Scientific Research Publishing

Advances in Alzheimer's Disease, 2020, 9, 77-86
https://www.scirp.org/journal/aad
ISSN Online: 2169-2467
ISSN Print: 2169-2459

Alzheimer's Disease: The Relative Importance Diagnostic

Maryam Habadi[1], **Christos P. Tsokos**[2]

[1]Departmentof Statistics, King Abdulaziz University, Jeddah, KSA
[2]Department of Mathematics and Statistics, University of South Florida, Tampa, USA
Email: mhabadi@kau.edu.sa

How to cite this paper: Habadi, M. and Tsokos, C.P. (2020) Alzheimer's Disease: The Relative Importance Diagnostic. *Advances in Alzheimer's Disease*, **9**, 77-86.
https://doi.org/10.4236/aad.2020.94006

Received: November 9, 2020
Accepted: December 28, 2020
Published: December 31, 2020

Abstract

As the population ages, Alzheimer's disease is rapidly increasing, and the diagnosis of the disease is still poorly understood. In comparison to cancer, 90% of patients become aware of their diagnosis, but only 45% of the people with Alzheimer's are aware. Thus, the need for biomarkers for reliable diagnosis is tremendous to help in finding treatment for this serious disease. Hence, the main aim of this paper is to utilize information from baseline measurements to develop a statistical prediction model using multiple logistic regression to distinguish Alzheimer's disease patients from cognitively normal individuals. Our optimal predictive model includes six risk factors and two interaction terms and has been evaluated using classification accuracy, sensitivity, specificity values and area under the curve.

Keywords

Alzheimer's Disease, Multiple Logistic Regression, Predictive Model, Classification Accuracy

1. Introduction

Alzheimer's disease causes memory loss, and it is not a normal part of aging. It is the only disease that cannot be prevented, treated or even slowed. A recent fact from Alzheimer's Association report in 2018 shows that only deaths from Alzheimer's disease have increased significantly while from other major causes of death in the United States have decreased significantly. The bar chart in Figure 1 shows the percentage changes in the top causes of death between 2000 and 2015. As we can see, the number of deaths from heart disease, the number one cause of death in the United States, decreased by 11%; however, recorded death from Alzheimer's disease increased by 123% [1].

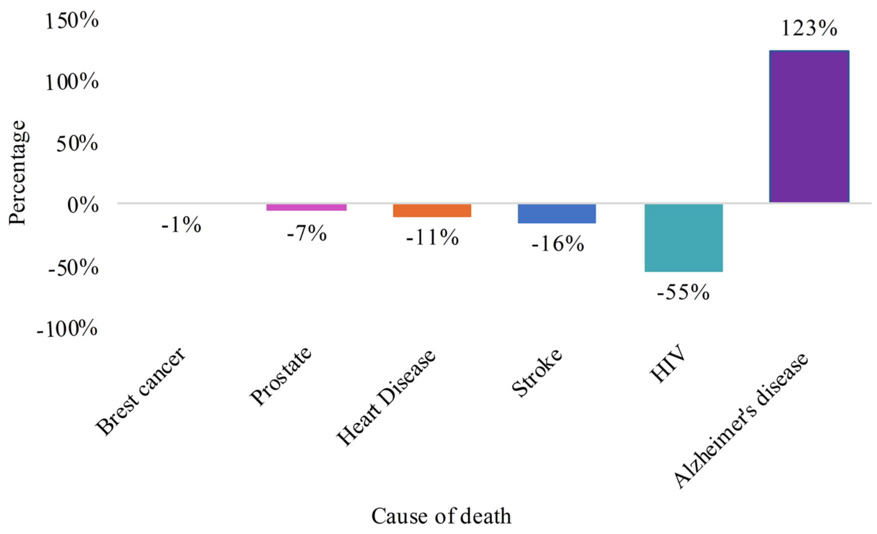

Figure 1. Percentage of selected causes of death between 2000-2015. Source: 2018 Alzheimer's Disease Facts and Figures.

