Implementation of Chi Square Automatic Interaction Detection (CHAID) Method to Identify Type 2 Diabetes Mellitus in Tuberculosis Patient. A Case Study in Cipto Mangunkusumo Hospital

T Siswantining¹, Maghfiroh¹, T Kamelia², D Sarwinda¹

¹Department of Mathematics, FMIPA Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia  
²Division of Pulmonology, Department of Internal Medicine, FK Universitas Indonesia / Dr. Cipto Mangunkusumo Hospital, Jl. Diponegoro No. 71, Central Jakarta 10430, Indonesia

*Corresponding Author: titin@sci.ui.ac.id

Abstract. A Pulmonary Tuberculosis (pulmonary TB) is a chronic infectious disease caused by Mycobacterium tuberculosis. Chronic infection cause the body in a state of oxidative stress. In the state of stress, stress hormone production increases and can affect the increase of blood sugar levels which then trigger the occurrence of diabetes mellitus (DM). The purpose of this study is to determine the factors associated with the emergence of DM Type 2 and make a classification to characterize DM Type 2 in patients with pulmonary TB. In this research, the data used are secondary data of pulmonary TB patients obtained from Cipto Mangunkusumo Hospital (RSCM) for six years, 2012 - 2017. Chi Square Automatic Interaction Detection (CHAID) method is used to classify categorical data by dividing the data set into subgroups. The dependent variable is type 2 DM status, and the independent variables are gender, age, body mass index, level of neutrophil, level of lymphocytes, erythrocyte sedimentation rate, and anti-tuberculosis medicines such as rifampicin, isoniazid, pyrazinamide, ethambutol, and streptomycin. There are seven classes are obtained by classification using CHAID. The result from analysis of CHAID method shows that the factors related to the occurrence of type 2 DM on TB patients are age, body mass index, gender and pirazinamid (Z). Based on CHAID method, a classification of the occurrence of type 2 DM on pulmonary TB patients is obtained, which is pulmonary TB patients who are ≥ 40 years old, have body mass index ≥ 25 kg/m², and the most patients are male (Class-4th).

1. Introduction
Tuberculosis (TB) and diabetes mellitus (DM) are two major health problems which are epidemiologically and globally significant because they are chronically related diseases [1]. Tuberculosis (TB) is an infectious chronic disease caused by Mycobacterium tuberculosis. Most TB bacteria invade the lungs, but can also invade other organs, such as skin, eyes, lymph glands, bone, lining of the brain, and so on [2]. Based on WHO data in 2013, Indonesia ranks fourth after India, China and South Africa as the country with the highest incidence of TB in the world, while diabetes mellitus (DM) is a metabolic disorder disease characterized by elevated blood sugar levels and carbohydrate metabolism abnormalities, fat, and protein are caused by insulin secretion abnormalities, insulin work
or both [3]. The number of DM patients worldwide is estimated at 285 million people and this number will increase up to 438 million people by 2030 [4].

Chronic infection causes the body in a state of oxidative stress. Oxidative stress is a condition where there is an imbalance between pro oxidants (free radicals) and antioxidants. When the body is in a state of stress, the production of stress hormones increases as epinephrine, cortisol, and glucagon which work synergistically. Increased production of these hormones can affect the increase in blood sugar levels (hyperglycemia). Increased blood sugar levels will affect the production of insulin from pancreatic beta cells. As a result, high blood sugar levels in prolonged time will trigger insulin resistance in muscle tissue causing diabetes [5,6].

Research conducted by [7] conducted in Nigeria that TB patients with impaired blood glucose tolerance had normal stress results after three months of tuberculosis treatment, while in Turkey the oral glucose tolerance test (TTGO) was examined to 58 TB patients and 23 patients with pneumonia. In TB patients, 10% of glucose intolerant and 9% had diabetes, while pneumonia patients had 17% of DM and no glucose intolerance. Both studies show that infection causes reversible glucose intolerance.

The above-mentioned study suggests that TB patients have a tendency to develop DM. This needs to be investigated in Indonesia regarding to the variables suspected to be associated with DM status in TB patients, especially in Cipto Mangukusumo Hospital (RSCM) so that the management of DM-TB patients can be improved.

Problems:

How is the classification of Type 2 DM patients based on the factors which are most significantly related to the appearance of Type 2 DM in pulmonary tuberculosis patients?

