# Discovering potential blood-based cytokine biomarkers for Alzheimer's disease using Firth Logistic Regression

*Mohammad Nasir Abdullah [1], Yap Bee Wah [2], Yuslina Zakaria [3], Abu Bakar Abdul Majeed [3], Ong Seng Huat [4]*

*(1) Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 35400 Tapah, Perak, Malaysia*
*(2) Advanced Analytics Engineering Centre, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*
*(3) Faculty of Pharmacy, Universiti Teknologi MARA, 42300 Puncak Alam, Selangor, Malaysia*
*(4) Department of Actuarial Science and Applied Statistics, UCSI University, 56000 Kuala Lumpur, Malaysia*

***CORRESPONDING AUTHOR:*** *Mohammad Nasir Abdullah, Department of Statistics, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 35400 Tapah, Perak, Malaysia. Email: nasir916@uitm.edu.my*

***DOI:* 10.2427/13173**
*Accepted on October 11, 2019*

## ABSTRACT

**Background:** Alzheimer's disease (AD) is a neurodegenerative disorder where patients suffer from memory loss, cognitive impairment and progressive disability. Individual blood biomarkers have not been successful in defining the disease pathology, progression and diagnosis of AD. There is a need to identify multiplex panels of blood biomarkers for early diagnosis of AD with high sensitivity and specificity. This study focused on identification of cytokine biomarkers. The maximum likelihood estimates of the ordinary logistic regression model cannot be obtained when there is complete separation and the alternative is Firth logistic regression which uses a penalised Maximum Likelihood in parameter estimation.
**Methods:** This paper reports a Firth logistic regression application in finding potential blood-based cytokine biomarkers for Alzheimer's disease in a matched case control study. We used a principle component analysis to discriminate the correlated, completely separated covariates.
**Results:** The Firth logistic regression results showed that nine individual biomarkers IL-1$\beta$, IL-6, IL-12, IFN-$\gamma$, IL-10, IL-13, IP-10, MCP-1 and MIP-1$\alpha$ had a significant relationshipwith elevated risk for AD as compared to the healthy control (HC). Principal component analysis with varimax rotation for the nine biomarkers revealed four factors (total variance explained=85.5%). The main principal component biomarkers were IL-1$\beta$, IL-6, IL-13 and MCP-1 (total variance explained=62.3%). Firth's logistic regression model with the first principal component had accuracy of 78.2% with sensitivity and specificity of 71.8% and 75% respectively.
**Conclusion:** Firth's logistic regression is a useful technique in identification of significant biomarkers when there is an issue of data separation.

*Key words: Alzheimer's disease; Cytokine; Biomarkers; Firth Logistic Regression, Penalised Maximum Likelihood*

## INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder characterised by the gradual progression of memory loss, impairment of cognitive functions and progressive disability that accounts for 60% to 80% of all types of dementia [1–3]which is thought to be a powerful strategy to examine the influence of genetic variants (i.e., single nucleotide polymorphisms (SNPs. The disease is also commonly characterised by the development of amyloid-beta (Aβ) plaques and hyper-phosphorylated tau neurofibrillary tangles that leads to neuronal death or apoptosis and memory decline [4]. AD can incur tremendous social and economic costs, to the sufferer as well as the caregiver and family members. Genes associated with the development of AD might have an effect on chemical neurotransmitters, which allow message to be communicated between nerve cells in the brain [5].

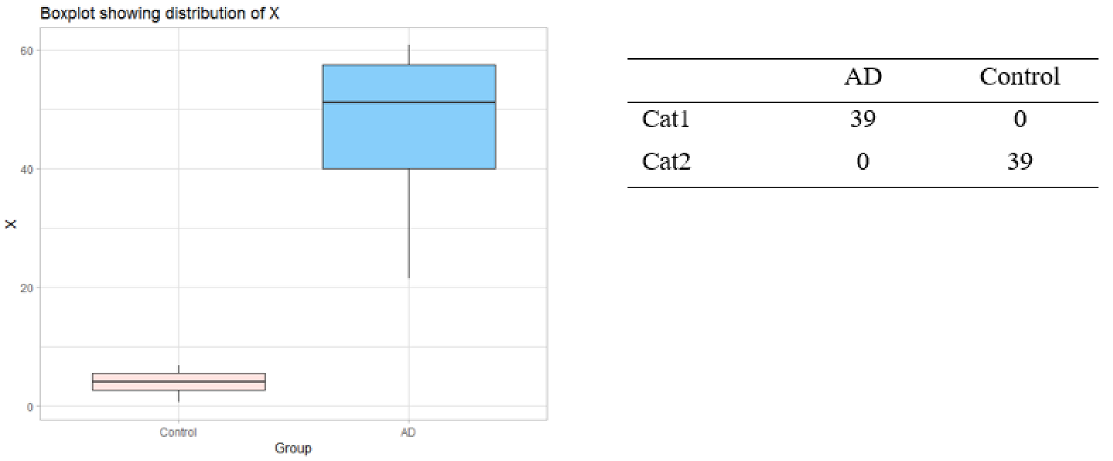
The prevalence of AD was estimated to be 47 million cases worldwide according to World Alzheimer Report 2016 [6]. The number is expected to increase to 131.5 million by 2050, affecting mostly low and middle income countries [7]. In Malaysia, AD cases are believed to be under-reported because most family members view its symptoms as normal aging and hence they do not seek any medical treatments. Cytokines are dissolved proteins or glycoproteins produced by the leukocytes. They act as chemical communicators between cells in a way similar to hormones but with the strongest activity in the microenvironment of the cells that they are contained within [8,9]. Cytokines are involved in both healthy biological processes such as cell growth, differentiation, inflammation, immunity, repair and fibrosis as well as pathological processes [10]. The word cytokine comes from Greek where cyto means cell and kinos means movement. These words reflect their role in cellular dynamics towards an infection [9].