In comparison to cancer, 90% of patients become aware of their diagnosis, but only 45% of the people with Alzheimer's are aware [2]. Thus, researchers and doctors are working to develop a diagnosis pattern of Alzheimer's disease that helps in early detection of the disease before symptoms increase. Different types of tests include neuropsychological test, blood tests, cerebrospinal fluid analysis, and brain imaging have been used to help understand and diagnosis this severe disease. Neuropsychological tests are an assessment of the brain function to evaluate numbers of areas including attention, problem-solving, memory, language, mood and behavior. Commonly used test tools include the Mini-Mental Status Examination (MMSE) and Dementia Rating Scale (CDR).

Brain imaging is used to detect some brain changes caused by Alzheimer's disease, that is, detecting the levels of plaques and tangles, the two types of disorders in the brain associated with the presence of Alzheimer's. Plaques are found between the dying cells in the brain from the buildup of a protein called beta-amyloid and tangles are twisted fibers within the dying cells from the other protein called tau. Beta-Amyloid and tau proteins are normally fragmented that the body produces, but in Alzheimer's the proteins are abnormal.

Cerebrospinal fluid analysis (CSF) is collecting the clear fluid that protects and surrounds the brain and spinal cord to determine the levels of beta-amyloid, total tau (T-tau) and phosphorylated tau (P-tau) proteins. Since CSF is in direct contact with the brain and spine, collecting a sample of the fluid can be a useful diagnostic tool for this neurodegenerative disease.

The primary goal of the present study is to develop the best statistical model to correctly predict Alzheimer's patients with their demographic, CSF, laboratory and brain imaging factors using logistic regression model. This model will allow us to accurately evaluate the probability that a patient is diagnosed with Alzheimer's disease. Moreover, we can rank the significant contributing risk factors based on their relative importance to the response. Hence, medical doctor can use our proposed data-driven model as a decision supportive before starting any treatment.

2. The Data

In the present study, we used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The primary goal of ADNI is to detect and track the progression of Alzheimer's disease by combining clinical, imaging, genetic and biological markers of participants to help researchers and doctors develop new treatments. More information about ADNI visits http://adni.loni.usc.edu.

Our data consist of 169 subjects with an age range from 58 - 94 years old. We have information about their demographic characteristics, neuropsychological test, laboratory data, cerebrospinal fluid analysis, and brain imaging data. Figure 2 below gives an extended detail of our data.

In the cerebrospinal fluid analysis, we have a concentration of P-tau and amyloid beta levels in picograms per milliliter (pg/ml) from the cerebrospinal fluid. The laboratory data consist of the levels of vitamin B12 in nanograms per milliliter (ng/mL), thyroid stimulating hormone in milliunits per liter (mU/L), Hemoglobin in grams per deciliter (g/dL) and cholesterol in milligram per deciliter (mg/dL) as they have been linked to Alzheimer's disease.

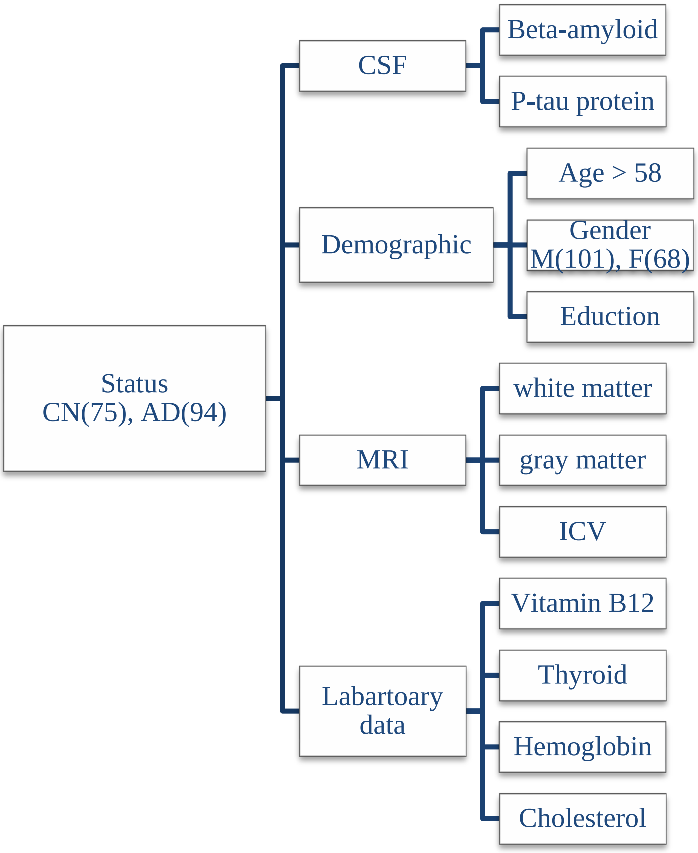

Figure 2. Schematic diagram of the data.

