International Journal of Community Medicine and Public Health
Krishnaiah S et al. Int J Community Med Public Health. 2015 May;2(2):137-148
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: 10.5455/2394-6040.ijcmph20150514

Research Article

Modeling the risk of age-related macular degeneration and its predictive comparisons in a population in South India

Sannapaneni Krishnaiah[1]*, Bapi Raju Surampudi[2,3], Jill Keeffe[4]

[1]Health Promotion Division, Public Health Foundation of India, Delhi NCR, Gurgaon-22003, Haryana, India
[2]School of Computer and Information Sciences, University of Hyderabad, Hyderabad-500046, Telangana, India
[3]International Institute of Information Technology (IIIT), Hyderabad-500032, Telangana, India
[4]Department of Ophthalmology, University of Melbourne, Melbourne-8002, VIC, Australia

Received: 12 February 2015
Accepted: 21 March 2015

***Correspondence:**
Dr. Sannapaneni Krishnaiah,
E-mail: skrishna_s@yahoo.com

ABSTRACT

Background: Objective of current study was to develop and cross-validate the prediction models for Age-related Macular Degeneration (AMD) by using Logistic Regression (LR) and Artificial Neural Networks (ANN).
Methods: A population based cross-sectional epidemiologic study. The data (n=3723) were analyzed on participants aged ≥40 years in Andhra Pradesh, South India. Sub-population data from this sample was drawn by using random under sampling and random over sampling techniques to derive a risk score from the LR model. The models were compared for their predictive abilities by an Area under the Receiver Operating Characteristic Curve (AUROC).
Results: The LR risk score was built with a score ranging from 0 to 60 for a sub-population dataset (n=213). A cut-off score of ≥30 had a sensitivity of 79% and a specificity of 69%. The predictive performance of ANN and LR was statistically equivalent (76% vs. 78%; P = 0.624). Both the models were stable and consistently obtained the same predictive accuracies in a 30-fold split-sample cross validation.
Conclusions: The sensitivity analysis of the ANN model indicated the relative importance of prioritizing modifiable risk factors for AMD in order to base preventive interventions to reduce the impact of the modifiable factors on AMD.

Keywords: Artificial neural network, Logistic regression model, Age-related macular degeneration, Population-based cross-sectional study, Risk score, South India

INTRODUCTION

Age-related Macular Degeneration (AMD) is one of the leading causes of irreversible vision loss among the elderly and is becoming a major public health problem in the developing world due to an increased size of the older population resulting from increased life expectancy. An estimated 25 to 30 million people around the world are believed to have been affected by AMD, and this number is expected to triple within 25 years. The etiology of AMD is not fully understood but is assumed to be of multi-factorial origin. Modeling the risks of AMD using the new approach is one way of finding the non-modifiable and modifiable risk factors that suggests ways of prevention strategies for delaying the onset of AMD or early diagnosis of AMD, thereby preventing the further deterioration of the visual loss due to AMD.

Many techniques have been used to derive clinical algorithms to infer the presence or severity of disease. The accuracy of a statistical method is finally reflected by its ability to predict outcome from training as well as test samples. Modeling a validated risk score based on a set of risk factors that cause AMD in the population is necessary for several reasons. The validated risk score yields a threshold score which can potentially identify

those at risk of AMD to refer high risk individuals for diagnosis and management. Artificial neural networks (ANN) have been used extensively to predict diseases, treatment outcomes and prognosis for a variety of diseases. Various architectures of ANN have been used in different medical diagnoses and their results are compared with the existing classification methods and physicians' diagnoses.[1] With the development of artificial intelligence, data-mining tools like ANN can be used to derive more nuances from patient data in predicting disease. The traditional Logistic Regression (LR) model estimates the probability to obtain a result as a function of several predictive features that are suspected to influence the outcome. The good fit of a logistic regression equation can be assessed by means of the χ^2 goodness of fit test or the Hosmer-Lemeshow statistic.

We have previously published the associations of various risk factors for occurrence of AMD using the standard multivariable techniques.[2] To the best of our knowledge, ANN models have not been established till date for predictions in the areas of ophthalmology. Therefore, the specific purpose of the present report is to train and test the ANN model including developing and validating LR risk score on persons aged 40 to 102 years in a total population (n=3723) and sub-populations drawn using random under sampling technique (n=213) and combination of random under sampling and random over sampling technique (n=1420) for the accurate prediction of AMD. The predictive performances of these two models are compared using an Area under the Receiver Operating Characteristic Curves (AUROC). A 30-fold split-sample cross validation was also done to make a reliable assessment of generalization errors in the population in South India.

METHODS

The data for this analysis were obtained from the Andhra Pradesh Eye Disease Study (APEDS) database which was a population based cross-sectional epidemiological study conducted in the South Indian state of Andhra Pradesh (AP) during 1996 and 2000. The details of the design of the APEDS and other related methods have been described earlier.[2-5] The approval of the ethics committee of the L. V. Prasad eye institute was obtained for the study design and informed consent was obtained from each participant after explaining the purpose of the study. The study was conducted in accordance with the tenets of the Helsinki declaration. A total of 10293 individuals of all ages were participated in this study from one urban and three rural areas from different parts of AP, chosen to roughly represent the population of the state. All participants were interviewed in detail and examined in a masked manner by trained professionals. A structured questionnaire was used to elicit the information on various risk factors of systemic diseases and personal habits such as smoking and chewing tobacco. The various details about the interview and ophthalmic examination procedures were published in detailed previously.[2]

In order to achieve the best performance of the LR and ANN models, in the present report, we drew a random sub-population from a total of 3723 participants ≥40 years old from the original APEDS by using the procedures "Random Under Sampling technique (RUS)" and combination of RUS and "Random Over Sampling technique (ROS)". The details of these procedures are described below.

Random under sampling and random over sampling techniques

As part of the RUS technique, the majority class is randomly under sampled by using the systematic random sampling consisting of removing the samples from the majority class population until the majority class becomes the 200% sample of the minority class sample size. In the minority class, there were 71 patients with AMD, therefore, the majority sample was reduced to the number equal to 142 participants. To select these 142 participants, systematic random sampling is used on the majority class that has not been arranged in any specific periodic order, to obtain an unbiased sample. This technique allowed for the large majority sample to be equally represented in the re-sampled dataset. This technique gives more accurate prediction models compared to the ones that have been used in medical literature till date, and to the best of our knowledge it has not been used in the academic research community till date. We have also done a combination of random under sampling and over sampling techniques by randomly selecting the neighboring similar minority cases from the minority class and with random over representing the minority class (n=710). That is from 71 in the minority class, a total of 710 cases were generated which was equal to 10 times the original minority class. From the majority class, the same sample size was reduced by systematic random sampling to make the total sample size equal to 1420 samples for this experiment.

