328335c23c

Cummings, J. L., & Jeste, D. V. (1999). Alzheimer's disease and its management in the year 2010. *Psychiatric Services*, *50*(9), 1173-1177. http://www.ncbi.nlm.nih.gov/pubmed/10478903

D'Agostino, G., Russo, R., Avagliano, C., Cristiano, C., Meli, R., & Calignano, A. (2012). Palmitoylethanolamide protects against the amyloid-B25-35-induced learning and memory impairment in mice, an experimental model of alzheimer disease. *Neuropsychopharmacology*, *37*(7), 1784-1792. https://doi.org/10.1038/npp.2012.25

Dayana, S. M. H., Lim, S. M., Tan, M. P., Chin, A. V, Poi, P. J. H., Kamaruzzaman, S. B., Majeed, A. B. A., & Ramasamy, K. (2014). IP-10 and IL-13 as potentially new, non-classical blood-based cytokine biomarker for Alzheimer's disease. *Neuorology and Neurosciences*, *43*(April), Article 115. https://doi.org/10.1093/ageing/afu045.2

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*(1), 27-38.

Gomez-Ramirez, J., & Wu, J. (2014). Network-based biomarkers in Alzheimer's disease: Review and future directions. *Frontiers in Aging Neuroscience*, *6*, Article 12. https://doi.org/10.3389/fnagi.2014.00012

Gross, A. L., Jones, R. N., Habtemariam, D. A., Fong, T. G., Tommet, D., Quach, L., Schmitt, E., Yap, L., & Inouye, S. K. (2012). Delirium and long-term cognitive trajectory among persons with dementia. *Archives of Internal Medicine*, *172*(17), 1324-1331. https://doi.org/10.1001/archinternmed.2012.3203

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 1-15. https://doi.org/10.1186/s13059-017-1215-1

Hasni, D. S. M., Lim, S. M., Chin, A. V., Tan, M. P., Poi, P. J. H., Kamaruzzaman, S. B., Majeed, A. B. A., & Ramasamy, K. (2016). Peripheral cytokines, C-X-C motif ligand10 and interleukin-13, are associated with Malaysian Alzheimer's disease. *Geriatrics and Gerontology International, 17*(5), 839-846. https://doi.org/10.1111/ggi.12783

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409-2419. https://doi.org/10.1002/sim.1047

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Humpel, C. (2011). Identifying and validating biomarkers for Alzheimer's disease. *Trends Biotechnol*, *29*(1), 26-32. https://doi.org/10.1016/j.tibtech.2010.09.007

Ibáñez, C., Simó, C., & Cifuentes, A. (2013). Metabolomics in Alzheimer's disease research. *Electrophoresis*, *34*(19), 2799-2811. https://doi.org/10.1002/elps.201200694

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecological Society of America*, *74*(8), 2204-2214.

Jang, H., Kim, M., Hong, J. Y., Cho, H. J., Kim, C. H., Kim, Y. H., Sohn, M. H., & Kim, K. W. (2020). Mitochondrial and nuclear mitochondrial variants in allergic diseases. *Allergy, Asthma and Immunology Research*, *12*(5), 877-884. https://doi.org/10.4168/aair.2020.12.5.877

Kocak, E. (2020). Evaluation of ms-dial and mzmine2 softwares for clinical lipidomics analysis. *Communications Faculty of Sciences University of Ankara Series*, *62*(1), 100-114.

Konis, K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models* (PhD Thesis). University of Oxford, UK.

Kosmidis, I., & Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, *4*, 1097-1112. https://doi.org/10.1214/10-EJS579

Kuehl Jr, F. A, Jacob, T. A, Galey, O. H., Ormond, R. E., & Meisinger, M. A. P. (1957). The identification of N-(2-hydroxyethyl)-palmitamide as a naturally occuring anti-inflammatory agent. *Journal of the American Oil Chemists' Society*, *79*(8), 5577-5578.

