Modeling tuberculosis patients using loglinear models

Noorliyana Nasseruddin¹, Zamalia Mahmud¹ and Khairil Anuar Md Isa²

¹Faculty of Computer and Mathematical Sciences Universiti Teknologi Mara, 40450 Shah Alam, Selangor, Malaysia
²Faculty of Health Sciences Universiti Teknologi Mara Selangor, Puncak Alam Campus, 42300 Bandar Puncak Alam, Selangor, Malaysia

E-mail: noorliyananasseruddin@yahoo.com

Abstract. Tuberculosis (TB) patients usually get infected by tuberculosis bacteria called Mycobacterium tuberculosis which can spread from one person to another through air. TB cases normally infect the lungs (Pulmonary TB) and in some other cases it infects other organs such as brain, heart, spine, lymph nodes, adrenal gland and intestines (Extrapulmonary TB). However, patients who are infected with tuberculosis bacteria may have complications due to other factors that can delay the process of healing. Using loglinear modelling, this study will examine whether types of TB is strongly associated with patient’s age, gender, whether or not they smoke, have HIV and into drug use. Records of 477 TB patients obtained from a previous related study are used in the modelling process where five loglinear models are compared to find the most parsimonious model that will determine the strength of the association between the variables. Likelihood ratio $G^2$, p-values, expected values and standardized residuals are used in the analysis. Based on the comparisons of the loglinear models, homogeneous model is found to be the most parsimonious compared to independence and joint models. The analysis based on the homogenous loglinear model show that the estimated odds of patients who smoke is 2 times more likely to have a Pulmonary TB compared to the non-smokers. Also patients who are drug users are 8 times more likely to have a Pulmonary TB compared to patients who are non-drug users. Hence, a recommendation for the healthcare providers is to give more attention to patients in these categories.

1. Introduction
Tuberculosis (TB) is an infectious disease that can spread from one person to another through air. It is caused by the bacteria called Mycobacterium tuberculosis. A person with TB disease can spread the disease to other people when they cough, sneeze or spit the TB germs into the air.
TB is one of the top 10 causes of death worldwide and it is a leading killer for people with HIV positive. About one quarter of the world population has been infected by TB bacteria but has not fallen ill to the disease.

In 2017, there was an estimated of 87% of new TB cases worldwide. India has the highest number of TB cases with 64% of the total, followed by China, Indonesia, Philippines, Pakistan, Nigeria and South Africa [1]. The deputy director-general of Public Health, Malaysia stated that in 2017, there are 26,168 TB cases reported an increase as 8% from 24,220 cases in year 2015. There were 6.5 death rates of over 100,000 of the population in 2017. The second highest number of deaths in government hospitals is the disease of respiratory system which includes TB infections [2]. Tuberculosis (TB) is divided into two categories which are Pulmonary and Extrapulmonary [3]. The Mycobacterium tuberculosis bacteria will cause tuberculosis (TB) which can destroy the human body tissue. Pulmonary TB happens when a Mycobacterium tuberculosis bacterium attacks the lungs and possibly to other organs. This bacterium can be transmitted through airborne. It may be curable if diagnosed early and treated with antibiotic. Extrapulmonary TB involves organs other than the lungs such as skin, joints, bones, pleura, lymph nodes, abdomen, genitourinary tract, and meninges. Even though Extrapulmonary TB cases are not major compared to Pulmonary TB, it also contributes to the burden of the disease and does not receive any specific attention internationally. The proportions of Extrapulmonary TB cases have increased [4].

Tuberculosis TB is a disease that can infect other people if there is no control over its spread. Controlling the spread of TB is difficult because it is difficult to diagnose and treat. Based on the analysis of previous study on the Prevalence of Comorbidities and The Risk Factors for Multiple Drug Resistant TB (MDRTB) Patients at the Institute of Respiratory Medicine [5], they only measure the significant risk factors of TB using Multiple Logistic Regression. They found that BCG status and patients with history of TB infection are the significance risk factors of MDRTB. However, this study will look at different perspective of the study by examining the strength of relationship between the risk factors (HIV, Drug Use and Smoking), characteristics (Age) of TB patients and TB Categories (Pulmonary and Extrapulmonary) by using Loglinear modeling.