Individual cytokine-based biomarkers of AD have been established in several studies. Mire-Sluis [11] serum and cerebrospinal fluid (CSF, reviewed potential biomarkers that contribute to the pathology of several neurodegenerative diseases, including AD by comparing more than 50 cytokines from over 100 publications. The potential blood-based biomarkers identified for AD are FGF1, IL-11, IL-18, ACT, IL-1β, GMCSF, HGF, IFN-Y, IL-1RA, IL-2, IL-2R, IL-10, IL12, MIP-1α, SDF-1α, sTNF-R1 and sTNF-R2. AD related down-regulated cytokines, namely IL-6R, IL-6, TNF-α and transforming growth factor beta (TGF-β), were also discovered. Furthermore, interleukin 1 beta (IL-1β) was reported to be an important contributor to AD where it is a master regulator of neuroinflammation produced by active inflammatory cells of myeloid lineage in microglia. In addition, serum amyloid A (SAA) was detected together with 1L-1β-immunopositive microglia for AD patients. Thus a link was proposed between P2X7R, SAA and 1L-1β in the central nervous system (CNS) pathophysiology [12].

Several cytokines such as INSR, VEGF-A, PRKACB, DLG4 and BCL2 are presumed to be involved in manganese-inducing AD [13] while 1L-1β and TGFβ can increase the amount of released VEGF (A to E) in certain cells [14]. In addition, according to [14], IL-1β and TGFβ act as mediators in paracrine VEGF-A production. Dayana et al. [15] investigated 12 cytokines and reported that CXCL-10 and IL-13 were promising cytokine biomarkers for AD. CXCL-10 was found to be significantly negatively correlated with Mini-Mental State Examination (MMSE) scores while IL-13 had a significant positive correlation with MMSE in AD patients.

In statistical modelling, logistic regression is commonly used for modelling a binary dependent variable. However, when the dataset is small, the maximum likelihood estimation (MLE) for logistic regression faces several problems such as biased or infinite estimates of the regression coefficients and frequent convergence failure of the likelihood due to separation [16–18]. When the data pattern shows complete separation or quasi complete

**FIGURE 1. Complete separation data example**



| | AD | Control |
|---|---|---|
| Cat1 | 39 | 0 |
| Cat2 | 0 | 39 |

(a) Boxplot of complete separation of continuous variable, X (b) Separation due to dichotomous X variable separation, then the MLE is non-existent (this phenomenon is called monotone likelihood) [19].

Figure 1 illustrates the complete separation for a continuous and categorical variable. Figure 1(a) shows an example of separation due to a continuous variable while Figure 1(b) shows separation due to a dichotomous covariate.

When data separation occurs, three alternatives are frequently employed: (1) increasing the sample size, (2) combining multiple categorical variables and (3) omitting the category (for more than 2 categories). However, increasing the sample size in a clinical study is not financially and administratively practicable, while to combine categories is not always possible especially when there are only two categories and each category is mutually exclusive. To omit the category might be dangerous if the category is important in the study. The Firth's penalised MLE method and the exact logistic regression method can address the separation issue. The exact method is computationally demanding, where it is infeasible when the sample size is greater than 100 [20,21]this study analyzes the effects of individual and built environment characteristics on the route choice using binary logistic regression of 524 survey responses. Conducted in a strategic area, the survey, as often is the case, collects data that are skewed and face the separation issue—the same outcome always occurs for a particular value of a predictor—according to which estimates by the conventional maximum likelihood (ML. The exact method permits replacement of the unsuitable maximum likelihood estimate by a median unbiased estimate [19].

The Firth logistic regression (penalised MLE for logistic regression) works well with multiple predictors and large sample size and it produces nearly unbiased estimates of the coefficients [19,22,23]. Rainey [24] suggested to select between a range of priors in logistic regression such as informative normal prior, Cauchy (0, 2.5) prior, Jeffrey's invariant prior, skeptical normal (0, 2) prior and an Enthusiastic normal (0, 8) prior. Detailed explanation about Firth penalised maximum likelihood is presented in Supplementary Material.

The aim of this study is to illustrate the use of Firth logistic regression in discovering potential cytokine biomarkers data for AD. This paper is organised as follows; the next section covers the description of the data and the Firth's penalised maximum likelihood, the results and discussions are presented in Section 3 while Section 4 concludes the paper.

## METHODS

The data for this paper were obtained from a study "Towards Useful Ageing -Neuroprotective model for healthy longevity among Malaysian elderly" or TUA (the Malay language word for ageing) under the Long-Term Research Grant Scheme (LRGS) programme of Ministry of Education, Malaysia. The ethical approval for TUA programme was obtained from the ethics committees of both Universiti Teknologi MARA (reference no: 600-RMI [5/1/6/01]) and University of Malaya Medical Center (UMMC; reference no: PPUM HU-61/12/1-1]). The dataset consists of 39 people living with AD and 39 age matched healthy control (HC) who were recruited from the Memory and Geriatric Clinic, UMMC. The inclusion criteria for AD patients were age 65 years or older and fulfilled all the conditions of probable AD based on the Revised National Institute of Neurological and Communication Disorders (Alzheimer's disease and Related Disorders Association criteria). Patients were diagnosed as possible or probable AD by a neurologist or geriatrician. The selection criteria were also based on mini mental state examination (MMSE) score (less than or equal to 26). The exclusion criteria were as follows: (1) age less than 65 years old; (2) functionally independent patients as measured by Katz basic activities of daily living and Lawton instrumental activities of daily living (IADL) scales; (3) the MMSE score of more than 27; and (4) patients could communicate and give informed consent.

The inclusion criteria for HC were age more than 65 years old and absence of any documented history of memory or other cognitive impairment, major psychiatric illnesses or mental disorders and concomitant diseases. Selected HCs were also functionally independent as measured by Katz basic activities of daily living and the Lawton instrumental activities of daily living scales, with MMSE of more than 27.

After going through the study information sheet and obtaining consent, 8.5 ml of blood were withdrawn from each patient. Within 30 minutes after venipuncture, the blood samples were centrifuged at 1050g for 3 minutes. The resultant supernatant (serum fraction) was transferred and divided into four divisors of 400 μl of serum. Next, 25μl of serum was added to platinum enzyme-linked immunoassay (ELISA) kits to bind and detect specific targeted cytokines. The detail of the extraction process can be found in [15,25]. Thirteen cytokines were found to be relevant in AD which could be grouped into classical and non-classical inflammatory cytokines. The former were IL-1$\beta$, IL-6, IL-12, IFN$\gamma$, TNF-$\alpha$ and TGF-$\beta$, while the latter, CXCL-1, IL-8, IL-10, IP-10 or CXCL-10, MIP-1$\alpha$, MCP-1 and IL-13.