Kussmann, M., Raymond, F., & Affolter, M. (2006). OMICS-driven biomarker discovery in nutrition and health. *Journal of Biotechnology*, *124*(4), 758-787. https://doi.org/10.1016/j.jbiotec.2006.02.014

Li, J., Liu, Y., Li, W., Wang, Z., Guo, P., Li, L., & Li, N. (2018). Metabolic profiling of the effects of ginsenoside Re in an Alzheimer's disease mouse model. *Behavioural Brain Research*, *337*(April 2017), 160-172. https://doi.org/10.1016/j.bbr.2017.09.027

Li, N. J., Liu, W. T., Li, W., Li, S. Q., Chen, X. H., Bi, K. S., & He, P. (2010). Plasma metabolic profiling of Alzheimer's disease by liquid chromatography/mass spectrometry. *Clinical Biochemistry*, *43*(12), 992-997. https://doi.org/10.1016/j.clinbiochem.2010.04.072

Licastro, F., Grimaldi, L. M. E., Bonafè, M., Martina, C., Olivieri, F., Cavallone, L., Giovanietti, S., Masliah, E., & Franceschi, C. (2003). Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain. *Neurobiology of Aging*, *24*(7), 921-926. https://doi.org/10.1016/S0197-4580(03)00013-7

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., Davies, G., Ritchie, C. W., Gale, C. R., Starr, J. M., Goate, A. M., Porteous, D. J., Yang, J., Evans, K. L., Deary, I. J., Wray, N. R., & Visscher, P. M. (2018). GWAS on family history of Alzheimer's disease. *Translational Psychiatry*, *8*(1), 0-6. https://doi.org/10.1038/s41398-018-0150-6

Maskery, M., Goulding, E. M., Gengler, S., Melchiorsen, J. U., & Rosenkilde, M. M. (2020). The dual GLP-1 / GIP receptor agonist DA4-JC shows superior protective properties compared to the GLP-1 analogue liraglutide in the APP / PS1 mouse model of Alzheimer's disease. *American Journal of Alzheimer's Disease & Other Dementias*, *35*, 1-11. https://doi.org/10.1177/1533317520953041

Minter, M. R., Taylor, J. M., & Crack, P. J. (2016). The contribution of neuroinflammation to amyloid toxicity in Alzheimer's disease. *Journal of Neurochemistry*, *136*(3), 457-474. https://doi.org/10.1111/jnc.13411

Mrak, R. E., & Griffin, W. S. T. (2005). Potential inflammatory biomarkers in Alzheimer's disease. *Journal of Alzheimer's Disease: JAD*, *8*(4), 369-375.

Nazarian, A., Yashin, A. I., & Kulminski, A. M. (2020). Summary-based methylome-wide association analyses suggest potential genetically driven epigenetic heterogeneity of Alzheimer's disease. *Journal of Clinical Medicine*, *9*(5), Article 1489. https://doi.org/10.3390/jcm9051489

Park, J. C., Han, S. H., & Mook-Jung, I. (2020). Peripheral inflammatory biomarkers in Alzheimer's disease: A brief review. *BMB Reports*, *53*(1), 10-19. https://doi.org/10.5483/BMBRep.2020.53.1.309

Rahayu, S., Sugiarto, T., Madu, L., Holiawati, & Subagyo, A. (2017). Application of principal component analysis (PCA) to reduce multicollinearity exchange rate currency of some countries in Asia period 2004-2014. *International Journal of Educational Methodology*, *3*(2), 75-83. https://doi.org/10.12973/ijem.3.2.75

Rahman, M. S., & Sultana, M. (2017). Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. *BMC Medical Research Methodology*, *17*(1), 1-15. https://doi.org/10.1186/s12874-017-0313-9

Romero, R., Espinoza, J., Gotsch, F., Kusanovic, J. P., Friel, L. A., Erez, O., Mazaki-Tovi, S., Than, N. G., Hassan, S., & Tromp, G. (2006). The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. *BJOG: An International Journal of Obstetrics & Gynaecology*, *113*(s3), 118-135.

Rosner, B. (1975). On the detection of many outliers. *Technometrics, 17*(2), 221-227. https://doi.org/10.1080/00401706.1975.10489305

Rougoor, C. W., Sundaram, R., & Van Arendonk, J. A. M. (2000). The relation between breeding management and 305-day milk production, determined via principal components regression and partial least squares. *Livestock Production Science*, *66*(1), 71-83. https://doi.org/10.1016/S0301-6226(00)00156-1

Rousseaux, M., Rénier, J., Anicet, L., Pasquier, F., & Mackowiak-Cordoliani, M. A. (2012). Gesture comprehension, knowledge and production in Alzheimer's disease. *European Journal of Neurology*, *19*(7), 1037-1044. https://doi.org/10.1111/j.1468-1331.2012.03674.x

Rubio-Perez, J. M., & Morillas-Ruiz, J. M. (2012). A review: Inflammatory process in Alzheimer's disease, role of cytokines. *The Scientific World Journal*, *2012*, Article 756357. https://doi.org/10.1100/2012/756357