2. Research design and methodology
This study had used 477 TB patient’s data obtained from a team of researchers at the Faculty of Health Sciences UiTM Puncak Alam Selangor who were involved in a study on Prevalence of comorbidities & risk factors for Multidrug Resistance among Tuberculosis patients in the Institute of Respiratory Medicine, Kuala Lumpur. The data consist of patient’s demographic profile, the patient health history and other related information about TB. Loglinear modeling was applied on the data in order to measure the strength of association between the categorical variables. Loglinear modeling has the ability to simultaneously test the association between multiple outcome and multiple explanatory variables.

2.1. Building loglinear models
Loglinear analysis deals with the association between categorical variables data by examining the entire main and interaction effects. In loglinear analysis, there is no distinction between the response and explanatory variables as all the variables are treated the same. The primary purpose is to find the most parsimonious model by comparing the saturated model (full model) with the reduced model such as (homogeneous and independence models). Loglinear models are associated with ANOVA (Analysis of Variance) and regression analysis but they are more flexible and interpretable. Loglinear models are mainly used with at least two response variables in the contingency tables [8]. Loglinear analysis would give interpretations and pattern of association interns of odds and odds ratio where it can be expressed as

\[
\text{Odds} = \frac{P}{1-P} = \frac{\text{Probability (Event Occurs)}}{\text{Probability (Event Does Not Occurs)}}
\]  

(1)
The larger the odds value, the larger the probability [6]. The other rescaling is the log odds. A log odd is a natural transformation of the odds.

There are two types of loglinear analysis which are General Loglinear Analysis (GENLOG) and Hierarchical loglinear analysis (HILOG). General loglinear modeling is used if we wish to search manually the parsimonious model among a finite set of models. Hierarchical loglinear analysis begins with a saturated model. It uses backward elimination stepwise procedure which can be used to screen alternative models automatically. General loglinear analysis (GENLOG) is often used to refine the best hierarchical model to be more parsimonious by dropping terms. This study had used the backward elimination that begins with the saturated model. For three-way contingency tables, the models would present various association patterns as described in the following sections.

2.1.1. Saturated model - The saturated model contains the maximum number of parameters. Any set of expected frequency can be described perfectly and saturated model fits to the data. This model shows no constraints of the data and reproduces the observed cell frequencies. Therefore saturated model is used as a basis for loglinear analysis. The model structure with variables X,Y and controlled by Z. Age of patients is the controlled variable in this study. Previous study used age and demographic background as a controlled variable [9].are shown below;

\[
\text{log}(E_{ijk}) = \lambda + \lambda_X^i + \lambda_Y^j + \lambda_Z^k + \lambda_{XZ}^i + \lambda_{YZ}^j + \lambda_{XY}^i + \lambda_{XYZ}^i
\]

where,
\[
\begin{align*}
\log(\mu_{ijk}) &= \text{log of the expected cell frequency of the cases for cell } i,j,k \\
\lambda &= \text{represents overall effect or a grand mean of the logarithms of the expected frequency.} \\
\lambda_X^i, \lambda_Y^j, \lambda_Z^k &= \text{represent the main effects of variables X, Y, and Z.} \\
\lambda_{XZ}^i, \lambda_{YZ}^j, \lambda_{XY}^i &= \text{represent the interaction or association between two variables X and Y, Y and Z and X and Y.} \\
\lambda_{XYZ}^i &= \text{represents the interaction or association between three variables X,Y and Z.}
\end{align*}
\]

In this study, the saturated model with three-way effect of smoking (S), TB categories (Pulmonary and Extrapulmonary) (C) and age (A) shows the following model structure;

\[
\text{log}(\mu_{ijk}) = \lambda + \lambda_S^i + \lambda_C^j + \lambda_A^k + \lambda_{SA}^i + \lambda_{CA}^j + \lambda_{SC}^i + \lambda_{SCA}^i
\]