Statistical analysis

Logistic regression modeling

The Logistic Regression (LR) model was constructed to identify the variables associated with AMD.[6,7] The model was constructed using a randomly selected sub-sample with 80% of the participants included (developmental model). The model obtained was tested with data derived from the remaining 20% of the total population (validation sample). Developmental and validation samples were created by assigning each patient a random number between 0 and 1. Patients with a random number of 0.80 or less formed the developmental sample, and the remaining patients formed the validation sample.

Developing a risk score from logistic regression model

A series of statistical analyses were conducted to develop a risk score. The associations between each risk factor

and AMD were first evaluated in bivariable analysis. Any factor with a P value ≤0.20 was eligible for addition into the multivariable logistic regression analysis. To retain the statistical power of the database, the missing data were handled using the modified 'hot deck imputation' procedure.[8] We then created a simple and logical index for risk stratification by generating a numerical weighted score by rounding all regression coefficients up to the nearest integer. This method was based on the 'β coefficients' rather than 'ORs', which can be excessively influenced by only a few factors.[9] Interaction terms between independent variables were not considered, because the model should be simple and easy to use as a model in clinical medicine. Once the final LR model was fitted, the risk score was obtained from the fitted LR model in two stages. In the first stage, the score was obtained by multiplying the model coefficients by 10. The reference category of the risk factor will be equal to '0'. In the second stage, the score was obtained by dividing the "β-coefficients" by an absolute value of the smallest coefficient in the model and rounding up to the nearest integer. Finally the total score was calculated as the sum of the risk scores for each risk factor present for a participant. However, both the two stages of obtaining risk scores yielded with same score, therefore, either one of the procedures was finally used in this analysis to obtain the risk scores.

Cross-sectional internal validation

The performance of the Logistic Regression (LR) and Artificial Neural Network (ANN) models were made according to the AUROC analysis. The larger area under the AUROC curve reflects the better performance of a diagnostic test. Sensitivity and specificity were calculated for various cut-off points of the calculated LR risk score. The LR cut-off risk score that gave the maximum sum of sensitivity and specificity was taken as an optimum. This procedure was performed for the entire sample (n=3723) and sub-populations that were drawn using RUS (n=213) and combination of RUS & ROS techniques (n=710). These score models were constructed by SPSS software. The SPSS software version 17.0 for Windows (SPSS, Chicago, IL, USA) was used for statistical analysis. A two-tailed P value of less than 0.05 was considered statistically significant.

Validation of logistic regression model using split-sample cross validation

The model should have a good accuracy for prediction for the general use of results. The method of 80-20 cross-validation was applied in this study in which 80% of the cases used to derive the model and its accuracy would be evaluated on the remaining 20% of the cases. Classification accuracy of the model as an accuracy rate for holdout sample is no more than 10% lower than the accuracy rate for the training sample. This would rule out the accurate prediction of the model.

Artificial neural network (ANN) modeling

A Multi-Layer Feed-Forward Neural Network (MLFFNN), which is a popular architecture in the ANN literature, was used in this study. The network was trained using an error back propagation training algorithm.[10] This algorithm adjusts the connection weights according to the back propagated error computed between the observed and the estimated results. This is a supervised learning procedure that attempts to minimize the error between the desired and the predicted outputs. About 90% of ANNs presented in clinical medicine are MLFFNNs.[1] For the present study, we chose the MLFFNN model as our main ANN analysis tool. For this analysis, an ANN with a hidden layer and back propagation along with momentum was used as a machine-learning method. The network used consisted of three layers: one input layer of 8 neurons (one for each input variable), one hidden layer of four neurons (it is the number which gives the best prediction result based on our empirical investigations) and one output layer of one neuron which is the output variable (Figure 1). In Figure 1, each arrow in the figure is bias weights, which is not multiplied by any incoming value.

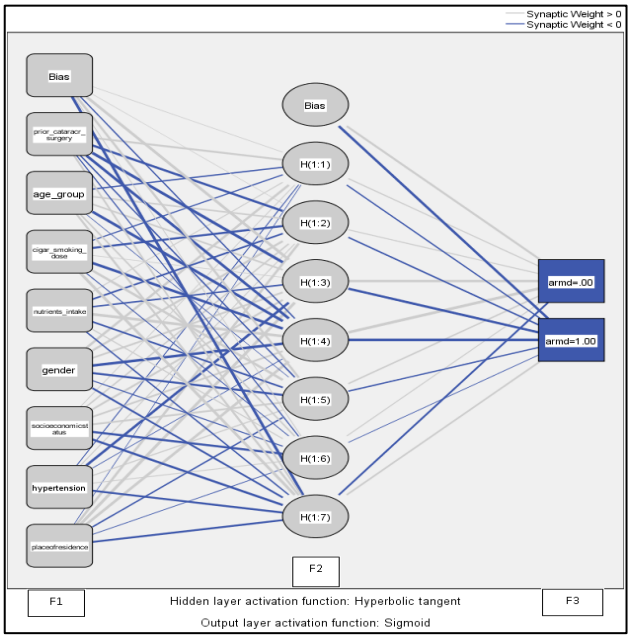

Figure 1: Structure of the neural network model used in this study. F1, input layer of neurons receiving values for independent variables such as *cigar smoking dose* (0 - Not a smoker, 1 - Light Smoker, 2 - Heavy smoker); food intake, etc.; F2, hidden layer of neurons whose number is determined empirically; F3, output layer of neurons with a single neuron corresponding to the single dependent variable. Nodes designated as *Bias* receive a constant input but whose weights are modified along with other weights. Bias node captures the mean value or baseline of the function being approximated. Figure generated from SPSS Ver 17.0 software.

Artificial Neural Network modeling was carried out in two steps: Firstly, by testing the model to calibrate the model parameters; random selection was used in the entire data sample (n=3723) to isolate the training set (70% of the records, i.e., 2606), an independent test set (15% of the records, i.e., 558) and holdout sample (15% of records, i.e., 559). The model was first adjusted with the training set and then tested with the test set by holding the holdout sample to determine the best ANN configuration by using the entire data sample (n= 3723).[11] Secondly, applying the methods used to study the contribution of the different variables at the input stage on the already calibrated ANN model (during the first step) by using the whole data set. The same procedure was repeated for the sub-populations with samples (n=213) and (n=1420), however, the ANN model in this situation was carried out using 75% (training), 15% (testing) and 10% (holdout) samples. During the training of MLFFNN, the corresponding known outputs of the system were held in the output nodes and compared with the actual outputs produced by the network. The nodes in the hidden layer had no prescribed initial values and helped to allow complex relationships between the input and output nodes to evolve. Information was transported from the input layer to the output layer by calculating the weighted sum at each node, which was derived by combining outputs of all the nodes in the previous layer. A non-linear activation function embedded in the nodes allowed them to learn non-linear relationships. This flexibility was useful when trying to learn complex relationships between biological features and clinical outcomes. In this study, sigmoid activation function (Φ) was used. Sigmoid function has a linear region over a small range of input variable values close to '0', but saturates to '1' for large values & to '0' for small values. In equation - 1, *v* is sum on node *j* for case *n*, nodes connecting *j* from previous layer *a* constant value of >0.

$$\emptyset\left(v_j(n)\right) = \frac{1}{1+\exp(-av_j(n))} \qquad (1)$$

Once the values on the output node had been calculated, they were compared with the desired values and a back propagation algorithm was used to adjust the weights to decrease the difference between the actual and desired predictions. This process was repeated iteratively using all cases in the training set until it met the least Mean Square Error (MSE) between the target and actual output values (equation 2).

$$MSE = \sqrt{\sum_{i=1}^{n}(O-T)^2} \qquad (2)$$

Where, *O* is actual outcome and *T* predictive outcomes, *i* is the number of patients.