Objectives:

The objective of this research is to classify the characteristics of Type 2 DM on pulmonary tuberculosis patients based on the factors that are most significantly related.

Scope of problem:
1. This study uses medical records of RSCM taken from 2012 to 2017.
2. The patients studied were patients with pulmonary tuberculosis who are inpatient or outpatient at pulmonology clinic of RSCM.

2. Research variables

The variables that will be involved in this research are : status of type 2 diabetes mellitus on patients, which is the presence or absence of type 2 DM on Pulmonary tuberculosis patients; gender of the patients, which consists of two categories namely female and male; age of the patients with pulmonary tuberculosis, which is calculated from the time the patient is born until admitted to the hospital with the diagnosis of Type 2 DM, expressed in years; Body Mass Index (BMI) is an assessment of adult nutritional status that is measured by body weight in kilograms (kg) divided by height in squared meters (m²);

Neutrophil level is the percentage of neutrophils in the blood of pulmonary tuberculosis patients based on the blood test results; Lymphocyte level is the percentage of lymphocyte in the blood of pulmonary tuberculosis patients based on the blood test results; Erythrocyte Sedimentation Rate (ESR) is the rate of erythrocyte cells sediment in the plasma expressed in millimeter (mm); Rifampicin (R) is an anti-tuberculosis drug which consists of two categories which are consuming or not consuming; Isoniazid (H) is an anti-tuberculosis drug which consists of two categories which are consuming or not consuming; Pyrazinamide (Z) is an anti-tuberculosis drug which consists of two categories which are consuming or not consuming; Etambutol (E) is an anti-tuberculosis drug which consists of two categories which are consuming or not consuming; Streptomycin (S) is an anti-tuberculosis drug given by injection.

3. Methods

The Population in this research is patient with pulmonary tuberculosis at RSCM who were registered in medical records in 2012-2017. The number of sample is 30 patients with pulmonary tuberculosis at RSCM, and used purposive sampling. The method for data analysis is Chi-square Automatic Interaction Detection (CHAID).
3.1 Analysis of Chi-square Automatic Interaction Detection (CHAID) Method

CHAID is a method to classify categorical data by dividing the data set into subgroups based on dependent variable [8]. According to [9] the CHAID method is an iterative technique that tests one by one independent variables used in classification and arrange them based on the significance level of the chi-square test of the dependent variable. The results of the classification are shown in the form of a tree diagram. The best independent variable that will form the first branch in the resulting tree diagram.

Before the process of calculating the CHAID algorithm, [10] divides the independent variables into three types:

1. Monotonic Variables
   The categories in these variables can be combined if they are adjacent to each other, i.e. variables whose categories indicate a sequence, such as ordinal data. Examples: age, income, and level of education.

2. Free Variable
   The categories in these variables can be combined even if they are not close to each other or do not pay attention to the sequence, such as nominal data. Examples: job, geographic area.

3. Floating Variable
   The categories in this variable will be treated like monotonic variables, except for the missing value which can be combined with any category.

3.2 Chi-square test ($\chi^2$)

Chi-square test is a non-parametric statistical test that can be used to test the equality of proportion, independence test, and goodness of fit test or suitability test. The chi-square test in the CHAID method according to [10] is used to determine whether the categories in the independent variables are uniform. Furthermore, the chi-square test statistic is then used to test the independence between variables, in this case it will be determined which independent variables are most significantly related to the dependent variable.

Suppose the first variable has a category b, i.e. $A_1,A_2,\ldots,A_b$ and the second variable has k categories, i.e. $B_1,B_2,\ldots,B_k$. The number of observations in the i-th category (i = 1,2,3,……b) on the first variable and the j-th category (j = 1,2,3,……k) in the second variable is expressed by $O_{ij}$ (Table 1).

| Table 1. Structure of Chi-Square Test Data $b \times k$ |
|-----------------------------------------------|
| Variable 2 | B1 | B2 | B3 | ... | Bk | Jumlah |
| Variable 1 |     |    |    |      |    |        |
| $A_1$     | $O_{11}$ | $O_{12}$ | $O_{13}$ | ... | $O_{1k}$ | $O_1$ |
| $A_2$     | $O_{21}$ | $O_{22}$ | $O_{23}$ | ... | $O_{2k}$ | $O_2$ |
| $A_3$     | $O_{31}$ | $O_{32}$ | $O_{33}$ | ... | $O_{3k}$ | $O_3$ |
| $A_b$     |    |    |    |      |    |        |
| Total     | $n_{11}$ | $n_{22}$ | $n_{33}$ | ... | $n_{bk}$ | $n_b$ |

Where: $O_{ij}$ = Number of observations on the i-th row and the j-th column; $n_i$ = Total observations on the i-th row; $n_j$ = Total observations on the j-th row; n = Total number of respondents.