2.1.2. Homogeneous model - Homogeneous model exists when there is a two-way interaction. The homogeneous association is when the conditional odds ratio between any two variables is same at each level of the third variable. The homogeneous model structure of of smoking (S), TB categories (Pulmonary and Extrapulmonary) (C) and age (A) is as follows;

\[
\text{log}(\mu_{ijk}) = \lambda + \lambda_S^i + \lambda_C^j + \lambda_A^k + \lambda_{SA}^i + \lambda_{CA}^j + \lambda_{SC}^i
\]

2.1.3. Independence model - Independence model has three different models that can capture the relationship between variables in three-way contingency table. They are mutual independence, joint independence, and conditional independence.
• Mutual independence
   This is the simplest model where all the variables are independent of one another. The model function \((S,C,A)\) is,
   \[
   \log(\mu_{ijk}) = \lambda + \frac{S}{i} + \frac{C}{j} + \frac{A}{k} \tag{5}
   \]

• Joint independence
   This model indicates that two variables are jointly independent of the third variable. The model function for \((SC, A)\) is,
   \[
   \log(\mu_{ijk}) = \lambda + \frac{S}{i} + \frac{C}{j} + \frac{A}{k} + \frac{SC}{ij} \tag{6}
   \]

• Conditional independence
   This model consists of three possible models with three random variables \((SC,SA), (SC,CA)\) and \((SA,CA)\). Considering the model function for \((SC,SA)\),
   \[
   \log(\mu_{ijk}) = \lambda + \frac{S}{i} + \frac{C}{j} + \frac{A}{k} + \frac{SC}{ij} + \frac{SA}{ik} + \frac{SC}{ij} \tag{7}
   \]

2.2. Inference of loglinear
   This section highlights the model selection from the models that fit well. This study focuses on the simplest possible model that explains the most variability in the data which is the parsimonious model. The process started with the saturated model and deletes the higher order until the model fits the data. This section describes the goodness of fit test and loglinear cell residual.

2.2.1. Goodness of fit
   The null hypothesis for likelihood-ratio goodness of fit test stated that the loglinear model is fit. If \(p\)-values is greater than 0.05, therefore the loglinear model is fit. The likelihood-ratio statistics for three-way case is,
   \[
   G^2 = 2 \sum n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) \quad i = 1, 2, ..., ith \ row \quad j = 1, 2, ..., jth \ column
   \tag{8}
   \]
   The \(G^2\) is the deviance for the model. Larger likelihood \(G^2\) values indicate that the model does not fit the data well while a good fitting model is when the \(G^2\) value is smaller and \(p\)-value is large [7].

2.2.2. Loglinear cell residual
   A good model should have small residuals [8]. The residual is used to find out how well are the cells fitted by the model. There is lack of fit if there are cells with small sample and adjusted residuals are greater than 2. The standardize residual for each cell is calculated by dividing the difference between observed and expected frequencies by the square root of the expected frequencies \((F_{obs} - F_{exp} / \sqrt{F_{exp}})\). The models are least appropriate if the cell shows the largest residuals. The residual should consist both positive and negative values of approximately the same magnitude that are distributed evenly across the cell of the table. Standardized residual is a ratio of the difference between the observed count and the expected count and the standard deviation of the expected count or standard error. The formula as follows;
   \[
   \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_i)(1-p_j)}} \tag{9}
   \]
   It is also used to measure the strength of the differences between the observed and the expected values which measures the significant of the cell to the chi-square value. The residual value less than -2 shows that the cell’s observed frequency is less than the expected frequency. For residual value greater than 2, the observed frequency is greater than the expected frequency.
2.3 Test of effect sizes on the most Parsimonious model

Test of effect sizes can precede after the most parsimonious model was selected. The effect sizes can be measured by odds and odds ratio. This section will discuss about the parameter estimates and odds ratio.