MRI scan includes measures about total brain volume, whole brain gray matter volume, whole brain white matter volume, and intracranial volume.

Our response in this Analysis is the status of the participants as cognitively normal individuals (CN) or Alzheimer's disease (AD) based on SPARE-AD score (Spatial Pattern of Abnormalities for Recognition of Early AD). SPARE-AD is an imaging analysis of the spatial patterns of brain atrophy to distinguish individuals with AD from CN. Positive diagnostics values indicate the presence of Alzheimer's disease and negative values indicate a normal pattern of brain structure [3] [4] [5].

Comparison of the Probability of Male and Female Diagnosed with Alzheimer's Disease

Several studies have mentioned that women are more likely than men, to be identified with Alzheimer's disease [6]. We proceed to investigate this issue by addressing the following question:

- Are male and female equality diagnosed with Alzheimer's disease?

To answer this question, we used the hypothesis test to determine whether the difference between the two proportions is significant. That is, to test the hypothesis that $H_0 : P_1 = P_2$ vs. $H_1 : P_1 \neq P_2$, where $P_1 = 0.5643 = \left(\frac{57}{101}\right)$ is the proportion of male with AD and $P_2 = 0.5441 = \left(\frac{37}{68}\right)$ is the proportion of female with AD. A *p*-value = 0.7951 indicate that at 5% level of significance, there is no statistically significant difference between the percentage of males and females diagnosed with Alzheimer's disease.

3. Statistical Method

For our analysis, we used multiple logistic regression to predict the status of the patients as CN or AD. The logistic regression is a method used to describe and explain the relationship between binary response and the statistically significant risk factors. It can answer questions like: do age, body weight, vitamin B12, cholesterol level, tau, and beta-amyloid proteins influence on the probability of having Alzheimer's disease?

Mathematically, let Y be the binary response and its possible outcome by 1 ("AD") and 0 ("CN"). The distribution of Y is specified by probability $P(Y=1)=\pi$ of AD and $P(Y=0)=(1-\pi)$ of CN, where $E(Y)=\pi$ is the mean of *Y*. Let $\pi(x)$ denote the probability of selecting AD patient given the risk factors *x*. The logistic regression model has a linear form for the logit of this probability defined as [7].

$$\operatorname{logit}\left[\pi(x)\right] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \sum \beta_j x_{ij}, \tag{1}$$

where β_j is the coefficient of the j^{th} risk factor $(j=1,\cdots,p)$, x_{ij} is the i^{th} observed value of the risk factor j $(i = 1, \cdots, n)$ and $\left(\frac{\pi(x)}{1-\pi(x)}\right)$ is the odds which expresses the ratio between the probability of predicting AD patient to the probability of CN.

The logistic regression model implies the analytic for the probability of selecting AD patient given by the risk factors as:

$$\pi(x) = \frac{\exp\left(\sum \beta_j x_{ij}\right)}{1+\exp\left(\sum \beta_j x_{ij}\right)}. \tag{2}$$

4. Implementation of the Multiple Logistic Model

We partition our data set into two parts training and testing with 75% and 25% of the data, respectively. We started with the full logistic regression model that includes all predictors and their possible interactions. Our logistic model with all independent variables and their possible interactions to predict whether the patient has Alzheimer's disease is given by:

$$\operatorname{logit}\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j, \tag{3}$$

where P denote the probability of selecting AD patient, β_j's denote the coefficients and X's are the risk factors and possible interactions. Using backward elimination algorithm to remove the term in the complex model that has the largest P_value and stop when any further elimination leads to poor fit. In addition to the minimum AIC (Akaike information criterion) that judges the quality of the model by how close the fitted values to the true expected values, that means, selecting the best statistical predictive model that minimize,