Calculation of relative importance

In order to find which independent variables have greater impact on predicting the dependent variable, we used a built-in procedure available in SPSS ver 17.0; ANN module) called "In put variable Importance". The importance value measures how much the ANN model - predicted output value changes for different values of the independent variable. Normalized importance is calculated by dividing the importance value by the largest value and expressed as percentage. The order of priority of each predictor in the form of relative importance is clearly understood with the use of the ANN model.

Model stability measurement

In order to check the stability of the ANN and LR methods, we have run the 30-fold split sample cross validation of the network and noted the relative contributions and normalized importance of the input variables on the output obtained for each method and each trained network. We then calculated the mean contribution of each variable for the different methods. The thirty training sessions allowed us to draw the Standard Error (SE) which gives an indication of the stability of this procedure.

RESULTS

Participants

The data (n=3723) were analyzed for ≥40 years age group. There was no significant difference found between participants and non-participants of the study. Details of various risk factors (input variables) and their association with AMD in the entire data (n=3723) (Table 1).

Logistic regression model

The adjusted LR model identified eight independent variables that associated with AMD in the training set. These eight variables were: age, gender, socioeconomic status, place of residence, prior cataract surgery, intake of nutrients, cigar smoking dose, and hypertension. The univariable associations of these eight risk factors for the subpopulation of data (n=213) drawn using RUS technique are shown in Table 2. The final logistic regression model that was used to derive the risk scores utilized these eight variables with a risk score assigned for each of this risk factor. These results are shown in Table 3. The Hosmer and Lemeshow goodness of fit statistic (Chi-square = 12.4; P = 0.133) for this multivariable LR model confirmed the good fit of the model in the prediction of AMD.

Considering the sensitivity and specificity of the AUROC analysis in respect of different risk values, a risk score of >30 with sensitivity of 79% and specificity of 69% selected as cut point. Therefore, if the risk score for an adult patient is greater than 30, the patient is more susceptible to AMD. Table 4 depicts the data on sensitivity, specificity, positive predictive value and negative predictive values for each cut-off of the LR risk scores.

Table 1: Details of various risk factors (input parameters) for AMD in entire sample (N = 3724).

Risk factors	Total sample	AMD N (%)	P value
Age (years)			
40-49	1424	13 (0.9)	<0.001
50-59	1047	14 (1.3)	
60-69	899	31 (3.4)	
≥70	353	13 (3.7)	
Gender			
Male	1751	31 (1.8)	0.632
Female	1972	40 (2.0)	
Socioeconomic status			
Extreme lower	417	4 (1.0)	0.103
Lower	1791	43 (2.4)	
Middle	1317	19 (1.4)	
Upper	145	2 (1.4)	
Place of residence			
Urban	934	19 (2.0)	0.782
Rural	2789	52 (1.9)	
High blood pressure			
No	1910	29 (1.5)	0.150
Yes	1813	41 (2.3)	
Nuclear cataract[$]			
No	2661	32 (1.2)	<0.001
Yes	895	29 (3.2)	
Cortical cataract[$]			
No	3022	39 (1.3)	<0.001
Yes	526	21 (4.0)	
Posterior subcapsular cataract[$]			
No	3002	46 (1.5)	0.071
Yes	548	15 (2.7)	
Prior cataract surgery			
No	3447	50 (1.5)	<0.001
Yes	276	21 (7.6)	
Any cataract[§]			
No	2271	22 (1.0)	<0.001
Yes	1442	49 (3.4)	
BMI			
Normal	1293	26 (2.0)	0.078
Underweight	1816	37 (2.0)	
Overweight	312	-	
Obese	153	4 (2.6)	
Cigarette smoking			
Never a smoker	3419	64 (1.9)	0.866
Current smoker	179	4 (2.2)	
Prior smoker	125	3 (2.4)	
Cigar smoking			
Never a smoker	3224	53 (1.6)	<0.001
Current smoker	362	10 (2.8)	
Prior smoker	137	8 (5.8)	
Alcohol consumption[£]			
Never a drinker	2658	61 (2.3)	0.023
Light drinker	805	8 (1.0)	
Heavy drinker	260	2 (0.8)	
Food intake[¶]			
Below recommended intake	1817	48 (2.6)	0.002
Above recommended intake	1906	23 (1.2)	

Table 2: Univariable analysis for associations between potential risk factors and AMD among sample drawn using random under sampling technique (n = 213).

Risk factors	Total sample	AMD N (%)	P value
Age (years)			
40-49	71	13 (18.3)	0.003
50-59	46	14 (30.4)	
60-69	66	31 (47.0)	
≥70	30	13 (43.3)	
Gender			
Male	93	31 (33.3)	1.000
Female	120	40 (33.3)	
Socioeconomic status			
Extreme lower	17	4 (23.5)	0.108
Lower	107	43 (40.2)	
Middle	79	19 (24.1)	
Upper	6	2 (33.3)	
Place of residence			
Urban	50	19 (38.0)	0.493
Rural	163	52 (31.9)	
High blood pressure			
No	106	30 (28.3)	0.146
Yes	107	41 (38.3)	
Nuclear cataract[$]			
No	140	32 (22.9)	<0.001
Yes	56	29 (51.8)	
Cortical cataract[$]			
No	154	39 (25.3)	0.002
Yes	41	21 (51.2)	
Posterior subcapsular cataract[$]			
No	164	46 (28.0)	0.063
Yes	33	15 (45.5)	
Prior cataract surgery			
No	188	50 (26.6)	0.014
Yes	25	21 (84.0)	
Any cataract[§]			
No	114	22 (19.3)	<0.001
Yes	99	49 (49.5)	
BMI			
Normal	79	26 (32.9)	0.031
Underweight	100	37 (37.0)	
Overweight	16	0 (0.0)	
Obese	10	4 (40)	
Cigarette smoking			
Never a smoker	197	64 (32.5)	0.641
Current smoker	5	2 (40.0)	
Prior smoker	11	5 (45.5)	
Cigar smoking			
Never a smoker	176	53 (30.1)	0.010
Current smoker	12	3 (25.0)	
Prior smoker	25	15 (60.0)	
Alcohol consumption[£]			
Never a drinker	162	61 (37.7)	0.058
Light drinker	40	8 (20.0)	
Heavy drinker	11	2 (18.2)	
Food intake[¶]			
Below recommended intake	117	48 (41.0)	0.009
Above recommended intake	96	23 (24.0)	

Symbolic details of Table 1 and 2 were given as below:

[§]: Includes significant nuclear, cortical, or posterior subcapsular cataract and/or history of prior cataract surgery and/or total cataract. [$]: Missing data on lens opacities is due to the presence of prior cataract surgery (pseudophakia or aphakia) or total cataract, the presence of phthisis bulbi, or the pupil not being dilated due to the risk of angle closure.