### Data pre-processing and preliminary analysis

The data were analysed using the R programming language, an open source software for statistical analysis [26], after the data were checked and cleaned to ensure the validity and reliability of all observations. The preliminary analyses included descriptive statistics such as frequency, percentage, mean and standard deviation to assess the distribution of the data.

**TABLE 1. Classical and non-classical cytokines (p = 13)**

| Classical Inflammatory Cytokines | Non-Classical Inflammatory Cytokines |
|---|---|
| Interleukin 1 beta (IL-1β) | Chemokine (C-X-C motif) ligand 1 (CXCL-1) |
| Interleukin 6 (IL-6) | Interleukin 8 (IL-8) |
| Interleukin 12 (IL-12) | Interleukin 10 (IL-10) |
| Interferon gamma (IFNγ) | Interferon gamma-induced protein 10 (IP-10) or C-X-C motif chemokine 10 (CXCL-10) |
| Tumor necrosis factor alpha (TNF-α) | Macrophage inflammatory protein 1 alpha (MIP-1α) |
| Transforming growth factor beta (TGF-β) | Monocyte chemoattractant protein-1 (MCP-1) |
| | Interleukin 13 (IL-13) |

The demographics of subjects such as age, gender, ethnicity, smoking history and history of alcohol consumption were recorded to outline their physiognomies. The binary logistic regression was employed to ascertain the association between AD and physiognomic variables. Initially, the univariable logistic regression was performed to assess the relationship of these variables, then the multivariable logistic regression was performed to evaluate relationship of the variables with the existence of other variables. The performance of this model was reported by using measures such as sensitivity, specificity, accuracy, Akaike Information Criterion (AIC) and area under the receiver-operating characteristic (ROC).

To establish the predictive model of blood-based cytokine biomarker for AD, we only focused on the cytokine variables only. We started the analysis by detecting the complete separation variables in cytokine using boxplot graphics and confirmed the results using linear programming method developed by [27]. These were done to know the suitability of Firth logistic regression application in the dataset.

Then, univariable Firth logistic penalised ML method was carried out to determine the significant individual biomarkers. In developing the univariable analysis for cytokines, data were not sliced into training and test set because the existence of the complete separation of variables in the dataset and due to the small sample size. Then only cytokines with p-value less than 0.25 [20] were selected to be in the multivariable Firth logistic regression. The forward selection, backward elimination and stepwise selection for Firth logistic regression were applied for feature selection.

However, since the data have covariates that were highly correlated which would lead to multicollinearity, we performed principle component analysis (PCA) to group the correlated cytokines together. Bartlett's test of sphericity was used to test if the correlation matrix is an identity matrix and then, a PCA using a varimax rotation was carried out.

At the final stage, we conducted the multivariable Firth logistic regression with the principal components. The performance of the final model was measured with Hosmer-Lemeshow test, sensitivity, specificity, AIC and area under ROC curve. We also did a comparison of the classification performance with different number of components.

## RESULTS AND DISCUSSIONS

Table 2 shows the frequency distribution of the physiognomic variables for HC and AD and univariable logistic regression analysis results. There is no significant relationship between AD and HC with gender, ethnicity and smoking status. However, history of alcohol consumption and age were statistically significantly related to AD (p-value < 0.05).

Multivariate logistic regression was employed to determine the relationship of physiognomic factors (gender, age and history of alcohol consumption) with AD using Forward and Backward LR (Likelihood ratio) selection method. Only Age and Alcohol were selected as significant predictors. Based on Hosmer-Lemeshow test, the model with two covariates (Age, Alcohol) best fits the data (Chi-Square (8 df) = 7.9134, p-value: 0.442). The Akaike Information Criterion (AIC) was found to be 75.55, where the accuracy and sensitivity were 80.77% and 82.05%, respectively. The area under the ROC curve was 0.87.

In Table 3, the odds-ratio for age (OR=1.29) indicates that for every one-year increase in age, the odds of getting AD increases by 29% when controlled for alcohol. Additionally, those who consumed alcohol are 3.8 (OR=1/0.26=3.8) time more likely to have AD.

Forward likelihood ratio method and backward likelihood ratio method were tested, multicollinearity and clinically plausible interaction checked. The model adequacy was checked using Hosmer-Lemeshow test. Diagnostic measures including outlier identification and influential statistical were done.

### Separation detection

In Figure 2, we present the five cytokines that showed complete separation between AD and HC: IL-1β, IL-6, IL-10, IL-13 and IP-10 (CXCL-10). In addition, the results in Table 4, confirmed that only these stated cytokines had complete separation issue since the intercept and coefficient were not equal to 0 (β ≠ 0) and were infinite [27]. Since these cytokines had complete separation between AD and HC, we can conclude that these

**TABLE 2. Univariable logistic regression of physiognomic factors for Alzheimer's disease subject (N=78)**

| Variable | HC<br>(N=39)<br>[n (%)] | AD<br>(N=39)<br>[n (%)] | $\beta$ [a]<br>(s.e.)[b] | Crude OR (95% CI)[c] | p-value |
|---|---|---|---|---|---|
| **Age** | 72.13[d]<br>(5.04)[e] | 80.69[d]<br>(6.41)[e] | 0.25<br>(0.06) | 1.28<br>(1.16, 1.44) | <0.0001 |
| **Gender** | | | | | |
| Female | 15<br>(40.54) | 22<br>(59.55) | | 1 | |
| Male | 24<br>(58.54) | 17<br>(41.46) | -0.73 (0.46) | 0.48<br>(0.19, 1.18) | 0.1145 |
| **Ethnicity** | | | | | |
| Chinese | 22<br>(45.83) | 26<br>(54.17) | | 1 | |
| Indian | 9<br>(52.94) | 8<br>(47.06) | -0.28 (0.57) | 0.75<br>(0.24, 2.29) | 0.6150 |
| Malay | 8<br>(61.54) | 5<br>(38.46) | -0.64 (0.64) | 0.53<br>(0.14, 1.82) | 0.3190 |
| **Alcohol** | | | | | |
| No | 17<br>(39.54) | 26<br>(60.46) | | 1 | |
| Yes | 22<br>(62.86) | 13<br>(37.14) | -0.95 (0.47) | 0.39<br>(0.15, 0.96) | 0.0425 |
| **Smoking** | | | | | |
| No | 12<br>(57.14) | 9<br>(42.86) | | 1 | |
| Yes | 27<br>(47.37) | 30<br>(52.63) | 0.39<br>(0.52) | 1.48<br>(0.54, 4.16) | 0.4450 |