$$\text{AIC} = -2\ln(L) + 2k,$$

where L is the value of the likelihood and k is the number of parameters in the model. Thus, our optimal data-driven statistical logistic model that predicts the patient's condition with minimum AIC is given by:

$$\begin{aligned} \log\left[\frac{P}{1-P}\right] &= 7.55 - 0.003\ \text{Abeta} + 0.170\ \text{PTau} + 10.18\ \text{Thyroid} \\ &\quad + 0.002\ \text{VB12} - 0.14\ \text{Chelost} - 0.44\ \text{Hem} \\ &\quad + 0.01(\text{Chelost} \cap \text{Hemog}) - 0.87(\text{Thyroid} \cap \text{Hemog}) \end{aligned} \tag{4}$$

The symbol ($\cap$) means interaction and as we can see from our proposed model, six risk factors and only two interaction terms are statistically significant contributing to the prediction of the patient's condition, namely, phosphorylated tau protein (P-tau), beta-amyloid protein, thyroid stimulating hormone, vitamin B12, cholesterol, hemoglobin, and the interaction between (cholesterol $\cap$ hemoglobin) and (thyroid stimulating hormone $\cap$ hemoglobin). Furthermore, as we can see, age is not one of the significant risk factors in our optimal predictive model, and this holds that Alzheimer's disease is not part of normal aging.

The coefficients in the logistic regression indicate the change in the expected log odds relative to the one-unit change in (X_j) holding all other predictors are constant [8] [9]. Thus, the interpretation of the coefficient (0.170) of P-tau protein means as the P-tau protein level increases, the odds of the participant diagnosed with AD will increases while holding all other variables constant. Alternatively, we can use the odds ratio $\exp(0.170) = 1.85$, and that means with all other predictors unchanged, every unit increase in the P-tau protein increase the odds of being Alzheimer's patient by a factor of 1.85.

Similarly, the interpretation of the coefficient (−0.003) of beta-amyloid protein means that as the beta-amyloid protein level decrease, the odds of the participant diagnosed with AD will increase while holding all other variables constant. Alternatively, by using the odds ratio $\exp(-0.003) = 0.997$, with all other predictors unchanged, every unit decrease in the beta-amyloid protein increases the odds of being Alzheimer's patient by a factor of 0.997.

Model Evaluation

To evaluate our optimal predictive model, we used classification accuracy, sensitivity, specificity values and area under the curve (AUC) for testing data. The proportions of correctly identified AD and CN participants from the multiple logistic model is called "accuracy". The proportions of actual Alzheimer's patients who are correctly identified from our predictive model as having the disease is known as "sensitivity" and the proportions of actual cognitively normal individuals who are correctly identified from the model is known as "specificity". A perfect predictive model would be described as 100% sensitive (that is predicting all sick people from Alzheimer's disease group as Alzheimer's) and 100% specific (that is predicting all normal individual as cognitively normal). For any test, however, there is usually a trade-off between these two measures and can be explored graphically by the receiver operating characteristic curve (ROC).

We used the confusion matrix of the testing data to get the values needed to assess the model. The confusion matrix is a classification table describe how well our multiple logistic regression model does in predicting Alzheimer's patients from cognitively normal individuals. Table 1 shows an illustration of a confusion matrix that we used to evaluate our proposed model on the test data. The four outcomes that formulated the table are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP is the number of Alzheimer's patients correctly identified as sick, and TN is the number of normal individuals correctly classified as healthy. FP is the number of healthy people incorrectly

Table 1. The confusion matrix.

		Actual class		Total
		CN	AD	
Predicted class	CN	$TN = 10$	$FN = 5$	15
	AD	$FP = 2$	$TP = 18$	20
Total		$N = 12$	$P = 23$	35

identified as sick, and *FN* is the number of Alzheimer's cases predicted incorrectly by our model as a healthy individual.

Using the confusion matrix, we found out that our model accuracy is $\left(\frac{TP+TN}{N+P}\right)=80\%$ and it correctly predicts 78.26% of all Alzheimer's disease cases (the sensitivity = $\left(\frac{TP}{P}\right)$). Also, it correctly identifies 83.33% of those who don't have Alzheimer's disease (the specificity = $\left(\frac{TN}{N}\right)$). A summary of our classification results is given in **Table 2** below.