2.3.1. Parameter estimates - Poisson distribution with unknown parameter $\lambda$ is the observation of $X_1, X_2, X_3, \ldots X_n$ where the likelihood function is;

$$L(\lambda; x) = \prod_{i=1}^{n} f(x_i; \lambda) = \prod_{i=1}^{n} \frac{x_i^{\lambda} e^{-\lambda}}{x_i!}$$  \hspace{1cm} (10)

Poisson likelihood function is obtained based on differentiation of the log of this function with respect to $\lambda$ [10].

$$l(\lambda; x) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda$$  \hspace{1cm} (11)

The parameter estimate is $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$ where the maximum likelihood estimation for $\lambda$ is the sample mean. The parameter estimates are the effect sizes which can be measured by odds and odds ratio.

2.3.2. Odds and odds ratio - The effect parameters estimates are related to odds and odds ratio. Another measure of association for two-way contingency tables is by using the odds ratio. Odds are the ratios of the probability of an event occurring ($\pi$) to the probability of the event not occurring ($1-\pi$) or probability of success ($\pi$) to the probability of failure as follow;

$$\text{Odds} = \frac{\pi}{1-\pi}$$  \hspace{1cm} (12)

The odds values are nonnegative, if it greater than 1.0 indicates that the success is more likely than failure. In 2 x 2 tables, the odds success for row 1 is $odds_1 = \frac{\pi_1}{1-\pi_1}$ and odds success for row 2 is $odds_2 = \frac{\pi_2}{1-\pi_2}$. The odds ratio is the odds from the two rows. The odds ratio between two odds;

$$\text{Odds Ratio, (OR)} \hat{\theta} = \frac{odds_1}{odds_2} = \frac{\pi_1}{1-\pi_1} \times \frac{1-\pi_2}{\pi_2}$$  \hspace{1cm} (13)

If $X$ and $Y$ are independent then, $\pi_1$ is equal to $\pi_2$ where the odds ratio is equal to 1. For $\hat{\theta}$ values greater than 1 represent there is a strong association. The farther the values from 1, the stronger is the association. The odds ratio is used to measure the association between exposure and the outcome. OR equal to 1 indicates that the exposure does not affect the odds of outcome, OR > 1 indicate exposure is associated with higher odds outcome and OR < 1 indicates the exposure is associated with lower odds outcome. If the odds ratio value is 0.5, the odds of success in row 1 is 0.5 times higher than the odds of success in row 2 or can be stated as $1/0.5 = 2.0$ times as high in row 2 as in row 2 [11].

3. Analysis and results

3.1 Comparison of loglinear models between Age (A), Smoking (S) and TB Categories (C)

$$\log(\mu_{ijk}) = \lambda + A_i + S_j + C_k + AS_{ik} + AC_{jk} + SC_{ij} + \lambda^{ASC}_{ijk}$$  \hspace{1cm} (14)

Three methods are used to compare the loglinear models which are fitted values, likelihood ratio $G^2$ statistics and cell residuals. Fitted values in the loglinear models are compared against the saturated model. The closer the fitted values to the saturated model indicate that the model fits the data well. Table 1 shows the fitted values for five loglinear models - mutual independence (A,S,C), joint independence (SC,A), conditional independence (SC,SA), homogeneous (AS,AC,SC) and saturated (SAC).
Table 1. Fitted values for loglinear models

| Age (A) | Smoking (S) | TB Categories (C) | Loglinear models |
|---------|-------------|-------------------|------------------|
|         |             | Pulmonary         | (S,C,A)          |
| <35     | Yes         | 86.43             | 93.86            |
|         |             | 22.54             | 15.10            |
| No      |             | 114.24            | 106.81           |
|         |             | 29.79             | 37.22            |
| ≥ 35    | Yes         | 73.79             | 80.14            |
|         |             | 19.24             | 12.89            |
| No      |             | 97.54             | 91.19            |
|         |             | 25.43             | 31.78            |

Table 1 shows the comparison of fitted values for several loglinear models against the saturated model (SAC). The results show that the homogeneous model (AS,AC,SC) has the closest fitted values to (SAC) an indication of the best fit model compared to other loglinear models. The others model found poorly fit due to the large discrepancies between the models fitted values and the saturated model fitted value.