Siino, M., Fasola, S., & Muggeo, V. M. R. (2018). Inferential tools in penalized logistic regression for small and sparse data: A comparative study. *Statistical Methods in Medical Research*, *27*(5), 1365-1375. https://doi.org/10.1177/0962280216661213

Sun, L. M., Zhu, B. J., Cao, H. T., Zhang, X. Y., Zhang, Q. C., Xin, G. Z., Pan, L. M., Liu, L. F., & Zhu, H. X. (2018). Explore the effects of Huang-Lian-Jie-Du-Tang on Alzheimer's disease by UPLC-QTOF/MS-based plasma metabolomics study. *Journal of Pharmaceutical and Biomedical Analysis*, *151*, 75-83. https://doi.org/10.1016/j.jpba.2017.12.053

Swardfager, W., Lanctot, K., Rothenburg, L., Wong, A., Cappell, J., & Herrmann, N. (2010). A meta-analysis of cytokines in Alzheimer's disease. *Biol Psychiatry*, *68*(10), 930-941. https://doi.org/10.1016/j.biopsych.2010.06.012

Tanzi, R. E. (2012). The genetics of Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(10), Article a006296. https://doi.org/10.1101/cshperspect.a006296

Von Schulze, A. T., Deng, F., Morris, J. K., & Geiger, P. C. (2020). Heat therapy: Possible benefits for cognitive function and the aging brain. *Journal of Applied Physiology, 129*(6), 1468-1476. https://doi.org/10.1152/japplphysiol.00168.2020

Waring, S. C., & Rosenberg, R. N. (2008). Genome-wide association studies in Alzheimer disease. *Archives of Neurology*, *65*(3), 329-334. https://doi.org/10.1001/archneur.65.3.329

Xie, L., Lai, Y., Lei, F., Liu, S., Liu, R., & Wang, T. (2015). Exploring the association between interleukin-1beta and its interacting proteins in Alzheimer's disease. *Molecular Medicine Reports, 11*(5), 3219-3228. https://doi.org/10.3892/mmr.2015.3183

Yin, Y., Liu, Y., Pan, X., Chen, R., Li, P., Wu, H. J., Zhao, Z. Q., Li, Y. P., Huang, L. Q., Zhuang, J. H., & Zhao, Z. X. (2016). Interleukin-1β Promoter polymorphism enhances the risk of sleep disturbance in Alzheimer's disease. *PLoS One*, *11*(3), 1-13. https://doi.org/10.1371/journal.pone.0149945

