Application of Binary Logistic Regression to the Prediction of Tuberculosis Disease in Rural Area of South India

Senthilvel Vasudevan1*, Jayanthi Surebabu2 and Sumathi Senthilvel3

1Lecturer in Biostatistics, Department of Pharmacy Practice, College of Pharmacy, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

2Formerly Tutor in Medical Entomology, Department of Community Medicine, Pondicherry Institute of Medical Sciences, Kalapet, Pondicherry – 605 014, South India

3Formerly Lecturer in Nursing (OBG), Kasturba Gandhi College of Nursing, Mahatma Gandhi College of Medical and Research Institute, Pillayarkuppam, Pondicherry – 607 402, South India

Abstract

Context

Tuberculosis (TB) is caused by Mycobacterium Tuberculosis most commonly affects the lungs in human beings. TB is common in India.

Aims

1. To find the risk factors of tuberculosis disease.
2. To estimate the probability of developing tuberculosis in an individual.

Settings and Design

A community based cross sectional study was done in the Chunnampet area of Tamil Nadu, South India.

Methods and Material

Adult populations (age ≥ 18, n = 1958) were analyzed. Variables considered included age, educational status, employment status and family income. A pre-designed and structured questionnaire was used for data collection. Study participants gave informed consent orally.

Statistical Analysis Used

Descriptive Statistics (Mean, SD and Proportion), Univariate Analysis and Binary Logistic Regression Analysis were performed.

Results

Participants were 41.90 ± 16.66 years old. Employment status, family income, tobacco smoking, overcrowding, diet habit and family history of tuberculosis were statistically associated with TB. Finally, the probability of getting or chance to develop TB in an individual was represented by a risk ratio of 0.71.

Conclusions

The present study revealed that the probability or chance to get tuberculosis in an individual was very high significant in the participant’s employment status, family income, tobacco smoking, family history, overcrowding, diet habit and weight except of the variables like age, educational status and cooking with wood.

Keywords: Tuberculosis; Prediction; Probability; Risk Factors; Overcrowding; Rural Area

Key Messages

1. The community people will have got some idea about the risk factors
2. They have to find themselves about the risk factors.
3. Finally, they will control themselves the prominent risk factors and controlled them in time. 4. By the recent advancement in the facility and treatment they will have to control and free from the communicable diseases.

Introduction

Tuberculosis (TB) is caused by Mycobacterium and most commonly affects the lungs. TB is also called as disease of poverty and potentiated by immunodeficiency. Thirty-seven million people have been saved by proper medication available since the year 2001 [1]. In 2013, there were an estimated of 13.7 million active chronic cases globally [2]. In India, the prevalence of TB was estimated as 230 per one lakh population at the year of 2012 [3]. Pulmonary tuberculosis is mostly commonly form but other symptoms are fever, chills, night sweats, fatigue, nail clubbing, loss of appetite and...
weight. Most of the infections lack of any symptoms and are called latent tuberculosis. Infections can be reduced by isolating infected TB patients in sanitoriums and pasteurization of milk, to improve the housing environment and decrease crowding to improve hygiene environment and decrease the crowding, to improve hygiene and sanitation with clean water and better nutritious [4,5,6]. In this study, the objectives are finding the risk factors of tuberculosis disease among the study population and to estimate probability of developing tuberculosis in an individual.

**Subjects and Methods**

**Study Design, Population, Sample Size and Study Variables**

A community based cross sectional study was done in the Chunnampet area of Tamil Nadu, South India in the year of 2012 with the sample size of 2005. In this particular study, adult population (age ≥ 18) only was taken into account for data analysis and finally the sample size was found n = 1958. The study variables age, educational status, employment status, family income, tobacco smoking, family history of tuberculosis, overcrowding, cooking with wood, diet habit and weight of the respondents were considered in this present study.

**Data collection**

The present study was conducted with a pre-designed and structured questionnaire was used for data collection with the oral consent of the study participants over six villages and the study was conducted by using door to door interview method. In this study, study participants were included those who are willing to participate in the study and residing in the study area with minimum one year and above. It was identified by seeing their respective ration card officially issued by the Government of Tamil Nadu.