[a] *Crude estimated slope (regression coefficient) for binary logistic regression;*
[b] *Standard Error for coefficient;*
[c] *95% confidence interval for odds-ratio (OR);*
[d] *mean for variable Age;*
[e] *Standard deviation for variable Age*

variables are potential biomarkers of AD.

Detection of separation was done using linear programming method developed by [27] via R programming package ("brglm2"). The intercept and coefficient values for each variable should equal to zero to indicate there are no separation. If the intercept or coefficient is not equal to zero, it indicates that the data has separated condition (either quasi-separation or complete separation).

## Fitting Firth's logistic regression

The univariable Firth logistic regression results in Table 5 show that all cytokines were significantly associated with AD except TGF-$\beta$, TNF-$\alpha$, CXCL-1 and IL-8 because the Firth's p-values were more than 0.05 and the AICs for these biomarkers were significantly higher compared to other cytokines. Next, we fitted multivariate Firth logistic regression model using variables with p-value ≤ 0.25 in Table 5. The forward selection, backward elimination and stepwise selection algorithms were applied for the Firth's logistics regression with nine biomarkers. The forward selection and stepwise selection selected only IL-13 in the final main effects model. On the contrary, only IL-6 was selected in the final main effects model for backward elimination. The non-agreement in selecting the best main effect model is due to high correlation of the covariates (biomarkers). Table 6 shows the Pearson's correlation between the covariates. IL-1$\beta$ has high correlations (0.80) with IL-6, IL-13 and IP-10. Meanwhile IL-6 has high correlation with IL-1$\beta$, IP-10 and IL-13.

Next, the PCA was applied to cluster correlated

**TABLE 3. Multivariable logistic regression of physiognomic factors for Alzheimer's disease subject (N=78) (Forward selection)**

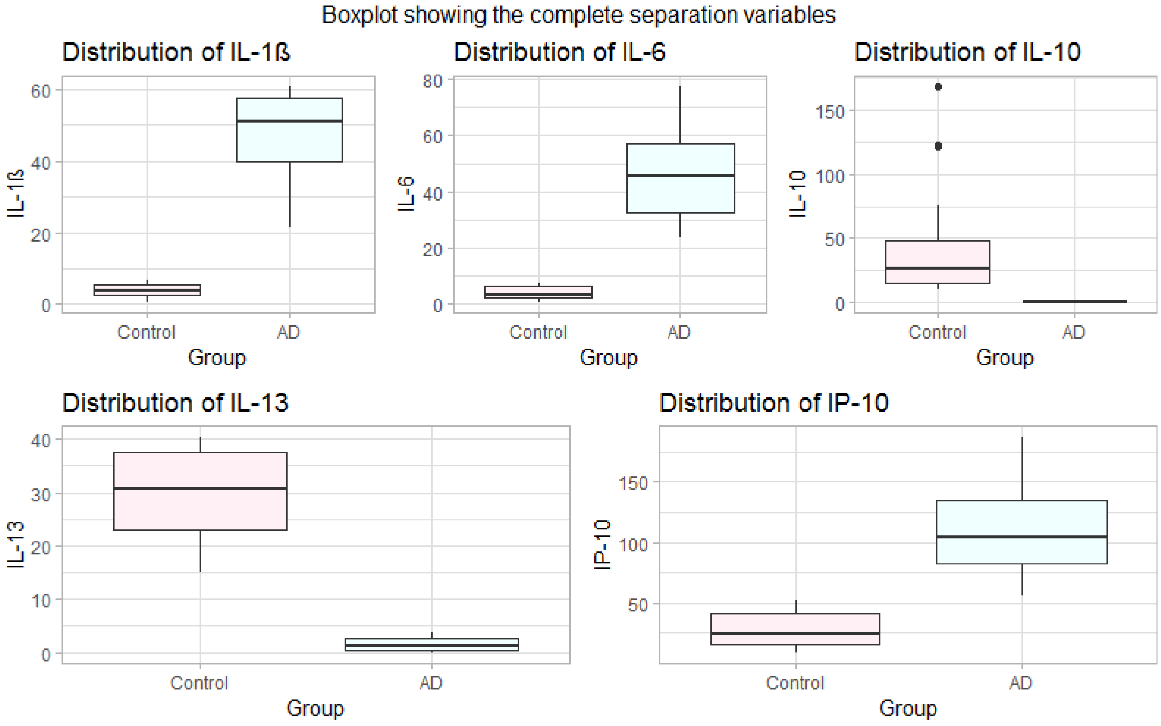
| Factors | β [a] | s.e.[b] | Wald Statistic[c] | Adjusted OR[d] (95% CI)[e] | p-value |
|---|---|---|---|---|---|
| **Age** | 0.25 | 0.06 | 4.50 | 1.29<br>(1.17, 1.46)) | <0.0001 |
| **Alcohol** | | | | | |
| No | | | | 1 | |
| Yes | -1.35 | 0.62 | -2.16 | 0.26<br>(0.07, 0.84) | 0.0303 |

[a] *Regression coefficient;*
[b] *Standard error;*
[c] *z-value* $= \frac{\hat{\beta}}{s.e}$
[d] *Adjusted odd ratio*
[e] *95% confidence interval constant = -18.81*