Another method to evaluate our model graphically is the receiver operating characteristic (ROC). Each point on the ROC curve represents a (sensitivity, 1-specificity) pair corresponding to a different decision cut-off point. The area under the ROC curve (AUC) is a measure of how well the model can distinguish between two diagnostic groups. For our proposed model, the AUC value is 87.68% which implies that our model does well in discriminating between the two classes of the patient's condition. **Figure 3** represents the receiver operating characteristic curve with the corresponding AUC value. After a careful investigation of our results, we can conclude that our predictive model provides a good prediction of the patient's condition.

Table 2. Classification summary of the multiple logistic regression model.

Evaluation value	Percentage
Accuracy	80%
Sensitivity	78.26%
Specificity	83.33%

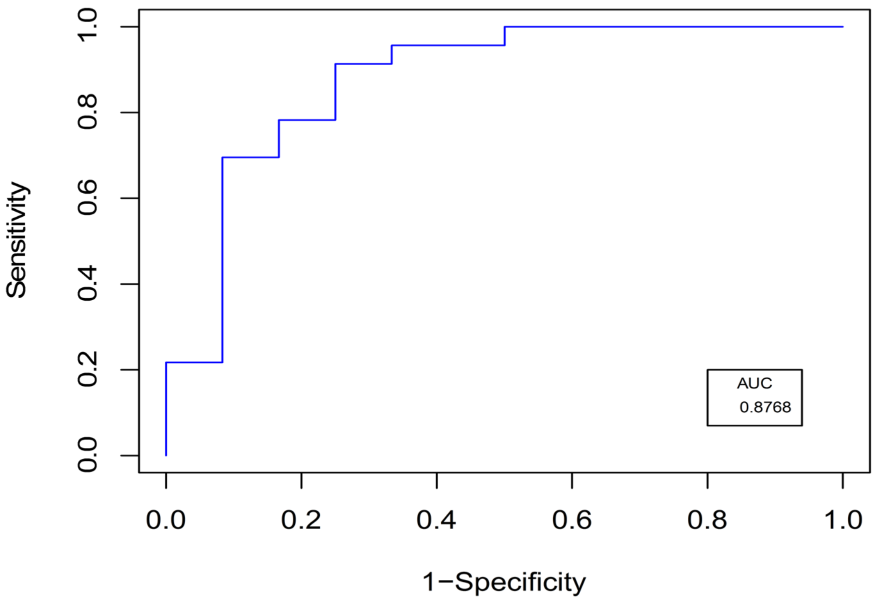

Figure 3. The receiver operating characteristic curve.

After validating our proposed model, we need to rank the risk factors in terms of their importance to Alzheimer's diagnostic. We identified the relative importance of the risk factors by the absolute value of their standardized coefficients (weights) and pseudo partial correlation. In the standardized coefficients, the higher the absolute value points to the greater strength of association with Alzheimer's diagnostic [10] [11]. The standardized weight is defined as:

$$\text{Standardized weight} = \frac{\beta_i}{s/sd_i}, \tag{5}$$

where β_i is the estimated coefficient (weight) for predictor *i*, sd_i is the sample standard deviation for predictor *i*, and $s = \pi/\sqrt{3}$.

The pseudo partial correlation is given by:

$$r = \pm\sqrt{(W_i - 2K)/-2LL_0} \tag{6}$$

where W_i is the Wald chi-square statistic for predictor *i*, *K* is the degrees of freedom of predictor *i*, and $-2LL_0$ is the log-likelihood of the model with only intercept term. The closer the value to 1 or −1, the stronger the association between a predictor and the outcome [12].

Thus, the relative importance of the significantly contributing risk factors in our predictive model is presented in **Table 3**. As can be seen, the result of the two methods is consistent, and we found out that P-tau protein is the most critical factor in diagnosing with Alzheimer's disease followed by beta-amyloid. These two proteins have been extensively studied by the author [13]. Also, the interaction between (thyroid $\cap$ hemoglobin) is ranked as number three significant predictor before the level of thyroid hormone alone and hemoglobin alone which they ranked as number 4th and number 8th significant risk factors, respectively.

Table 3. Relative importance of the risk factors.