3.1.1. Likelihood ratio goodness of fit test - Smaller likelihood ratio $G^2$ statistics value indicates a better model fit. Table 2 shows the results of the likelihood ratio goodness of fit $G^2$ statistics for all models.

Table 2. Likelihood ratio goodness of fit statistics

| Model      | $G^2$ | Degree of Freedom | P-value |
|------------|-------|-------------------|---------|
| (S,C,A)    | 26.055| 4                 | 0.000   |
| (SC,A)     | 15.657| 3                 | 0.001   |
| (SC,SA)    | 5.562 | 2                 | 0.062   |
| (AS,AC,SC) | 1.416 | 1                 | 0.234   |
| (SAC)      | 0     | 0                 | 0       |

It shows the most fitting model is homogeneous model (AS,AC,SC), since the $G^2$ statistics is the smallest ($G^2 = 1.416$) and largest p-value 0.234. The homogeneous model P-value equal to 0.234 are greater than 0.05 significant value, thus the model fit the data.

3.1.2. Loglinear cell residual - The larger residual indicates the data is poorly fit. The smaller positives and negatives residual indicate that there is a strong association between a particular relationship, while larger positives and negatives residual indicates a week association. Table 3 shows the loglinear cell residual for the homogeneous model (AS, AC, SC).
Table 3. Loglinear cell residuals for homogeneous model

| AGE (A) | SMOKING (S) | TB CATEGORIES (C) | Observed count | Model (AS,AC,SC) |
|---------|-------------|-------------------|----------------|------------------|
| <35     | Yes         | Pulmonary          | 74             | Fitted Count     |
|         |             | Extrapulmonary     | 18             | 76.34            |
|         | No          | Pulmonary          | 116            | 15.66            |
|         |             | Extrapulmonary     | 45             | 113.66           |
| ≥35     | Yes         | Pulmonary          | 100            | 47.34            |
|         |             | Extrapulmonary     | 24             | 97.66            |
|         | No          | Pulmonary          | 10             | 0.592            |
|         |             | Extrapulmonary     | 82             | 0.220            |

Table 3 shows that the standardized residual values are small. There is no extreme difference between the fitted and observed values. Both the positives and negatives values of residuals are small which indicates a good model fit. Finally, the homogeneous model of Age, Smoking and TB Categories (Pulmonary and Extrapulmonary) is found to be the most parsimonious model compared to other loglinear models.

3.2 Comparison of loglinear models between Age (A), Drug use (D) and TB Categories (C)

\[
\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^D + \lambda_k^C + \lambda_{ij}^{AD} + \lambda_{ik}^{AC} + \lambda_{jk}^{DC} + \lambda_{ijk}^{AD\cdot DC}
\]  

(15)

Table 4. Fitted values for loglinear models

| AGE (A) | Drug Use (D) | TB CATEGORIES (C) | Loglinear Models |
|---------|--------------|-------------------|-----------------|
|         |              |                   | (A,D,C) | (DC,A) | (DC,DA) | (AD,AC, DC) | (DAC) |
| < 35    | Yes          | Pulmonary          | 13.69   | 16.72  | 11.63   | 11.5      | 12     |
|         |              | Extrapulmonary     | 3.57    | 0.54   | 0.375   | 0.5       | 0      |
|         | No           | Pulmonary          | 186.98  | 183.95 | 188.06  | 178.5     | 178    |
|         |              | Extrapulmonary     | 48.76   | 51.79  | 52.94   | 62.5      | 63     |
| ≥ 35    | Yes          | Pulmonary          | 11.69   | 14.28  | 19.38   | 19.5      | 19     |
|         |              | Extrapulmonary     | 3.05    | 0.46   | 0.63    | 0.5       | 1      |
|         | No           | Pulmonary          | 159.64  | 157.05 | 152.94  | 162.5     | 163    |
|         |              | Extrapulmonary     | 41.63   | 44.21  | 43.06   | 33.5      | 33     |