[£]: A person who drinkes alcohol for at least 3 to 4 days a week, 5 to 6 days a week, and/or everyday was considered to be heavy drinker.

[η]: Those taken food (cereals, pulses, green leafy vegetables, milk products, fruits and milk products) below the recommended levels considered to be the ones consumed food below the recommendation.

Table 3: Multivariable Logistic Regression analysis for associations between potential risk factors and AMD among sample drawn using random under sampling technique (n=213).

Risk Factors	Total sample N	AMD N (%)	'B' coefficient	Odds ratio (95% CI)	Risk score
Age group (years)					
40-49	71	13 (18.3)	--	1.00	0
50-59	46	14 (30.4)	0.553	1.74 (0.66, 4.56)	6
60-69	66	31 (47.0)	1.195	3.30 (1.38, 7.94)	12
≥70	30	13 (43.3)	0.408	1.50 (0.60, 4.92)	4
Gender					
Female	120	40 (33.3)	0.402	1.49 (0.71, 3.13)	4
Area					
Urban	50	19 (38.0)	0.963	2.62 (1.13, 6.06)	10
Socioeconomic status					
Extreme lower + lower	124	47 (38.0)	1.156	1.94 (0.96, 3.94)	12
High blood pressure	107	41 (38.3)	0.412	1.51 (0.77, 2.96)	4
Prior cataract surgery[§]	34	21 (61.8)	1.451	4.27 (1.61, 11.27)	15
Cigar smoking dose[ৱ]					
Light	12	3 (25.0)	-0.864	0.42 (0.08, 2.01)	0
Heavy	25	15 (60.0)	1.353	3.87 (1.36, 10.99)	14
Food Intake below recommended[η]	117	48 (41.0)	1.156	3.18 (1.52, 6.65)	12

[§]: Replaced in the multivariable logistic regression model that adjusted for confounding variables such as age, cigar smoking, alcohol consumption and food intake.

[ৱ]: Those above the 25th percentile of pack years smoked were considered to be heavy smokers.

[£]: A person who drinkes alcohol for at least 3 to 4 days a week, 5 to 6 days a week, and/or everyday was considered to be heavy drinker.

[¥]: Abnormal body mass index includes participants of underweight, overweight and obese.

[η]: Those food consumed daily (cereals, pulses, green leafy vegetables, milk products, fruits and milk products) below the recommended levels considered to be the ones consumed food below the

Neural network model

The significant predictors that were identified in the LR model were also studied using the ANN model and these results are shown in Tables 5 & 6. Figure 1 shows a schematic diagram of the ANN model. As we were interested in identifying which input parameters are most important in predicting the outcome of AMD, the sensitivity analysis function was set. The validation of ANN model revealed that the model is correct in three out of four times (Tables 5a and 6a). The ANN model (Table 5b) revealed that daily food intake below the recommendation, history of prior cataract surgery, history of heavy cigar smoking, increased age, female gender, urban place of residence, presence of hypertension and belonging to extreme lower and lower socioeconomic status were in order of priority and significantly contributed to the increased risk of AMD. When we trained, tested and validated the ANN model for the entire data (n=3723), the order of importance of the

variable in predicting AMD were history of prior cataract surgery (normalized importance: 100%), heavy cigar smoking (43.8%), food intake below the recommendation (39%), belonging to extreme lower and lower socioeconomic status (37.6), being in an urban area (34.1%), increased age (31.6%), hypertension (29%), and female gender (12.2%) (Data not shown). However, the predictive accuracy of validation of the model in this instance was not satisfactory (AUROC: 66%) (Figure 2a). The learning and training including the validation of ANN model on the data (n=1420) generated using the combination of under and over sampling method has also shown more or less similar results (data not shown).

Table 4: Sensitivity, specificity, positive predictive value, negative predictive value at different risk score cut points, among the sample drawn using RUS technique for AMD predictions (n=213).