**Statistical Analysis used**

The quantitative variables are expressed as follows: (a). **Descriptive statistics**: mean, S.D, proportions. (b). **Univariate Analysis**: The associations between the symptom of TB and the study variables like age, educational status, employment status, family income (in monthly), tobacco smoking, family history of tuberculosis, overcrowding, cooking with wood, diet habit and weight are appropriately evaluated through spearman's rank correlation coefficient or through non-parametric statistics Mann-Whitney-Wilcoxon on test for independent samples. (c). **Binary Logistic Regression Analysis (BLR Analysis)** [7]. Prediction of probability was found by finding the Odds ratio and finding the risk factors of tuberculosis disease among the adult study population by using Binary Logistic Regression Analysis (BLR Analysis). Symptoms of Tuberculosis was taken as dependent variable and other factors like age, educational status, employment status, family income (monthly), tobacco smoking, family history of tuberculosis, overcrowding, cooking with wood, diet habit and weight were taken as independent variables. The significance was taken in this study as p – value less than 0.05.

**Consent from the Participants**

The oral consent was only obtained from the study participants.

**Ethical Issue**

In this study, there were no ethical issues.

**Variables Included in the Study**

The variables were taken and included in the BLR Analysis. They were as follows:

**Symptom of Tuberculosis**: If any person have the symptom of TB = 1  
= 0, otherwise

**Age** (X₁) - in years

**Educational Status** (X₂): Literate = 1

Illiterate = 0 otherwise.

**Employment Status** (X₃): If employed = 1

= 0 otherwise.

**Family Income** (X₄): If high income = 1 (> Rs. 5000)

= 0 otherwise

**Tobacco Smoking** (X₅): If smoking = 1

= 0 otherwise

**Family History of TB** (X₆): If an individual having TB = 1

= 0 otherwise

**Overcrowding** (X₇): If overcrowding = 1

= 0 otherwise

**Cooking** (X₈): If Cooking with wood = 1

= 0 otherwise

**Diet Habit** (X₉): If an individual took nutritious food = 1

= 0 otherwise

**Weight (X₁₀)** – in Kgs.

**Model Used in this Study**

By using the BLR Analysis, find out the significant variables and fitting a Binary Logistic Regression Analysis equation and as follows:

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k 
\]

And find out the \( Y \) and \( e^Y \) values. Then, finally substitute the \( e^Y \) – value in the following equation and find the \( P \) – value.

\[
P = \frac{1}{1 + e^{-Y}}
\]

Finally, found out the value of \( P \). This gives the probability of
getting or chance to develop TB in an individual.

Results

In the present study, the average age of the adult study participants’ was found as 41.90 ± 16.66 (Range: 18 – 90) years and the average weight was found as 55.63 ± 11.17 (Range: 15 – 98) kgs. In Univariate Analysis, the symptom of TB was positively associated with the patients’ age with Spearman’s rho = 0.071 and p – value < 0.01, negatively associated with weight with Spearman’s rho = -0.098 and p – value < 0.001. The other variables like gender (p – value < 0.05, MWW), educational status (p – value < 0.01, MWW), employment status (p – value < 0.001), family income (monthly) (p – value < 0.001, MWW), tobacco smoking (p – value < 0.001, MWW), family history of tuberculosis (p – value < 0.01, MWW), overcrowding (p – value < 0.001, MWW) and diet habit (p – value < 0.001, MWW). The study variables are taken and fixed in the BLR equation and found the required equation, significance, odd ratios and 95% Confidence Interval are shown in table 1.

In BLR Analysis, the study variables: educational status (OR: 0.812, C.I: 95% 0.670 – 0.984), employment status (OR: 0.593, C.I: 95% 0.478 – 0.737), family income (OR: 1.436, C.I: 95% 1.147 – 1.797), tobacco smoking (OR: 1.681, C.I: 95% 1.292 – 2.186), family history of tuberculosis (OR: 0.613, C.I: 95% 0.417 – 0.899), overcrowding (OR: 1.413, C.I: 95% 1.164 – 1.715), diet habit (OR: 0.444, C.I: 95% 0.355 – 0.557) and weight was found as 5.493, df = 8, and p – value = 0.704 which was greater than 0.05 and it wasn’t significant. This was showed that the model was good fit with these listed variables in the present study.