**TABLE 4. Separation detection using linear programming method (p=13)**

| Variable | Intercept | Coefficient | Variable | Intercept | Coefficient |
|---|---|---|---|---|---|
| **IL-1β** | -∞ | ∞ | **IL-8** | 0 | 0 |
| **IL-6** | -∞ | ∞ | **IL-10** | ∞ | -∞ |
| **IL-12** | 0 | 0 | **IL-13** | ∞ | -∞ |
| **IFN-γ** | 0 | 0 | **IP-10** | -∞ | ∞ |
| **TGF-β** | 0 | 0 | **MCP-1** | 0 | 0 |
| **TNF-α** | 0 | 0 | **MIP-1α** | 0 | 0 |
| **CXCL-1** | 0 | 0 | | | |

**FIGURE 2. Box plot for five cytokines**



covariates for further analysis. All nine covariates that were statistically significant in Table 5 were considered for PCA. The correlation matrix was used to calculate Bartlett's test of sphericity. The Bartlett's test result was significant and thus we could proceed with PCA [chi-squared (df): 506.53 (36)].

A PCA using a varimax rotation of all nine covariates was carried out. The covariates with a factor loading of 0.4 and higher (indicating satisfactory loading) were regarded as valid and significant contributors to the component. In PCA, only the first component (PC) had eigenvalue more than 1 (eigenvalue: 5.61) and the percentage of variance explained was 62.32%. Based on factor loadings, 4 PCs were extracted, where the total variance explained was 85.81%.

The first principal component (PC1) represents four biomarkers, IL-1β, IL-6, IL-13 and MCP-1 while PC2 represents two biomarkers, IFN-γ and MIP-1α. The third PC (PC3) represents two biomarkers, IL-12 and IP-10 and the fourth PC (PC4) represents only IL-10. The rotated component loadings for the nine covariates are presented in Table 7.

Then, a Firth's logistic regression model using the four principal components was fitted. Table 8 showed that all four components were statistically significant. All the PCs

**TABLE 5. Univariable Firth's logistic regression in measuring cytokines association with AD (nAD = 39, nHC = 39)**

| Cytokines | AD [Mean (SD)] | HC [Mean (SD)] | Crude OR[a] (AIC[b]) | AUC[c] | (95% CI)[d] | p-Value[e] |
|---|---|---|---|---|---|---|
| IL-1β | 46.69 (12.54) | 3.94 (1.76) | 1.43 (6.01) | 1.00 | (1.15, 1.79) | 0.0017* |
| IL-6 | 46.71 (16.00) | 3.62 (2.02) | 1.40 (6.03) | 1.00 | (1.16, 1.70) | 0.0006* |
| IL-12 | 33.99 (22.57) | 7.07 (4.65) | 1.36 (48.68) | 0.95 | (1.17, 1.58) | 0.0001* |
| IFN-γ | 1.41 (1.09) | 0.25 (0.16) | 42.51 (70.80) | 0.85 | (2.65, 681.09) | 0.0081* |
| TGF-β | 1.51 (2.36) | 0.66 (2.61) | 1.14 (109.73) | 0.69 | (0.91, 1.43) | 0.2558 |
| TNF-α | 0.06 (0.10) | 0.31 (1.35) | 0.09 (109.25) | 0.68 | (0.00, 7.61) | 0.2829 |
| CXCL-1 | 5.32 (5.14) | 5.42 (7.01) | 1.00 (112.13) | 0.52 | (0.93, 1.07) | 0.9559 |
| IL-8 | 0.28 (0.59) | 0.22 (0.14) | 1.19 (111.86) | 0.44 | (0.40, 3.56) | 0.7512 |
| IL-10 | 0.78 (0.54) | 39.65 (34.79) | 0.52 (6.00) | 1.00 | (0.36, 0.76) | 0.0006* |
| IL-13 | 1.64 (1.25) | 29.57 (7.61) | 0.63 (6.11) | 1.00 | (0.47, 0.83) | 0.0011* |
| IP-10 | 112.54 (36.03) | 28.17 (14.01) | 1.37 (7.61) | 1.00 | (1.04, 1.82) | 0.0274* |
| MCP-1 | 10.09 (4.195) | 3.89 (2.62) | 1.74 (64.52) | 0.91 | (1.35, 2.24) | <0.0001* |
| MIP-1α | 0.56 (0.42) | 0.20 (0.15) | 155.53 (85.04) | 0.80 | (11.29, 2142.13) | 0.0002* |

[a] *Crude odds ratio (OR) were calculated based on exponential coefficient of each cytokine.*

[b] *A lower value of Akaike Information criteria (AIC) is preferred as it indicates the model fits better.*

[c] *Area under the Receiver Operating Curve (AUC) range from 0 to 1, where the higher value of AUC indicates the model fits better.*

[d] *The variable is considered significant if the 95% confidence interval (95% CI for odds ratio) does not include 1 in the interval range.*

[e] *Simple Firth's logistic regression was done for on all individual cytokines.*

* *indicates $p < 0.05$ and the cytokine that have significant association with AD*

**TABLE 6. Pearson's correlation between covariates (p=9)**

| | IL-1β | IL-6 | IL-12 | IFN-γ | IL-10 | IL-13 | IP-10 | MCP-1 | MIP-1α |
|---|---|---|---|---|---|---|---|---|---|
| IL-1β | 1.00 | 0.84 | 0.66 | 0.63 | -0.59 | -0.86 | 0.80 | 0.64 | 0.55 |
| IL-6 | | 1.00 | 0.59 | 0.44 | -0.54 | -0.84 | 0.80 | 0.67 | 0.45 |
| IL-12 | | | 1.00 | 0.42 | -0.39 | -0.60 | 0.68 | 0.38 | 0.41 |
| IFN-γ | | | | 1.00 | -0.35 | -0.56 | 0.54 | 0.36 | 0.51 |
| IL-10 | | | | | 1.00 | 0.63 | -0.52 | -0.32 | -0.37 |
| IL-13 | | | | | | 1.00 | -0.77 | -0.60 | -0.47 |
| IP-10 | | | | | | | 1.00 | 0.60 | 0.50 |
| MCP-1 | | | | | | | | 1.00 | 0.33 |
| MIP-1α | | | | | | | | | 1.00 |