Criterion	Sen	95% CI	Spe	95% CI	+PV	95% CI	-PV	95% CI
≥0	100.00	94.7- 100.0	0.00	0.0 - 2.6	32.5	26.2 - 39.3		
>0	100.00	94.7 - 100.0	2.13	0.4 - 6.1	33.0	26.6 - 39.9	100	29.2 - 100.0
>4	100.00	94.7 - 100.0	5.67	2.5 - 10.9	33.8	27.3 - 40.8	100	63.1 - 100.0
>6	100.00	94.7 - 100.0	6.38	3.0 - 11.8	34.0	27.5 - 41.0	100	63.1 - 100.0
>8	100.00	94.7 - 100.0	8.51	4.5 - 14.4	34.5	27.9 - 41.6	100	73.5 - 100.0
>10	100.00	94.7 - 100.0	11.35	6.6 - 17.8	35.2	28.5 - 42.4	100	79.4 - 100.0
>12	98.53	92.1 - 100.0	15.60	10.0 - 22.7	36.0	29.1 - 43.4	95.7	77.4 - 99.9
>14	95.59	87.6 - 99.1	19.15	13.0 - 26.6	36.3	29.3 - 43.8	90.0	73.1 - 98.0
>16	94.12	85.6 - 98.4	26.95	19.8 - 35.1	38.3	30.9 - 46.2	90.5	77.4 - 97.3
>18	92.65	83.7 - 97.6	30.50	23.0 - 38.8	39.1	31.5 - 47.1	89.6	77.2 - 96.6
>20	91.18	81.8 - 96.7	36.88	28.9 - 45.4	41.1	33.1 - 49.3	89.7	78.8 - 96.1
>22	88.24	78.1 - 94.8	41.13	32.9 - 49.7	42.0	33.8 - 50.5	87.9	77.5 - 94.6
>23	88.24	78.1 - 94.8	41.84	33.6 - 50.4	42.3	34.0 - 50.9	88.1	77.8 - 94.7
>24	86.76	76.4 - 93.8	46.10	37.7 - 54.7	43.7	35.2 - 52.5	87.8	78.1 - 94.3
>26	86.76	76.4 - 93.8	51.06	42.5 - 59.6	46.1	37.2 - 55.1	88.9	79.9 - 94.8
>28	82.35	71.2 - 90.5	62.41	53.9 - 70.4	51.4	41.6 - 61.1	88.0	80.0 - 93.6
>29	82.35	71.2 - 90.5	63.12	54.6 - 71.1	51.9	42.0 - 61.6	88.1	80.2 - 93.7
>30	79.41	67.9 - 88.3	66.67	58.2 - 74.4	53.5	43.3 - 63.5	87.0	79.2 - 92.7
>32	70.59	58.3 - 81.0	73.05	64.9 - 80.2	55.8	44.6 - 66.6	83.7	76.0 - 89.8
>34	66.18	53.7 - 77.2	77.30	69.5 - 83.9	58.4	46.6 - 69.6	82.6	75.0 - 88.6
>35	64.71	52.2 - 75.9	78.01	70.3 - 84.5	58.7	46.7 - 69.9	82.1	74.5 - 88.2
>36	58.82	46.2 - 70.6	80.14	72.6 - 86.4	58.8	46.1 - 70.7	80.1	72.6 - 86.4
>37	58.82	46.2 - 70.6	80.85	73.4 - 87.0	59.7	47.0 - 71.5	80.3	72.7 - 86.5
>38	48.53	36.2 - 61.0	82.98	75.7 - 88.8	57.9	44.1 - 70.9	77.0	69.5 - 83.4
>39	45.59	33.5 - 58.1	85.82	78.9 - 91.1	60.8	46.0 - 74.3	76.6	69.2 - 82.9
>40	39.71	28.0 - 52.3	88.65	82.2 - 93.4	62.8	46.7 - 77.0	75.3	68.0 - 81.7
>41	38.24	26.7 - 50.8	89.36	83.1 - 93.9	63.4	46.9 - 77.9	75.0	67.7 - 81.3
>42	32.35	21.5 - 44.8	91.49	85.6 - 95.5	64.7	46.5 - 80.3	73.7	66.5 - 80.1
>43	30.88	20.2 - 43.3	92.20	86.5 - 96.0	65.6	46.8 - 81.4	73.4	66.3 - 79.8
>44	29.41	19.0 - 41.7	95.04	90.0 - 98.0	74.1	53.3 - 89.1	73.6	66.6 - 79.9
>47	27.94	17.7 - 40.1	96.45	91.9 - 98.8	79.2	57.8 - 92.9	73.5	66.5 - 79.7
>48	26.47	16.5 - 38.6	97.16	92.9 - 99.2	81.8	59.7 - 94.8	73.3	66.3 - 79.5
>49	19.12	10.6 - 30.5	97.16	92.9 - 99.2	76.5	50.1 - 93.2	71.4	64.4 - 77.6
>50	17.65	9.5 - 28.8	97.16	92.9 - 99.2	75.0	47.6 - 92.7	71.0	64.0 - 77.3
>51	14.71	7.3 - 25.4	97.16	92.9 - 99.2	71.4	41.9 - 91.6	70.3	63.3 - 76.6
>52	13.24	6.2 - 23.6	97.16	92.9 - 99.2	69.2	38.6 - 90.9	69.9	62.9 - 76.2
>53	11.76	5.2 - 21.9	97.16	92.9 - 99.2	66.7	34.9 - 90.1	69.5	62.6 - 75.9
>54	8.82	3.3 - 18.2	99.29	96.1 - 100.0	85.7	42.1 - 99.6	69.3	62.4 - 75.6
>57	4.41	0.9 - 12.4	99.29	96.1 - 100.0	75.0	19.4 - 99.4	68.3	61.4 - 74.6
>59	1.47	0.04 - 7.9	100.00	97.4 - 100.0	100.0	2.5 - 100.0	67.8	61.0 - 74.1
>61	0.00	0.0 - 5.3	100.00	97.4 - 100.0			67.5	60.7 - 73.8

Sen: Sensitivity; Spe: Specificity; +PV: Positive predictive value; -PV: Negative predictive value; 95% CI: 95% Confidence Intervals

Table 5 (a): Classification results for most successful experiment of ANN model (n=213).

Sample	Observed	Predicted		Percent correct
		No	Yes	
Training	No	92	20	82.1
	Yes	26	31	54.4
	Overall %	69.8	30.2	72.8
Testing	No	18	4	81.8
	Yes	4	7	63.6
	Overall %	66.7	33.3	75.8
Holdout	No	8	0	100
	Yes	1	2	66.7
	Overall %	81.8	18.2	90.9

Table 5 (b): Sensitivity analysis of the input variables for the most successful calculation with the ANN model (n=213).

Importance priority	Input variables	Relative importance	Normalized importance (%)
1	Food intake below recommendation	0.189	100
2	Prior history of cataract Surgery	0.180	95.0
3	History of heavy cigar smoking	0.177	93.5
4	Age	0.167	88.2
5	Gender	0.119	38.2
6	Place of residence	0.072	32.8
7	Hypertension	0.059	30.9
8	Extreme lower & lower SES	0.036	19.0

SES: Socioeconomic status

Comparison of neural networks and logistic regression

Figures 2a and 2b and figures 3a and 3b present the AUROC curves for the ANN and LR models respectively based on the entire population and sub-population data analyses (n=3723 & n=213) respectively. In the analysis of using the complete population (n=3723), the LR model outperformed the ANN model (AUROC = 76% vs. 66%; P <0.0001) (Figures 2a & 2b). However, in the analysis of using sub-population (n=213), the performance of the ANN and LR models were statistically equivalent (AUROC: 0.76 vs. 0.78; P = 0.624) (Figures 3a & 3b). The ANN model, however, outperformed the LR model in a sub-population drawn using the combination of RUS and ROS techniques and the predictive ability was statistically significant between these two models (AUROC = 89% vs. 79%; p < 0.0001) (Figures 4a & 4b).

Thirty-fold split sample cross-validation

To evaluate the efficacy of the prediction of AMD by the ANN as well as the LR models, we have trained and tested the model thirty consecutive times with inputs of eight predictors. Tables 6a, 6b, and 7 show these results of the model stability measurements. Both the models that were trained and tested showed a good model stability (Tables 6a, 6b, and 7). When we ran a 30-fold of the 80-20 split sample cross validation of LR model, the model has obtained better stability with the mean (SD) of training and holdout samples; 70.5 (2.0) and 67.4 (1.4) respectively.

Table 6 (a): Sensitivity analysis of the input variables for the most successful calculation with the ANN model (n=213).

Training Mean ± SD	Testing Mean ± SD	Holdout Mean ± SD
72.9 ± 5.9	72.9 ± 6.1	64.7 ± 9.9

Table 6 (b): Sensitivity analysis of the input variables for the most successful calculation with the ANN model (n=213).