Example

According to the data analysis, the Binary Logistic Regression equation (1) has been modified and it is given by as follows:

\[ Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_10 \]

Here, \( \beta_0 = 0.917 \); \( \beta_2 = -0.209 \); \( \beta_3 = -0.522 \); \( \beta_4 = 0.355 \); \( \beta_5 = -0.811 \); \( \beta_6 = -0.009 \);

and \( X_2 = 0 \); \( X_3 = 0 \); \( X_4 = 0 \); \( X_5 = 1 \); \( X_6 = 1 \); \( X_7 = 1 \); \( X_8 = 0 \); \( X_9 = 0 \); \( X_{10} = 45 \) Kgs.

From equation (1), we found as follows:

\[ Y = 0.887 \]

Therefore, \( e^Y = e^{0.887} \)

\[ = 2.428 \]

Then, from equation (2), we got,

\[ p = \frac{1}{1 + e^Y} \]

\[ = \frac{1}{1 + e^{0.887}} \]

Table 1: Distribution of variables in the Binary Logistic Regression equation, significance, odds ratios and 95% Confidence Interval (Lower and Upper Limits)

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>S.E</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I for EXP(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>Age (X_1)</td>
<td>-0.002</td>
<td>0.003</td>
<td>0.332</td>
<td>1</td>
<td>0.565*</td>
<td>0.998</td>
<td>0.992</td>
</tr>
<tr>
<td>Educational Status (X_2)</td>
<td>-0.209</td>
<td>0.098</td>
<td>4.526</td>
<td>1</td>
<td>0.033</td>
<td>0.812</td>
<td>0.670</td>
</tr>
<tr>
<td>Employment Status (X_3)</td>
<td>-0.522</td>
<td>0.111</td>
<td>22.188</td>
<td>1</td>
<td>0.000</td>
<td>0.593</td>
<td>0.478</td>
</tr>
<tr>
<td>Family Income (X_4)</td>
<td>0.362</td>
<td>0.114</td>
<td>9.985</td>
<td>1</td>
<td>0.002</td>
<td>1.436</td>
<td>1.147</td>
</tr>
<tr>
<td>Tobacco Smoking (X_5)</td>
<td>0.519</td>
<td>0.134</td>
<td>14.954</td>
<td>1</td>
<td>0.000</td>
<td>1.681</td>
<td>1.292</td>
</tr>
<tr>
<td>Family History of TB (X_6)</td>
<td>-0.490</td>
<td>0.196</td>
<td>6.258</td>
<td>1</td>
<td>0.012</td>
<td>0.613</td>
<td>0.417</td>
</tr>
<tr>
<td>Overcrowding (X_7)</td>
<td>0.346</td>
<td>0.099</td>
<td>12.192</td>
<td>1</td>
<td>0.000</td>
<td>1.413</td>
<td>1.164</td>
</tr>
<tr>
<td>Cooking with wood (X_8)</td>
<td>0.046</td>
<td>0.100</td>
<td>0.214</td>
<td>1</td>
<td>0.644*</td>
<td>1.048</td>
<td>0.860</td>
</tr>
<tr>
<td>Diet Habit (X_9)</td>
<td>-0.811</td>
<td>0.115</td>
<td>49.562</td>
<td>1</td>
<td>0.000</td>
<td>0.444</td>
<td>0.355</td>
</tr>
<tr>
<td>Weight (X_{10})</td>
<td>-0.009</td>
<td>0.004</td>
<td>3.901</td>
<td>1</td>
<td>0.048</td>
<td>0.991</td>
<td>0.983</td>
</tr>
<tr>
<td>Constant</td>
<td>0.917</td>
<td>0.335</td>
<td>7.475</td>
<td>1</td>
<td>0.006</td>
<td>2.502</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p &lt; 0.05</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>p &gt; 0.05</td>
<td>Not Significant</td>
</tr>
<tr>
<td>p &lt; 0.01</td>
<td>Highly Significant</td>
</tr>
<tr>
<td>p &lt; 0.001</td>
<td>Very Highly Significant</td>
</tr>
</tbody>
</table>

S. E - Standard Error  df - degrees of freedom  Sig. - Significant  Exp (B) - Expected Beta  C.I - Confidence Interval
been explained that the crowding was a high significant factor for developing TB in human population has been reported by Stein (1952) [15]. In the present study, we have got a result as weight increase in the population, and then the chance of occurrence or the development of TB was low in an individual. The similar type of results have been obtained in a study was conducted in Hong Kong by Leung CC et al (2007) [16] were mentioned in that the obesity was associated with a lower risk of active pulmonary TB in the older population.