By comparing the closest fitted values with the saturated model (DAC) in Table 4, it is found that the best fit model is homogeneous model (AD,AC,DC) compared to other loglinear models.
3.2.1. Likelihood ratio goodness of fit test

| Model     | \( G^2 \) | Degree of Freedom | P-value |
|-----------|----------|------------------|--------|
| (A,D,C)   | 19.304   | 4                | 0.001  |
| (DC,A)    | 10.253   | 3                | 0.017  |
| (DC,DA)   | 6.511    | 2                | 0.039  |
| (AD,AC,DC)| 1.437    | 1                | 0.231  |
| (DAC)     | 0        | 0                | 0      |

Table 5 shows the likelihood ratio goodness of fit test for age (A), drug use (D) and TB categories (C). The smallest likelihood \( G^2 \) statistics 1.437 with the largest p-value 0.231 correspond to the most fitting and parsimonious homogeneous model (AD,AC,DC).

3.2.2. Loglinear cell residual.

Table 6 shows no extreme difference between the fitted and observed values. Both the positives and negatives values for residuals are relatively small, an indication of model fit.

| AGE (A) | Drug Use (D) | TB CATEGORIES (C) | Observed count | Model (AD,AC,DC) |
|---------|--------------|-------------------|----------------|-------------------|
|         |              |                   | Fitted count   | Standardized residual |
| <35     | Yes          | Pulmonary          | 12             | 11.5              | 0.148             |
|         |              | Extrapulmonary     | 0              | 0.5               | -0.707            |
| No      | Pulmonary    | 178               | 178.5          | -0.037            |
|         | Extrapulmonary | 63            | 62.5           | 0.063             |
| ≥35     | Yes          | Pulmonary          | 19             | 19.5              | -0.113            |
|         | Extrapulmonary | 1             | 0.5            | 0.708             |
| No      | Pulmonary    | 163               | 162.5          | 0.039             |
|         | Extrapulmonary | 33            | 33.5           | -0.086            |

3.3 Test of effect sizes on the most Parsimonious model

Table 7. Parameter estimates for fitting homogeneous model

| Age (A), Smoking (S), TB Categories (C) | Age (A), Drug Use (D), TB Categories (C) |
|---------------------------------------|----------------------------------------|
| Parameter                             | Estimate | Parameter | Estimate |
| Constant                              | 3.075    | Constant | 3.512 |
| Age = <35                             | 0.782    | Age = <35 | 0.624 |
| Smoking = Yes                         | -0.562   | Drug Use= Yes | -4.205 |
| TB Categories = Pulmonary             | 1.360    | TB Categories = Pulmonary | 1.579 |
| [Age = <35]*[Smoking = Yes]           | -0.544   | [Age = <35]*[Drug Use = Yes] | -0.622 |
| [Age = <35]*[TB Categories = Pulmonary] | -0.484 | [Age = <35]*[TB Categories = Pulmonary] | -0.530 |
| [Smoking = Yes]*[TB Categories = Pulmonary] | 0.709 | [Drug Use = Yes]*[TB Categories = Pulmonary] | 2.085 |
Table 7 shows the parameter estimates for the most parsimonious homogenous model. The parameters estimates were obtained from fitting the homogeneous loglinear model Age (A), Smoking (S), TB Categories (C) as follows:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^S + \lambda_k^C + \lambda_{ij}^{AS} + \lambda_{ij}^{AC} + \lambda_{jk}^{SC}$$  \hspace{1cm} (16)

$$= 3.075 + 0.782A - 0.562S + 1.360C - 0.544AS - 0.484AC + 0.709SC$$

For parameters interaction $\lambda_{jk}^{SC}$ Smoking and TB Categories, the estimated values is 0.709 with p-value 0.005 (< 0.01) and odds ratio $e^{0.709}$ equal 2.03. The value 2.03 indicates that for patients who smoked, the estimated odds of Pulmonary TB infection is 2.03 times more likely than patients who do not smoke.