Importance priority	Input variables	Mean ± SD	Standard error	Average normalized Importance (%)
1	Food intake below recommended	0.127 ± 0.037	0.007	58.7
2	Prior history of cataract surgery	0.183 ± 0.052	0.009	83.0
3	History of heavy cigar smoking	0.172 ± 0.053	0.009	77.6
4	Age	0.170 ± 0.049	0.009	78.1
5	Gender	0.079 ± 0.023	0.004	36.6
6	Place of residence	0.111 ± 0.035	0.006	50.7
7	Hypertension	0.067 ± 0.030	0.006	30.6
8	Extreme lower & lower SES	0.089 ± 0.040	0.007	40.8

SES: Socioeconomic status

Table 7: Classification results for most successful experiment of LR model (n = 213).

Sample	Observed	Predicted AMD		Percent correct
		No	Yes	
Training AMD	No	86	14	86
	Yes	31	31	50
				72.2
Holdout AMD	No	29	13	69
	Yes	3	6	66.7
				68.6

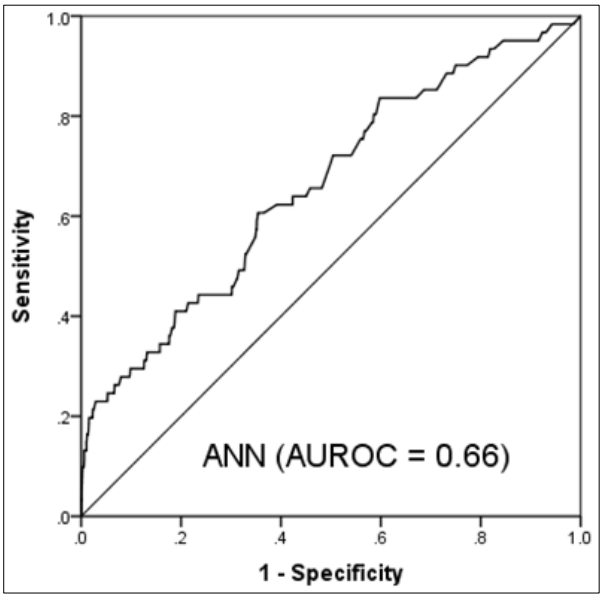

Figure 2a: AUROC curve for ANN model in predicting AMD in the entire sample (n=3723).

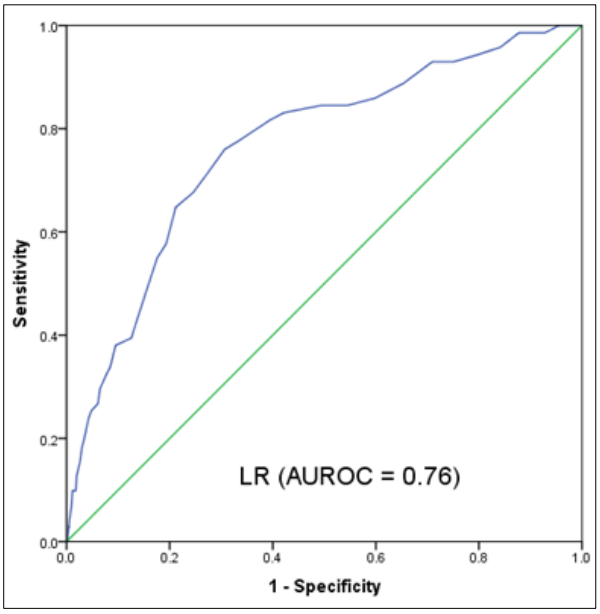

Figure 2b: AUROC curve for LR model in predicting AMD in the entire sample (n=3723).

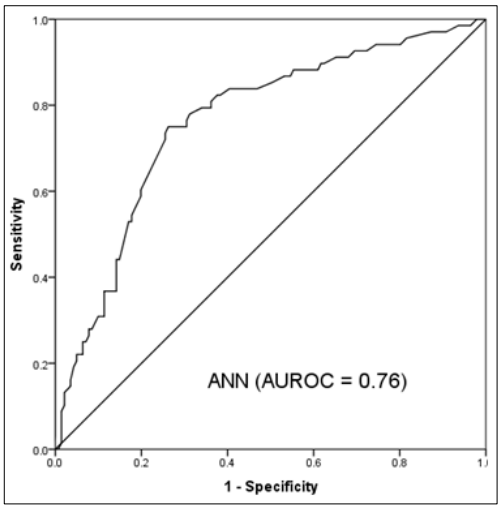

Figure 3a: AUROC curve for ANN model in predicting AMD in a subpopulation (n=213).

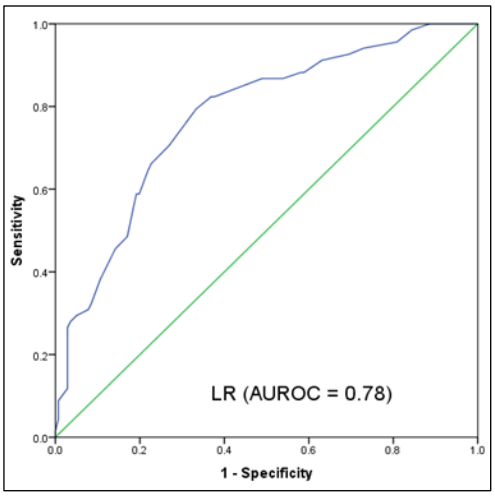

Figure 3b: AUROC curve for LR model in predicting AMD in a subpopulation (n = 213).

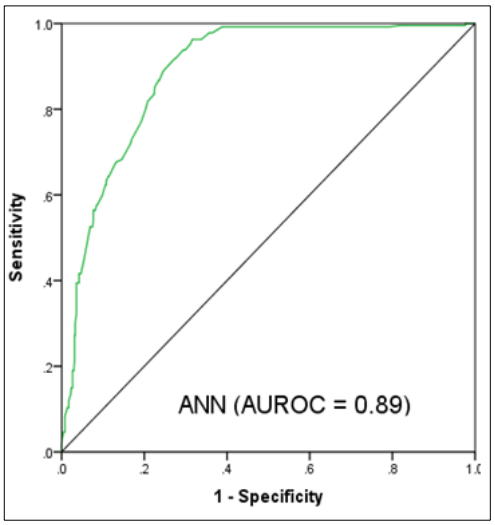

Figure 4a: AUROC curve for ANN model in predicting AMD in a sub-population (n=1420).