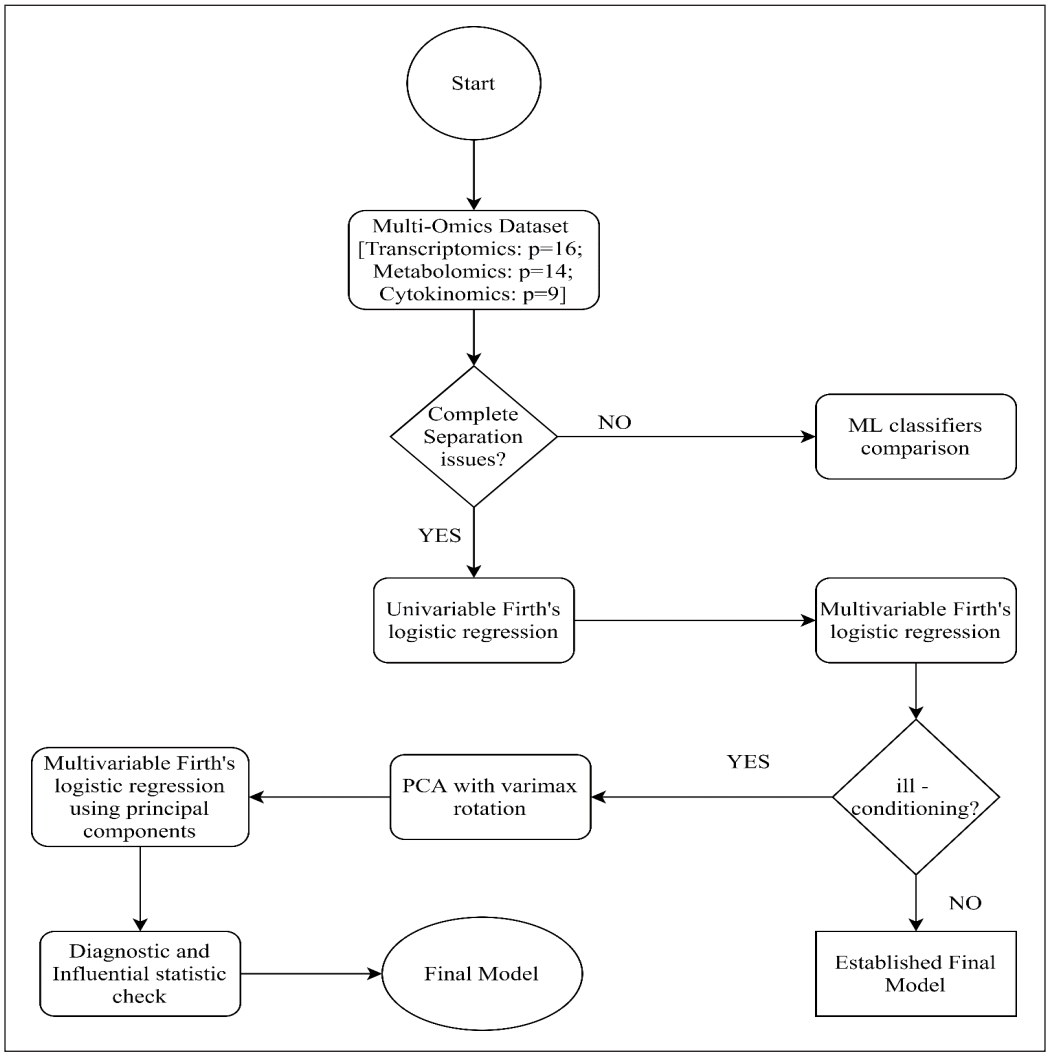
Yin, Z., Raj, D., Saiepour, N., Van Dam, D., Brouwer, N., Holtman, I. R., Eggen, B. J. L., Möller, T., Tamm, J. A., Abdourahman, A., Hol, E. M., Kamphuis, W., Bayer, T. A., De Deyn, P. P., & Boddeke, E. (2017). Immune hyperreactivity of Aβ plaque-associated microglia in Alzheimer's disease. *Neurobiology of Aging*, *55*, 115-122. https://doi.org/10.1016/j.neurobiolaging.2017.03.021

Zhang, X. (2011). *Omics technologies in cancer biomarker discovery*. CRC Press.