[a] *Pearson's correlation coefficient was employed.*

**TABLE 7. Rotated component loadings using varimax (p=9)**

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| **IL-1β** | 0.585 | 0.428 | 0.465 | 0.383 |
| **IL-6** | 0.687 | 0.212 | 0.447 | 0.382 |
| **IL-12** | 0.183 | 0.224 | 0.912 | 0.138 |
| **IFN-γ** | 0.234 | 0.797 | 0.214 | 0.118 |
| **IL-10** | -0.170 | -0.190 | -0.161 | -0.929 |
| **IL-13** | -0.574 | -0.324 | -0.419 | -0.492 |
| **IP-10** | 0.534 | 0.331 | 0.581 | 0.304 |
| **MCP-1** | 0.925 | 0.167 | 0.101 | 0.060 |
| **MIP-1α** | 0.141 | 0.825 | 0.160 | 0.183 |

*[a] Rotation was done using varimax method.*

**TABLE 8. Multivariable Firth's logistic regression of 4 PCs for Alzheimer's disease subject (n=78)**

| Rotated Scores | β [a] | s.e.[b] | Wald Statistic[c] | Adjusted OR[d] (95% CI)[e] | p-value |
|---|---|---|---|---|---|
| **PC 1** | 3.97 | 1.37 | 2.88 | 52.75 (3.56, 780.87) | 0.0039 |
| **PC 2** | 1.34 | 0.59 | 2.26 | 3.83 (1.20, 12.27) | 0.0236 |
| **PC 3** | 2.19 | 0.82 | 2.68 | 8.94 (1.81, 44.27) | 0.0073 |
| **PC 4** | 1.50 | 0.52 | 2.85 | 4.47 (1.60, 12.50) | 0.0043 |

*[a] Regression coefficient*

*[b] Standard error*

*[c] z-value* = $\frac{\hat{\beta}}{s.e}$

*[d] Adjusted odd ratio*

*[e] 95% confidence interval constant = 0.4663*

*Forward likelihood ratio method and backward likelihood ratio method were tested, multicollinearity and clinically plausible interaction checked. The model accuracy was checked using Hosmer-Lemeshow test, classification table and ROC curve. Diagnostic measures including outlier identification and influential statistical were done.*

**TABLE 9. Hierarchical of model performance on each PCs in Firth's logistic regression (n=78)**

| Number of PCS | Sensitivity (%) | Specificity (%) | Accuracy (%) | Precision (%) |
|---|---|---|---|---|
| PC1 | 71.79 | 84.62 | 78.21 | 82.35 |
| PC1 + PC2 | 79.49 | 89.74 | 84.62 | 88.57 |
| PC1 + PC2 + PC3 | 92.31 | 92.31 | 92.31 | 92.31 |
| PC1 + PC2 + PC3 + PC4 | 100.00 | 100.00 | 100.00 | 100.00 |

show an impact on AD since the adjusted odds ratio were larger than 1. The biggest contribution is PC 1, which has 4 covariates involved (IL-1β, IL-6, IL-13 and MCP-1α), followed by PCs 3, 4 and 5.

The model with four PCs fitted the data best based on Hosmer-Lemeshow test (Chi-Square (8 df) =2.78, p-value: 0.9472). The AIC was found to be 15.441, where sensitivity and the under the ROC curve were 97.22% and 1.00, respectively.

We then compared the classification performance using different numbers of components. Firth's logistic regression model with PC1 only had an accuracy of 78.2% with sensitivity and specificity of 71.8% and 84.6% respectively. Table 9 shows the hierarchy of model performance starting from PC1 and adding all PCs in the Firth's logistic regression model. The classification performance increased for every PC added and achieved 92.3% accuracy, sensitivity, specificity and precision when we used three PCs.

## CONCLUSION

This paper focused on the issue of complete separation of data and illustrates the use of Firth Logistic regression as an alternative to the classical logistic regression model.

The aim was to establish a reliable prediction model for AD cytokine biomarkers. In the presence of complete separation, the classical binary logistic regression would produce infinite MLE estimates of the coefficients (non-existence of an MLE for non-overlapped data) [24]. Firth's logistic regression was used to investigate the relationship between cytokines and AD as it can cater for complete separation issue. Due to the presence of multicollinearity among the biomarkers, we used principal component analysis techniques to cluster the biomarkers.

This study found that IL-1β, IL-6, IL-13 and MCP-1α were the important biomarkers. Yin [28] reported that IL-1β and homozygous APOE 4 combined are associated with increased hazard in developing AD. IL-1β was also reported with six accompanying pathways in Cytoscape that linked them to AD [29]. The importance of IL-6 was also supported by [30] which reported that the levels of IL-6 and IFN-γ were significantly higher in altered T-lymphocytes of AD compared to HC. In our 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 [31,32].

Furthermore, tumor necrosis factor (TNF-α) was found to be insignificant and did not affect AD based on univariable Firth's logistic regression. This result was supported by [33] and the authors concluded that the alterations in immunological conditions involving tumor necrosis factor mediated signaling were not the primary events in commencing AD pathology including amyloid plaques and tangle development.

Next, IL-10 (PC 4) was also found to be associated with an increase in risk for AD in multivariable Firth (OR: 4.47). The IL-10 and IL-13 are said to be anti-inflammatory cytokines by virtue of their ability to suppress genes for pro-inflammatory cytokines [34]. These results are in line with the findings by Dayana et al. [15].

Then, the odds-ratio for interferon gamma-induced protein 10 (IP-10) or C-X-C motif chemokine 10 (CXCL-10) in PC 3 indicated an elevated risk for AD. These results were supported by [35], where the authors stated that CXCL-10 were positively correlated with the severity of cognitive decline in AD patients. Furthermore, in animal studies, CXCL-10 has been implicated in disease progression of APPSWE mouse. It had been demonstrated that the ablation of CXCL-10 receptor, chemokine (C-X-C motif) receptor 3 in APPSWE/PS1ΔE9 mice ameliorated amyloidosis and cognitive decline [35].