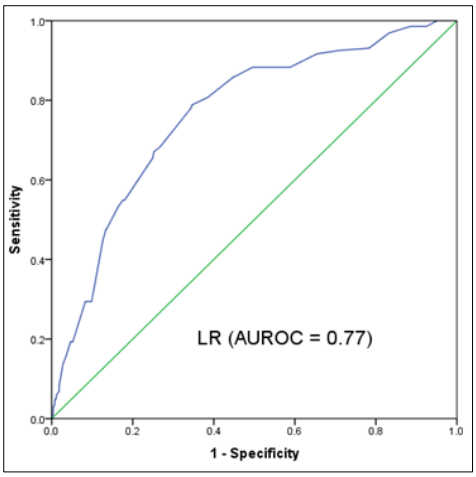

Figure 4b: AUROC curve for LR model in predicting AMD in a sub-population (n=1420).

DISCUSSION

An improved understanding of the various factors predictive of the development of AMD will take an increasing importance when developing the new therapeutic and novel strategies for the prevention and/or further deterioration of vision loss due to AMD.

The LR model is the most commonly used method to analyse the ecological data because of its capacity to give predictive and explanatory results. However, its incapability of taking into account of non-linear relationships between the dependent variable and each independent variable is its principal drawback. That is why the use of artificial intelligent methods such as ANN model is wholly justified and is becoming popular in ecology where the relationships between dependent and independent variables are principally non-linear. The ANN model in this work has been developed with the purpose to clarify the 'black-box' approach. The ANN models are able to make perfect predictions and thus are more powerful in many fields including health research, because it is an assumption free modeling approach. The ANN model has the opportunity to work on methods such as sensitivity analysis to add power to ANNs in their explanatory capacity.[12] A total of eight variables were used as input against which a hidden layer of two neurons were developed in this work. The ANN model which was utilized has been shown to be satisfactory in performance because of its capacity in identifying and learning the complex relationships among response variables and output variable AMD. Our analysis identified the modifiable risk factors predicted by the ANN model and their importance in order were history of heavy cigar smoking, daily food intake below the recommendation and presence of hypertension. The non-modifiable factors in order of priority were history of prior cataract surgery, belong to extreme lower and lower socioeconomic status, living in an urban area, increased age and female gender. However, female gender and urban place of residence were relatively less important as learnt from ANN model.

It is of great importance that developing a risk model for prediction of AMD will be useful in clinical practice for two important reasons: First, there will be as few factors as possible that must be remembered when the score is applied to a patient. Second, the recorded risk factors must be easily transferable to the patient's individual risk. This will in a way help the clinician in early diagnosis of the patient with AMD and therefore prevent further visual loss in the patient by appropriate timely and cost effective intervention. The accuracy of the risk estimates is likely to be reduced when they are not calculated directly from LR models, particularly when continuous variables form part of the input data for the model. In the present model we developed, however, all the categories which had a significant impact on outcome were nominal or discrete variables, therefore, minimizing the loss of information and precision in this way. Moreover, the parameters or predictors used in this model were cautiously chosen, to facilitate easier documentation in a clinical setup or in community level screening, when the patient walks into the clinic or primary health setup, thus limiting the need for extensive training of the ANN model network. This study demonstrated the good predictive validity with respect to LR risk score to identify individuals at the risk of AMD. With a risk score of >26, sensitivity was 87% and specificity was 51%. Whether our suggested cut-off of 30, which demonstrated the sensitivity of 79% and specificity of 69% is ideal for different population settings is not known, however, as mentioned earlier this data provides the opportunity to understand the ANN and LR model's accurate diagnostic predictive abilities and its comparisons. Since this data was population based cross-sectional in nature, one approach is to use the cut-off that maximizes predictive power mathematically and the prediction may only be consistent among patients with similar characteristics, resulting generalization of these results across some settings and may not suit others.

On the other hand, the computer-assisted tool based on ANN in our study predicts the outcome of AMD for patients with the presence or absence of various risk factors with greater performance compared to the multiple LR model. This is because the ANN model is equipped with the advantage of translating multivariable nonlinear relationships into continuous functions with various interactions between predictors and output variable, without the need of understanding exactly the underlying relationships between the variables.[11,12] In addition, ANN models are rich flexible nonlinear systems with a robust performance in dealing with cumbersome data and have the ability in generalizing from the input. We have trained and tested the ANN model on the entire data (n=3723) and on the subsample (n=213) drawn using RUS technique had different priority for the order of predictors. Additionally, we have used a 30-fold cross validation to generate the multiple replicate estimates of the performance of both LR and ANN models in this study. We have also used 10-fold cross validation, but there appears to be, in general, lower fold cross validation procedures that tend to provide lower estimates of the

relative importance of sensitivity analysis due to their relatively smaller sizes of training sets when compared to the higher fold (20-fold, 30-fold) partitions. Hence, we have performed a 30-fold higher cross validation procedure in this study. Therefore, due to higher accurate predictions, ANN predictive algorithms have potential to serve as promising metrics for diagnosis, prognosis, and therapeutic evaluation of irreversible eye disorders.

In our study, the predictive ability of the ANN model obtained from the original data decreased substantially. One of the reasons for this poor predictive ability was due to the lower error rates with the balancing of our original sample. Therefore, the results reported in this work may illustrate the importance of testing and validating the predictive models by means of resampling techniques, such as the ones used in this work. This approach would be more useful when small and poorly balanced samples are available related to the outcome variable studied.

Comparison of the LR and ANN models in predicting performance

As is shown in Figures 3a & 3b from results, the predictive performance of the LR and ANN models were the same in case of sub-population analysis. However, the predictive performance of ANN model was better when compared to the predictive ability of the LR model and the predictive ability was statistically significant between these two models (AUROC = 89% vs. 79%; P <0.0001) (Figures 4a & 4b). These models benefit from the learning of continuous functions, therefore, these models were applicable to the current data. The relationship between dependent variable and independent variable presents a linear result approximately after logistic transformation. Hence, the predictive performance of models depends on the relationship of variables. ANN shows better performance than LR when a nonlinear relationship occurs between independent and dependent variables regardless of logistic transformation. It is more appropriate to use a decision tree model[13] when there exists a step function relationship between independent and dependent variables, however, our data did not exhibit this kind of relationship and hence we did not use this model for prediction.

Risk score and sensitivity analysis

According to ANN, besides food intake below the recommendation, the second and third most important factors in predicting the occurrence of AMD were history of prior cataract surgery and history of heavy cigar smoking respectively. Both the models showed that history of heavy cigar smoking was a significant risk factor for increased risk of AMD. The risk score computed by the LR model in this study was 14 for an individual who has the habit of heavy cigar smoking (Table 4). Moreover, history of heavy cigar smoking is the first most modifiable risk factor in order of priority as identified by a 30-fold sensitivity analysis of the ANN model (Table 6b). Smoking is the most consistently identified risk factor for AMD in some of the white populations,[14-16] but not a risk factor in the few studies reported from Asian populations.[17,18] Our earlier report revealed that heavy cigar smoking was a significant modifiable risk factor for AMD.[2] Good eating habits and abstinence from smoking have a large impact on not only eye health but on general health too. Earlier studies have shown that nutritional supplements can help prevent certain age-related eye diseases. The Age-Related Eye Disease Study (AREDS) tested a food supplement combination of vitamin C, beta carotene (the precursor of vitamin A), vitamin E and zinc and the effect of this combination on macular degeneration.[19] On the other hand, the presence of combined risk factors of smoking and the most commonly seen gene of Compliment Factor H (CFH), is thought to increase the risk of AMD to almost a 34-fold risk.[20] In this ANN model, the calculation of various interactions were performed with the use of multi-layer feed-forward neural network architecture while calculating the relative importance of these smoking and nutrition variables, therefore, the network was trained better using an error back propagation training algorithm which showed that food intake below the recommendation is the second most important modifiable risk factor. Hence, the ANN model as in this case suggests that addressing the modifiable risk factors may be given the priority by policy makers in order to prevent or at least arrest the progression of AMD in this south Indian population.