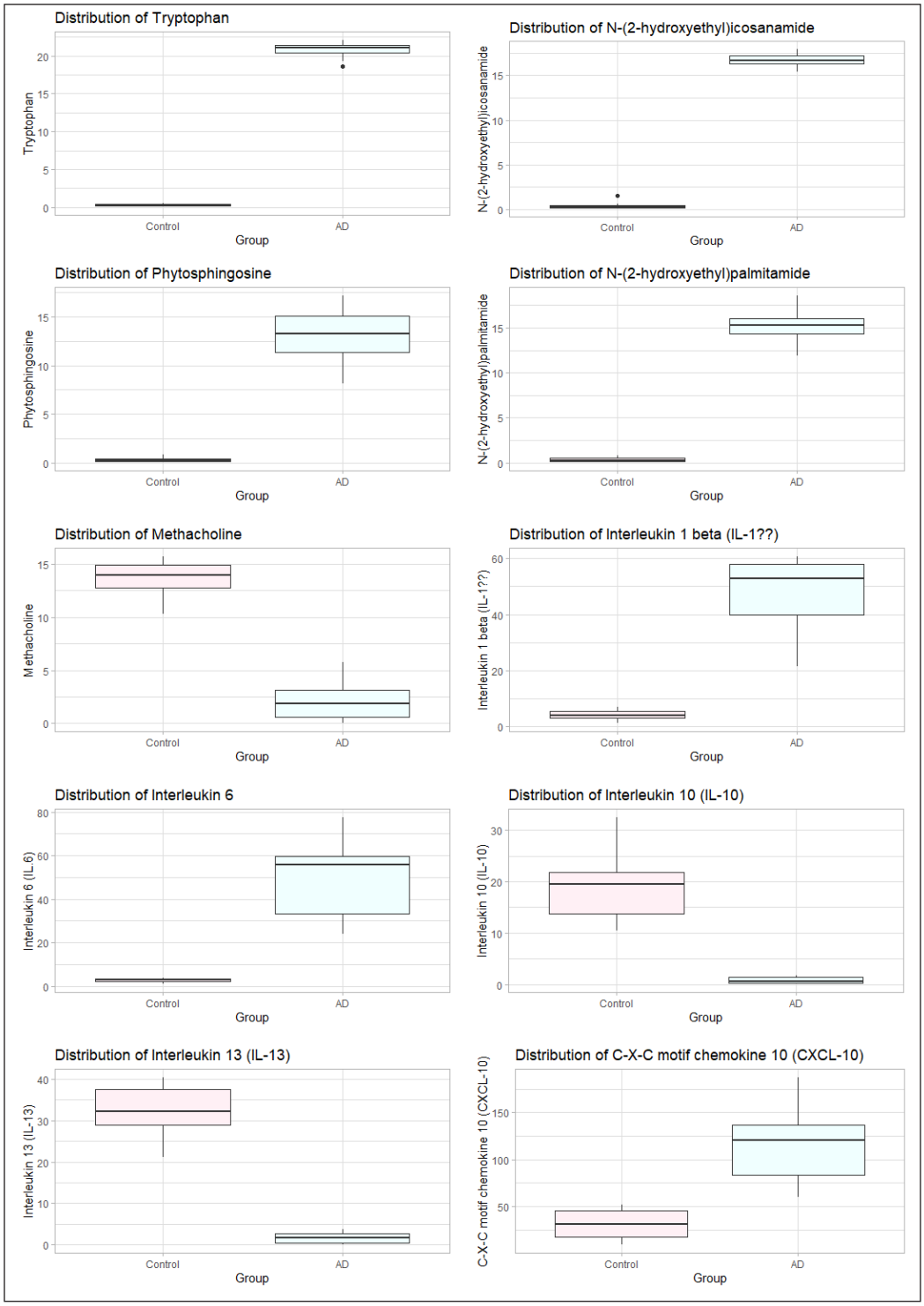
Zheng, C., Zhou, X. W., & Wang, J. Z. (2016). The dual roles of cytokines in Alzheimer's disease: Update on interleukins, TNF-α, TGF-β and IFN-γ. *Translational Neurodegeneration*, *5*(1), 1-15. https://doi.org/10.1186/s40035-016-0054-4

Zhou, J., Zhu, Z., & Ji, Z. (2014). A Memetic algorithm based feature weighting for metabolomics data classification. *Chinese Journal of Electronics*, *23*(4), 706-711.

APPENDIX



Supplementary Figure 1. Flow of modelling multi-omics dataset on Alzheimer's disease



Supplementary Figure 2. Boxplot of ten complete separation biomarkers

Supplementary Table 1
Univariable Firth's (1993) *logistic regression in measuring potential biomarkers association with AD* (n_{AD} = 32, $n_{non\text{-}AD}$ = 16)