In conclusion, Firth's logistic regression is a useful technique for the identification of significant biomarkers when there is an issue of data separation. The results of this study can be validated by increasing the sample size of study. In future work, we seek to develop an efficient prediction model for AD by combining cytokines, transcriptomics and proteomics biomarkers.

## Acknowledgement

The authors would like to thank Universiti Teknologi MARA (UiTM) for supporting this research under the BESTARI grant (600-IRMI/DANA 5/3/BESTARI (113/2018).

## Availability of data and material

The dataset used in this study was obtained from the TUA-Long Term Research Grant (LRGS) programme.

## Author's contribution

MNA and YBW analysed the data and drafted the manuscript. YBW and OSH supervised the analysis of the methods. ABAM and YZ conceived the project. All authors read and approved the final manuscript.

## Competing interest

The authors declare that they have no competing interests.

## Consent for publication

Not applicable

## Ethics approval and consent to participate

Not applicable

## Funding

This research was supported by the Ministry of Education (MOE) under Long Term Research Grant Scheme (Reference no: 600-RMI/LRGS 5/3 [3/2012]) and Bestari Grant Phase 2/2018 (Reference no: 600-IRMI/DANA 5/3/BESTARI (113/2018)).

## References

1. Hao X, Yao X, Yan J, Risacher SL, Saykin AJ, Zhang D, et al. Identifying Multimodal Intermediate Phenotypes Between Genetic Risk Factors and Disease Status in Alzheimer's Disease. Neuroinformatics. 2016;14(4):439–52.
2. Liu M, Cheng D, Wang K, Wang Y. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis. Neuroinformatics. 2018;16(3–4):295–308.
3. Wang C, Fei G, Pan X, Sang S, Wang L. High thiamine diphosphate level as a protective factor for Alzheimer ' s disease.

Neurol Res [Internet]. 2018;6412(May):1–7. Available from: https://doi.org/10.1080/01616412.2018.1460704

4. Silzer TK, Phillips NR. Etiology of type 2 diabetes and Alzheimer's disease: Exploring the mitochondria. Mitochondrion [Internet]. 2018;(February):0–1. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1567724917303392
5. Cayton H, Graham N, Warner J. Alzheimer's and other dementias. London: Class Publishing; 2008.
6. Pelleieux S, Picard C, Lamarre-Théroux L, Dea D, Leduc V, Tsantrizos YS, et al. Isoprenoids and tau pathology in sporadic Alzheimer's disease. Neurobiol Aging [Internet]. 2018;65:132–9. Available from: https://doi.org/10.1016/j.neurobiolaging.2018.01.012
7. Nuri THM, Hong YH, Ming LC, Joffry SM, Othman MF, Neoh CF. Knowledge on Alzheimer's disease among public hospitals and health clinics pharmacists in the State of Selangor, Malaysia. Front Pharmacol. 2017;8(OCT):1–6.
8. Dembic Z. The role of cytokines in disease related to immune response [Internet]. Methods for the Study of Marine Benthos. 2013. i–xvii. Available from: http://dx.doi.org/10.1002/9781118542392.fmatter
9. Fitzgerald KA, O'Neill LAJ, Gearing AJH, Callard RE. The Cytokine Facts Book. Vol. 53, Journal of Chemical Information and Modeling. 2013. 1689–1699 p.
10. Mire-Sluis A. Cytokines and disease. Trends Biotechnol. 1993;11(3):74–7.
11. Brosseron F, Krauthausen M, Kummer M, Heneka MT. Body Fluid Cytokine Levels in Mild Cognitive Impairment and Alzheimer's Disease: a Comparative Overview. Mol Neurobiol. 2014;50(2):534–44.
12. Facci L, Barbierato M, Zusso M, Skaper SD, Giusti P. Serum amyloid A primes microglia for ATP-dependent interleukin-1β release. J Neuroinflammation. 2018;15(1):1–11.
13. Ling J, Yang S, Huang Y, Wei D, Cheng W. Identifying key genes, pathways and screening therapeutic agents for manganese-induced Alzheimer disease using bioinformatics analysis. Medicine (Baltimore). 2018;97(22):1–6.
14. Stocks J, Bradbury D, Corbett L, Pang L, Knox AJ. Cytokines upregulate vascular endothelial growth factor secretion by human airway smooth muscle cells: Role of endogenous prostanoids. FEBS Lett. 2005;579(12):2551–6.
15. Mohd Hasni DS, Lim SM, Chin AV, Tan MP, Poi PJH, Kamaruzzaman SB, et al. Peripheral cytokines, C-X-C motif ligand10 and interleukin-13, are associated with Malaysian Alzheimer's disease. Geriatr Gerontol Int. 2016;1–8.
16. Rahman MS, Sultana M. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. BMC Med Res Methodol. 2017;17(1):1–15.
17. Cordeiro MG, McCullagh P. Bias Correction in Generalized Linear Models. Vol. 53, Journal of the Royal Statistical Society: Series B (Methodological). 1991. p. 629–43.
18. M. J. Box. Bias in nonlinear estimation. J R Stat Soc Ser B. 1971;33(2):171–201.
19. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21(16):2409–19.
20. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Vol. 398. John Wiley & Sons; 2013.
21. Gim THT, Ko J. Maximum Likelihood and Firth Logistic Regression of the Pedestrian Route Choice. Int Reg Sci Rev. 2017;40(6):616–37.
22. Kosmidis I, Firth D. A generic algorithm for reducing bias in parametric estimation. Electron J Stat. 2010;4:1097–112.
23. Firth D. Bias Reduction of Maximum Likelihood Estimates. 1993;80(1):27–38.
24. Rainey C. Dealing with separation in logistic regression models. Polit Anal. 2016;24(3):339–55.
25. Dayana SMH, Lim SM, Tan MP, Chin A V, Poi PJH, Kamaruzzaman SB, et al. IP-10 and IL-13 as potentially new, non-classical blood-based cytokine biomarker for Alzheimer's disease. Neuorology Neurosci. 2014;43(April):i32.
26. RStudio Team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA; 2018. Available from: http://www.rstudio.com/
27. Konis KP. Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models. 2007; Available from: http://ora.ox.ac.uk/objects/ora:2848
28. Yin Y, Liu Y, Pan X, Chen R, Li P, Wu HJ, et al. Interleukin-1β Promoter polymorphism enhances the risk of sleep disturbance in Alzheimer's disease. PLoS One. 2016;11(3):1–13.
29. Xie L, Lai Y, Lei F, Liu S, Liu R, Wang T. Exploring the association between interleukin-1beta and its interacting proteins in Alzheimer's disease. Vol. 11, Molecular medicine reports. 2015. p. 3219–28.
30. Azad FJ, Talaei A, Rafatpanah H, Yousefzadeh H. Association between Cytokine Production and Disease Severity in Alzheimer ' s Disease. Iran J Allergy, Asthma Immunol. 2014;13(6):433–9.
31. Licastro F, Grimaldi LME, Bonafè M, Martina C, Olivieri F, Cavallone L, et al. Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain. Neurobiol Aging. 2003;24(7):921–6.
32. Mrak RE, Griffin WST. Potential inflammatory biomarkers in Alzheimer's disease. J Alzheimers Dis [Internet]. 2005;8(4):369–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16556968
33. Nagae T, Araki K, Shimoda Y, Sue LI, Beach TG, Konishi Y, et al. Cytokines and Cytokine Receptors Involved in the Pathogenesis of Alzheimer ' s Disease. J Clin Cell Immunol. 2016;7(4):1–31.
34. Rubio-Perez JM, Morillas-Ruiz JM. A review: Inflammatory process in Alzheimer's disease, role of cytokines. Sci World J. 2012;2012.
35. Minter MR, Taylor JM, Crack PJ. The contribution of neuroinflammation to amyloid toxicity in Alzheimer's disease. J Neurochem. 2016;136(3):457–74.