However, there are some limitations of the ANN model that exhibits the week performance of the model. These were that the standardized coefficients and odds ratios corresponding to each variable cannot be easily calculated and presented as they are in regression models. ANN generates weights, which are difficult to interpret as they are affected by the program used to generate them. This lack of interpretability at the level of individual variables is one of the most criticized features in neural network models. Another limitation of our analysis is linked to the relative small sample size of AMD cases. From an epidemiological point of view because of this small sample size of AMD cases it is clear that the results presented in this work are only valid for this particular environment and cannot be generalized. Further replication of our study in cohort study of APEDS will help to correct upon these limitations and improve generalizability.

In conclusion, ANN and LR risk score models can be successfully used to model the risk of AMD and these models have given better predictive capability of AMD in the studied population. Risk score models may be capable of providing prognostic information with satisfactory explanatory power. Though this data is not readily useful for practical clinical usage for the accurate diagnosis of AMD this can, however, provide insights into the development of such validated risk scores based on the cohort study of APEDS that is currently being conducted

in AP. This cohort study can give an opportunity to model accurate ANN and LR risk score, which may help to establish early diagnosis of AMD in the population. The ANN model in this analysis provided the relative importance of input neurons both in terms of modifiable phenomena, which gives a basis to focus on modifying the modifiable factors and thus contributing to prevention strategies of AMD in this South Indian population.

ACKNOWLEDGEMENTS

This manuscript is part of the PhD work of the corresponding author titled "Modeling the risks of age-related eye diseases in a population in South India" in school of optometry and Vision science, University of New South Wales, Sydney. The authors thank the entire APEDS team, in particular Dr. Lalit Dandona who designed and conducted the detailed study; Dr. McCarty A for help with study design and guidance; Dr. Sreedevi Yadavalli for language editing; Mrs. Sabera Banu for library assistance; and all the volunteers for participating in this study.

Funding: The study was funded by Hyderabad eye research foundation, Hyderabad, India & Christoffel-Blinden mission, Germany
Conflict of interest: None declared
Ethical approval: The study was approved by the institutional review board of L. V. Prasad eye institute

REFERENCES

1. Finne P, Finne R, Stenman UH. Neural network analysis of clinicopathological factors in urological disease: a critical evaluation of available techniques. Br J U Int. 2001;88:825-31.
2. Krishnaiah S, Das TP, Nirmalan PK, Nutheti R, Shamanna BR, Rao GN, et al. Risk Factors for age-related macular degeneration: findings from the Andhra Pradesh eye disease study. Invest Ophthalmol Vis Sci. 2005;46:4442-9.
3. Dandona L, Dandona R, Srinivas M, Giridhar P, Vilas K, Prasad MN, et al. Blindness in the Indian state of Andhra Pradesh. Invest Ophthalmol Vis Sci. 2001;42:908-16.
4. Dandona R, Dandona L, Naduvilath TJ, Nanda A, McCarty CA. Design of a population study of visually impairment in India: the Andhra Pradesh eye disease study. Indian J Ophthalmol. 1997;45:251-7.
5. Dandona L, Dandona R, Naduvilath TJ, McCarty CA, Nanda A, Srinivas M, et al. Is eye-care-policy focus almost exclusively on cataract adequate to deal with blindness in India? Lancet. 1998;21:1312-6.
6. Hosmer DW, Lemeshow S. Logistic regression. In: Hosmer DW, Lemeshow S, eds. Statistics in Medicine. 2nd ed. New York: Wiley; 1989;10:1162-3.
7. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med. 1986;5(5):421-33.
8. Aday LA. Bivariable analysis. In: Aday LA, eds. Designing & Conducting Health Surveys. 2nd ed. San Francisco: Jossey Bass Publishers; 1996.
9. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig M, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care. 2007;30:510-5.
10. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back propagation error. Nature. 1986;323;533-6.
11. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/valance dilemma. Neural Comput. 1992;4:1-58.
12. Garson GD. Interpreting neural-network connection weights. Artif Intell Expert. 1991;6:47-51.
13. Chang-ping LI, Xin-yue ZHI, Jun MA, Cui Z, Zhu ZL, Zhang C, et al. Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. Chin Med J. 2012;125(5):851-7.
14. Klein R, Klein BE, Linton KL, DeMets DL. The Beaver Dam Eye Study: the relation of age-related maculopathy to smoking. Am J Epidemiol. 1993;137:190-200.
15. Smith W, Mitchell P, Leeder SR. Smoking and age-related maculopathy. The Blue Mountains Eye Study. Arch Ophthalmol. 1996;114:1518-23.
16. Vingerling JR, Hofman A, Grobbee DE, de Jong PT. Age-related macular degeneration and smoking. The Rotterdam Study. Arch Ophthalmol. 1996;114:1193-6.
17. Miyazaki M, Nakamura H, Kubo M, Kiyohara Y, Oshima Y, Ishibashi T, et al. Risk factors for age related maculopathy in a Japanese population: the Hisayama Study. Br J Ophthalmol. 2003;87:469-72.
18. Xu L, Li Y, Zheng Y, Jonas JB. Associated factors for age related maculopathy in the adult population in China: the Beijing Eye Study. Br J Ophthalmol. 2006;90:1080-90.
19. Age-Related Eye Disease Study Research Group. The relationship of foodary carotenoid and vitamins A, E, and C Intake with AMD in a case control study: AREDS Report No 22. Arch Ophthalmol. 2007;125(9):1225-32.
20. Despriet DD, Klaver CC, Witteman JC, Bergen AA, Kardys I, de Maat MP, et al. Complement factor H polymorphism, complement activators, and risk of age-related macular degeneration. JAMA. 2006;296:301-9.

DOI: 10.5455/2394-6040.ijcmph20150514
Cite this article as: Krishnaiah S, Surampudi BR, Keeffe J. Modeling the risk of age-related macular degeneration and its predictive comparisons in a population in South India. Int J Community Med Public Health 2015;2:137-48.