Biomarkers	AD [mean (sd)]	Non-AD [mean (sd)]	Crude OR[a] (AIC)[b]	(95% CI)[c]	p-value[d]
Dihydroceramide C2	9.23 (5.86)	0.33 (0.24)	3.91(28.86)	(0.86,17.78)	0.0779
(Z)-N-(2-hydroxyethyl)icos-11-enamide	6.06 (3.88)	0.31 (0.21)	51.37(23.97)	(1.21,2178.64)	0.0394
Cholest-5-ene	3.94 (4.12)	0.23 (0.11)	30.63(39.39)	(1.17,803.11)	0.0401
Tryptophan	20.84 (0.85)	0.28 (0.14)	1.46(5.94)	(1.19,1.78)	0.0002
N-(2-hydroxyethyl) icosanamide	16.68 (0.66)	0.36 (0.35)	1.61(5.94)	(1.25,2.07)	0.0002
11,12-dihydroxy arachidic acid	5.50 (4.99)	0.27 (0.18)	185.51(30.29)	(2.85,12093.79)	0.0143
3-hydroxylidocaine	0.37 (0.27)	0.27 (0.18)	5.65(63.05)	(0.4,79.98)	0.2002
Phytosphingosine	13.14 (2.47)	0.32 (0.22)	1.99(6.09)	(1.33,2.98)	0.0009
N-(2-hydroxyethyl) palmitamide	15.28 (1.49)	0.35 (0.23)	1.7(5.98)	(1.27,2.27)	0.0003
1-hexadecanoyl-sn-glycerol	17.48 (0.72)	5.63 (5.05)	2.3(14.91)	(1.26,4.2)	0.0068
20 alpha-dihydroprogesterone glucuronide	0.57 (0.49)	0.48 (0.31)	1.56(64.64)	(0.38,6.38)	0.5367
Methacholine	1.96 (1.46)	13.65 (1.64)	0.47(6.21)	(0.3,0.74)	0.0011
2-oxo-docosanoic acid	1.22 (1.01)	0.27 (0.18)	10.73(48.1)	(1.54,74.69)	0.0165
cis-11-hexadecenal	9.05 (5.50)	0.33 (0.24)	1.68(32.86)	(1.15,2.46)	0.0072
ANKRD28	0.16 (0.54)	-0.29 (0.46)	4.67(57.71)	(1.28,17.03)	0.0197
CCDC92	-0.01 (0.75)	-0.89 (0.75)	3.65(53.01)	(1.55,8.61)	0.0031
DEFA3	-1.19 (1.55)	1.00 (1.57)	0.41(46.97)	(0.23,0.73)	0.0022
FBXO32	0.25 (0.45)	-0.09 (0.58)	3.76(60.34)	(0.97,14.62)	0.0563
GRIA4	-0.34 (0.82)	0.89 (0.71)	0.19(45.33)	(0.07,0.49)	0.0007
HDAC7	-0.04 (0.43)	0.10 (0.30)	0.39(63.52)	(0.07,2.12)	0.2740
IFITM3	0.28 (0.67)	0.08 (0.62)	1.56(64.11)	(0.6,4.05)	0.3635
LY6G6D	-0.41 (1.08)	0.51 (1.00)	0.45(57.43)	(0.23,0.89)	0.0211
MC1R	0.07 (0.74)	-0.04 (0.42)	1.28(64.78)	(0.49,3.33)	0.6160
MRPL18	0.13 (0.66)	-0.92 (0.81)	5.47(47.2)	(2.04,14.65)	0.0007
SPOCD1	0.27 (1.06)	-0.69 (1.33)	1.91(58.35)	(1.1,3.33)	0.0225
ST14	-0.39 (0.34)	1.43 (1.15)	0.09(30.98)	(0.02,0.45)	0.0033
TOR1AIP2	-0.16 (0.63)	0.02 (0.19)	0.52(63.73)	(0.16,1.74)	0.2913
TRIM16L	0.21 (0.59)	-0.33 (0.16)	24.93(50.33)	(2.47,251.47)	0.0064
UBXN7	-0.09 (0.67)	0.35 (0.41)	0.29(59.27)	(0.09,0.93)	0.0378
VEGFB	-0.11 (0.52)	0.23 (0.48)	0.29(60.54)	(0.08,1.03)	0.0556

Supplementary Table 1 *(continue)*

Biomarkers	AD [mean (sd)]	Non-AD [mean (sd)]	Crude OR[a] (AIC)[b]	(95% CI)[c]	p-value[d]
IL-1β	48.19 (12.15)	4.04 (1.88)	1.34(6.05)	(1.07,1.68)	0.0110
IL-6	49.04 (16.47)	2.71 (0.85)	1.3(5.97)	(1.1,1.55)	0.0027
IL-12	30.08 (17.27)	6.89 (4.94)	1.3(32.18)	(1.1,1.53)	0.0023
IFN-γ	1.49 (1.14)	0.15 (0.06)	380322206.99 (24.82)	(89.98, 1607462564390182)	0.0111
IL-10	0.81 (0.55)	18.91 (5.95)	0.56(6.02)	(0.39,0.81)	0.0019
IL-13	1.66 (1.28)	32.01 (6.28)	0.74(6.04)	(0.62,0.89)	0.0013
IP-10	115.49 (36.62)	31.26 (14.34)	1.25(6.86)	(1.03,1.52)	0.0259
MCP-1	10.72 (4.09)	3.92 (2.47)	1.97(32.45)	(1.31,2.97)	0.0011
MIP-1α	0.55 (0.38)	0.17 (0.07)	648.09(45.97)	(7.03,59748.96)	0.0050

[a]Crude odds ratio (OR) was calculated based on exponential coefficient of each biomarkers.
[b]A lower value of Akaike Information criteria (AIC) is preferred as it indicates the model fits better.
[c]The variable is considered significant if the 95% confidence interval (95% CI for odds ratio) does not include 1 in the interval range.
[d]Simple Firth's (1993) logistic regression was done for all individual biomarkers.