The parameter estimates for model Age (A), Drug Use (D), TB Categories (C) is

$$\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^D + \lambda_k^C + \lambda_{ij}^{AD} + \lambda_{ij}^{AC} + \lambda_{jk}^{DC}$$  \hspace{1cm} (17)

$$= 3.512 + 0.624A - 4.205D + 1.579C - 0.622AD - 0.530AC + 2.085DC$$

For parameters interaction $\lambda_{jk}^{DC}$ Drug Use and TB Categories, the estimated values is 2.085 with p-value 0.042 (< 0.05) and odds ratio $e^{2.085}$ equal 8.04. The value 8.04 indicates that patients who are drug users, the estimated odds of Pulmonary TB infection is 8.04 times more likely than patients who are not drug users.

4. Discussion and conclusion

Results of the analysis show that the homogeneous model is the most fitting and parsimonious model that can used to describe the strength of association between age, smoking, drug use and TB categories (Pulmonary and Extrapulmonary). The results found that patients who smoked are 2 times more likely to have Pulmonary TB than non-smokers while drug users are 8 times more likely to have Pulmonary TB compared to patients who are non-drug users. The study also found that patients who have Pulmonary TB are less than 35 years old while Extrapulmonary TB tend to infect older patients. Generally, the number of patients that are infected with Pulmonary TB is higher than Extrapulmonary TB. It is recommended that patients in these categories (smoking, drug users, ageing) require more attention and extensive treatment in order to heal.

5. Acknowledgement

The authors would like to extend their appreciation to Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam for the support on the conference fund and enable this paper to be presented at ICOAIMS 2019.

References

[1] WHO 2018 Tuberculosis and HIV. Retrieved April 2018, from World Health Organization: http://www.who.int/hiv/topics/tb/about_tb/en/

[2] Bernama 2018 Prolonged coughing may be due to TB. Retrieved December 2018, from The Sun Daily: https://www.thesundaily.my/archive/prolonged-coughing-may-be-due-tb-EUARCH534537
[3] Carey E, Higuera V, and Nall R 2018 Pulmonary Tuberculosis. Retrieved 2018, from Healthline: https://www.healthline.com/health/pulmonary-tuberculosis

[4] Kulchavenya E 2014 Extrapulmonary tuberculosis: are statistical reports accurate? Retrieved December 2018, from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4072044/

[5] Azmi A 2013 Prevalence of comorbidities & risk factors for Multidrug Resistance among Tuberculosis patients in Institute of Respiratory Medicine, Kuala Lumpur / Alia Azmi. Retrieved 2018, from Universiti Teknologi Mara: http://ir.uitm.edu.my/id/eprint/15474/

[6] Mahmud Z, Kamal I S M, Sulong S, and Azid N N N 2014 Examining potential risk factors to acute pancreatitis disease: A comparison of loglinear models in a Malaysian case study. Journal of Medical Sciences (Faisalabad), 14(4), pp 153-161. https://doi.org/10.3923/jms.2014.153.161

[7] Deni S M and Jemain A A 2008 Mixed geometric truncated poisson model for sequences of wet days. Journal of Applied Sciences, 8(21), pp 3975-3980. https://doi.org/10.3923/jas.2008.3975.3980

[8] Agresti A 2007 An Introduction To Categorical Data Analysis. (New Jersey: John Wiley)

[9] Ismail A, Campbell M J, Ibrahim H M and Jones G L 2006 Health related quality of life in Malaysian children with thalassaemia. Health and Quality of Life Outcomes, 4, [39]. https://doi.org/10.1186/1477-7525-4-39

[10] Hogg R V and Tanis E A 2010 Probability and Statistical Inference. (New Jersey: Pearson Prentice Hall)

[11] Szumilas M 2010 Explaining Odds Ratio. NCBI. Retrieved 2018 from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/