Conclusion

The present study revealed that the educational status, employment status, family monthly income, tobacco smoking, overcrowding and diet habit were the independent risk factors to the development of tuberculosis in an individual or a person or a human being. So, by avoiding regular smoking, by eating nutritious food recommended by a dietician and careful with other risk factors, then one can prevent and free from the development of TB or to prolong the diseases for a longer period. To eliminate the development of TB disease in the initial stage from the community by conducting IEC program and Group Discussion activities on smoking and other tuberculosis related risk factors among rural community through the Governmental and Non-Governmental organizations.

Acknowledgement

No funding of sources to our study. Authors are thankful to Dr. (Brig.) Zile Singh, Prof. and HOD, Dept. of Community Medicine, Pondicherry Institute of Medical Sciences, Kalapet, Puducherry – 605 014. South India for his excellent help and encouragement to our study. Authors are thankful to the faculties, students and to the interns those who were participated in this study. Authors are very thankful to the community people those who were participated in this study and for their great cooperation.

References

1. WHO: What is Tb? How is it treated?
3. WHO: Countries: India Available from the web site.

Discussion

In the present study, the educational status was significant with the development of TB. The factors like, low educational status, low income of family and low employment status were also independent significant risk factors in the development of TB. The similar types of things were mentioned in the Centers for Disease Control and Prevention (CDC) [8]. Age and cooking with wood were not risk factors for the development of TB. In a study, persons with age 60 and above were familial contact with pulmonary TB which was given by [9]. Even though, gender was a significant factor in Univariate Analysis but, it wasn’t included in the Binary Logistic Regression Analysis. Because of there was no evidence for that association from the previous studies. Similar results were mentioned by [10] that the cooking with biomass fuels was not significant with TB development. In our study, the tobacco smoking and diet are statistically highly significant with p < 0.001. The same type of results have been obtained in a review study on risk factors for TB was done by Padmanesan Narasimhan et al (2013) [11]. In a study by Davies P.D et al (2006) [12] was mentioned that the smoking of cigarettes of more than 20 per day has been increased the TB case rates and the development of progressive TB disease by the latent Mycobacterium tuberculosis infection. In our study, the family history of TB is a significant factor for developing TB in an individual. The same result was obtained by Richard Bellamy (1998) [13] has been reported that in his study there was some strong evidence that the host genes influence susceptibility to tuberculosis in human beings.

In this study, we have got the factor overcrowding was an independent and very highly significant risk factor for developing TB in a person or in an individual. In another one study related to overcrowding, the overcrowding was an independent risk factor for developing TB in human population has been reported by Bernard Larouze et. al. (2008) [14]. In a study by Lilli Stein has been explained that the crowding was a high significant factor of TB disease and with other social variables also. The strongest association being with crowding i.e., average no. of persons per house by Stein (1952) [15]. In the present study, we have got a result as weight increase in the population, and then the chance of occurrence or the development of TB was low in an individual. The similar type of results have been obtained in a study was conducted in Hong Kong by Leung CC et al (2007) [16] were mentioned in that the obesity was associated with a lower risk of active pulmonary TB in the older population.

Discussion

The probability of getting or chance to develop TB in an individual = 71% (approximately)

\[
P = \frac{2.428}{2.428 (1 - P)}
\]

\[
P = 2.428 - 2.428P
\]

\[
P (1 + 2.428) = 2.428
\]

\[
P = \frac{2.428}{3.428}
\]

\[
P = 0.7082 \text{ (or) } 70.82\% \sim 71\%
\]

References

1. WHO: What is Tb? How is it treated?
3. WHO: Countries: India Available from the web site.
8. CDC: Centers for Disease Control and Prevention. Social determinants of Health: Why is addressing the role of social determinants of health important.