The expected value for each observation is

$$E_{ij} = \frac{n_i n_j}{n}$$

The following are the independency test steps

1. Hypothesis
   $H_0$: Both variables are independent
   $H_1$: not so
2. Determine the level of significance $\alpha$
3. Statistical value of the test
4. Make a decision rule
5. Make a conclusion from the outcome of the decision

The chi-square test described above is a chi-square test of independence, whereas in the CHAID method chi-square test is also used to test the equality of proportion and independence test. To test the similarity of proportions, the hypothesis is

$$H_0: p_1 = p_2$$
$$H_1: p_1 \neq p_2$$

The next steps would be similar to the previous steps as described above.

3.3 CHAID Algorithm

The CHAID algorithm is used to separate and merge categories in the variables used in the analysis. Based on [11], the outline of this algorithm is divided into three stages, namely merging, splitting, and stopping.

a. Merging Stage

The merging stage is performed to combine categories of independent variables that have more than two categories. To perform this merger stage examined the significance of each category of independent variables to the dependent variable. The steps for the merger stage are as follows

1. For each variable $X_1, X_2, \ldots, X_k$, create a contingency table of size $2 \times J$ formed by a pair of categories from the category of independent variables with their dependent variables having as many as $J$ categories.
2. Calculate the statistical value of chi-square test on each pair of categories for each independent variable to test the equality of proportions.
3. Pay attention to the chi-square test statistic value for each category pair, if smallest $\chi^2_{\text{count}} < \chi^2_{\text{table}}$, then the category is combined into a new category.
4. Re-examine the significance of the new categories. If there are still categories of independent variables that have more than two categories then repeat steps 2 and 3, but if it is significant go to the next step.

Recalculate the statistical value of the chi-square test to determine which independent variables are most significantly related to the dependent variable. After the chi-square value is obtained, calculate the p-value of each significant independent variable.

5. Make a Bonferroni correction for the p-value obtained in step 5.

b. Partition Stage

In this stage of separation, the variables will be partitioned or divided groups. To perform the separation, the most significant independent variables will be used as split nodes by looking at the Bonferroni corrected p-value values of the merging stages on each independent variable by choosing the smallest p-value value $< \alpha$.

c. Termination Stage

Before the termination stage, repeat the merging stage to analyze subsequent subgroups. The termination will be performed if there are no more significant independent variables that indicate any difference in the dependent variables.

4. Results and Discussion

The data obtained in this study were analyzed using Chi-square Automatic Interaction Detection (CHAID) method with SPSS software version 23. The dependent variable in this study was the status of DM Type 2 in pulmonary TB patients divided into two categories, i.e. yes and no, while independent variables existed as many as 11 variables expressed in category form and determined in consultation with physician and laboratory clinical reference.
From Figure 1 above, we found that significant factors related to the emergence of Type 2 DM in pulmonary tuberculosis patients were age, body mass index (BMI), sex, pyrazinamide (Z). The CHAID tree diagram above illustrates that pulmonary TB patients are divided into 7 classifications or groups by reading the results of the tree diagram following the top-down stopping rule, i.e. starting from the parent node, i.e. age. Then under it is a subgroup (child node) which is obtained from the division of the parent node up to the bottom end of the tree called the terminal node (Table 2).

**Table 2. Classification of pulmonary TB patients**

| Classification | Characteristics |
|----------------|-----------------|
| 1              | Pulmonary TB patients with age ≥ 40 years and body mass index (BMI) ranged from 18.5-24.9 kg/m² |
| 2              | Pulmonary TB patients with age ≥ 40 years, BMI < 18.5 kg/m², and taking pyrazinamide (Z) |
| 3              | Pulmonary TB patients with age ≥ 40 years, BMI < 18.5 kg/m², and not taking pyrazinamide (Z) drugs. |
| 4              | Pulmonary TB patients with age ≥ 40 years with BMI ≥ 25 kg/m², and male sex. |
| 5              | Pulmonary TB patients with BMI ≥ 25 kg/m², and female sex. |
| 6              | Pulmonary TB patients aged 18-39 years with BMI < 25 kg/m². |
| 7              | Pulmonary TB patients aged 18-39 years with BMI ≥ 25 kg/m². |