*

# SUPPLEMENTARY MATERIAL

This Supplementary Material explained about the Firth penalised maximum likelihood.

## Firth penalised maximum likelihood

In the binary logistic model, for each observation, the response Y can take either 1 for success or 0 for failure values such that $p(Y_i = 1) = \pi_i$ or $p(Y_i = 0) = 1 - \pi_i$ are the probabilities of success and failure respectively. The structure component of the logistic model is based on the logit of the probability of success equal to a linear combination of the covariates such that $logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \beta^T x_i$, where is the elements containing the model parameter that we wish to estimate and $\beta = (\beta_1, \dots, \beta_p)^T$.

The goal of logistic regression is to estimate the outcome using the most parsimonious model. The MLE obtained the estimate for β maximizing the likelihood function of *l(β)*. The joint probability of $P(Y_1 = y_i, \ \dots, \ Y_n = y_n : \pi) = \prod[\pi_i{}^{y_i} \times (1 - \pi_i)^{(1-y_i)}]$, where is the probability of success for a given case with the values of the risk factors in the i<sup>th</sup> combination and $y_i$ is the response value for the target variable with $Y_i \sim binomial(1, \pi_i)$. Then, to get the ML, the logarithm of the likelihood function is taken, namely $\log l(\pi;\ y_1, \dots, y_n) = \sum[y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)]$. The log likelihood function for the model parameter can be obtained by substituting the logit equation, which leads to: $l(\beta) = l(\beta;\ y, X) = \sum\sum y_i x_{ij}\beta_j - \sum \log(1 + exp \sum x_{ij}\beta_j)$, where $x_{ij}$ is the subset in the *i<sup>th</sup>* row and *j<sup>th</sup>* column of the data matrix *x*.

Finding the MLE requires computing the derivatives of the likelihood function. To produce optimum β, the score function K(β) is taken into consideration where, by partial derivative on $K(\beta) = \frac{\partial \log l(\beta)}{\partial \beta} = \ \sum[(y_i - \pi_i) \times x_i]$. Setting the K(β)=0, the MLE of β is obtained using the Newton-Raphson [1–3].

The parameter estimates of a binary logistic regression model using ML sometimes does not converge to finite values when the sample points are separated (known as monotone likelihood). When dealing with monotone likelihood (biased MLE), [4] proposed the penalised method to handle separation issue. Fundamentally, Firth's penalised method is to use exact regular probability function with a bias term that is receptive to small sample size and rare targets [5]. Firth used the penalty term $\frac{1}{2} trace[l(\beta)^{-1} \partial I(\beta)/\partial \beta_j]$ in the ML-based score function to remove first order bias. The Firth's penalised likelihood is $l_{Firth}(\beta) = l(\beta) \times |I(\beta)|^{0.5}$, and the log of the likelihood are $\log l_{Firth}(\beta) = \log\big(l(\beta)\big) + 0.5 * \log(|I(\beta)|)$. I(β) denotes the Fisher information matrix. Next, the Firth's penalised score function can be interpreted as $U_{Firth}(\beta) = \frac{\partial}{\partial \beta}[\log(l(\beta)) + 0.5 * \log(|I(\beta)|)] = \sum([(y_i - \pi_i + h_i \ \times (0.5 - \pi_i)] \times \ x_i)$, where, $h_i$ is the diagonal elements in the Firth's likelihood structure of predicted matrix $H = W^{0.5} \times X \times (X' \times W \times X)^{-1} \times X' \times W^{0.5}$, where *W* is the diagonal matrix of $[\pi_i \times (1 - \pi_i)]$ and, the regular design matrix.

## References

1. Gim THT, Ko J. Maximum Likelihood and Firth Logistic Regression of the Pedestrian Route Choice. Int Reg Sci Rev. 2017;40(6):616–37.
2. Kosmidis I, Firth D. A generic algorithm for reducing bias in parametric estimation. Electron J Stat. 2010;4:1097–112.
3. Czepiel SA. Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. Cl Notes [Internet]. 2012;1–23. Available from: papers3://publication/uuid/4E1E1B7E-9CAC-4570-8949-E96B51D9C91D
4. Firth D. Bias Reduction of Maximum Likelihood Estimates. 1993;80(1):27–38.
5. Rahman MS, Sultana M. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. BMC Med Res Methodol. 2017;17(1):1–15.

*