Rank	Risk Factor	Standardized Weights	Pseudo Partial Correlation
1	P-Tau protein	4.384	0.542
2	Beta-amyloid	3.568	−0.410
3	Thyroid $\cap$ Hemoglobin	2.514	−0.243
4	Thyroid	2.171	0.212
5	Vitamin B12	1.665	0.196
6	Cholesterol	1.554	−0.154
7	Cholesterol $\cap$ Hemoglobin	1.496	0.147
8	Hemoglobin	0.349	−0.019

5. Conclusions

The importance of knowing the causes of the disease helps find the best way to cure it. While several top causes of death are decreasing, Alzheimer's deaths are on the rise. Thus, in the present study, we developed a statistical predictive model using multiple logistic regression to predict Alzheimer's disease patients by selecting the relevant risk factors using backward elimination. We found that six risk factors and only two interaction terms namely, phosphorylated tau protein (P-tau), beta-amyloid protein, thyroid stimulating hormone, vitamin B12, cholesterol, and the interaction between (cholesterol ⋂ hemoglobin) and (thyroid stimulating hormone ⋂ hemoglobin) were significantly contributing to Alzheimer's disease.

We evaluated the quality of the proposed model by classification accuracy, sensitivity, specificity values and area under the curve, the result of which attested to the effectiveness of the model. Then, we examine the relationship between the response and the significant contributing predictors and rank them based on their standardized coefficients. By defining and ranking the statistically significant risk factors, they will be useful as a screening tool to discriminate Alzheimer's disease patients from cognitively normal individuals.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Alzheimer's Association (2018) 2018 Alzheimer's Disease Facts and Figures. Includes a Special Report on the Financial and Personal Benefits of Early Diagnosis. *Alzheimer's & Dementia*, **14**, 367-429. https://doi.org/10.1016/j.jalz.2018.02.001

[2] Alzheimer's Statistics—United States & Worldwide Stats. Available: https://braintest.com/alzheimers-statistics-throughout-the-united-states-and-worldwide/

[3] Davatzikos, C. and Da, X. (2013) Spare-Mci Scores from Upenn/Sbia: MRI-Based Biomarker of Conversion from MCI to AD. 3-5.

[4] Davatzikos, C., Xu, F., An, Y., Fan, Y., and Resnick, S.M. (2009) Longitudinal Progression of Alzheimer's-Like Patterns of Atrophy in Normal Older Adults: The Spare-AD Index. *Brain*, **132**, 2026-2035. https://doi.org/10.1093/brain/awp091

[5] Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N. and Trojanowski, J.Q. (2011) Prediction of MCI to AD Conversion, via MRI, CSF Biomarkers, and Pattern Classification. *Neurobiology of Aging*, **32**, 2322.e19-2322.e27.

[6] Chapman, R.M., *et al.* (2011) Women Have Farther to Fall: Gender Differences between Normal Elderly and Alzheimer's Disease in Verbal Memory Engender Better

Detection of Alzheimer's Disease in Women. *Journal of the International Neuropsychological Society*, **17**, 654-662.
https://doi.org/10.1017/S1355617711000452

[7] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. Third Edition, John Wiley & Sons, Inc., Hoboken.
https://doi.org/10.1002/9781118548387

[8] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017) An Introduction to Statistical Learning with Applications in R. Springer, Berlin.

[9] Agresti, A. (2007) An Introduction to Categorical Data Analysis. Second Edition, Wiley, Hoboken.

[10] Zhang, D. (2018) Package 'Rsq' Title R-Squared and Related Measures.

[11] Thompson, D., Wi, M. and Health, A. (2009) LR, Ranking Predictors in Logistic Regression. Paper D10-2009, Assurant Health, West Michigan, 1-13.
http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Ranking+Predictors+in+Logistic+Regression#9

[12] Bhatti, I.P., Lohano, H.D., Pirzado, Z.A. and Jafri, IA. (2006) A Logistic Regression Analysis of the Ischemic Heart Disease Risk. *Journal of Applied Sciences*, **6**, 785-788.
https://doi.org/10.3923/jas.2006.785.788

[13] Habadi, M. and Tsokos, C.P. Alzheimer's: Probabilistic Approach to the Behavior of Beta-Amyloid and Phosphorylated Tau Proteins Levels. 1-23.