From the seven classifications that are formed, it can be seen in Table 3 that the number and percentage of each of the TB patients with Type 2 DM and not having Type 2 DM are as follows
Table 3. Percentage of results for each classification of pulmonary TB patients

| Classification Number | Number of DM Patients | Percentage | Number of non-DM Patients | Percentage |
|-----------------------|-----------------------|------------|---------------------------|------------|
| 1                     | 35                    | 39.33%     | 54                        | 60.67%     |
| 2                     | 7                     | 15.91%     | 37                        | 84.09%     |
| 3                     | 1                     | 100.00%    | 0                         | 0.00%      |
| 4                     | 13                    | 86.67%     | 2                         | 13.33%     |
| 5                     | 9                     | 45.00%     | 11                        | 55.00%     |
| 6                     | 0                     | 0.00%      | 54                        | 100.00%    |
| 7                     | 1                     | 14.29%     | 6                         | 85.71%     |

The largest percentage of pulmonary tuberculosis patients with Type 2 DM is in the 3rd classification, which is 100%, but in that classification or group there is only one pulmonary TB patient with Type 2 DM, so the classification does not represent the population well. Therefore, to see the characteristics of pulmonary tuberculosis patients with Type 2 diabetes, the 4th classification was selected with a percentage of 86.67% with the characteristics of patients aged ≥ 40 years, having BMI ≥ 25 kg / m², and male sex. This is in accordance to a study by [12] which shows that in adults aged 18-45 in China, it is known that obesity is positively correlated with the emergence of Type 2 DM in patients with pulmonary TB; the prevalence of DM incidence was higher when BMI ≥ 23.41 kg / m². Research conducted by [13] showed that factors associated with the emergence of Type 2 DM in patients with pulmonary tuberculosis were age and body mass index (BMI). [14] in Czech and Finland shows that the number of men with Type 2 diabetes is more dominant than women.

The following is an accurate classification result obtained from data analysis using CHAID

Table 4. Accuracy of the classification results

| Observed | Predicted | Percent Correct |
|----------|-----------|-----------------|
| not DM   | not DM    | 98.8%           |
| DM       | DM        | 21.2%           |
| Overall  | 93.0%     | 7.0%            | 76.5%                  |

From Table 4 above, it was found out that the accuracy of overall pulmonary tuberculosis patient predictions using CHAID was 76.5% with prediction error of 23.5%. On the accuracy of classification results to predict Type 2 diabetes by 21.2%. This is due to limitations on research data obtained in the RSCM medical record where the number of pulmonary TB patients with Type 2 DM and the number of pulmonary TB patients with Type 2 DM were out of balance.

5. Conclusion

Based on the results of analysis by using CHAID method, the conclusion obtained is that the main factor associated with the emergence of DM Type 2 in patients with pulmonary TB is age. In addition, there are also gender factors, body mass index (BMI), and pyrazinamide (Z). The results of the classification of pulmonary tuberculosis patients with Type 2 DM based on the factors that have been obtained are patients with characteristics aged ≥ 40 years, have BMI ≥ 25 kg / m², and male sex with percentage of accuracy is 86.67%.

Acknowledgements

This research is supported by PITTA UI 2018 Research Grant from DRPM Universitas Indonesia.
References
[1] Aregbesola A 2016 Gender difference in Type 2 Diabetes and the role of body Annals of Clinical Biochemistry pp. 1-24
[2] Bagozzi R P 1994 Advanced methods of marketing research. Blackwell Publishers Ltd Oxford
[3] Departemen Kesehatan Republik Indonesia 2011 Pedoman nasional penanggulangan Tuberkulosis (Jakarta: Direktorat Jendral Pengendalian Penyakit dan Penyehat Lingkungan)
[4] Departemen Kesehatan Republik Indonesia 2014 Pedoman nasional penanggulangan Tuberkulosis (Jakarta: Direktorat Jendral Pengendalian Penyakit dan Penyehat Lingkungan)
[5] Dooley K E and Chaisson R E 2009 Tuberculosis and Diabetes Mellitus: Convergen of two epidemics 9(12) pp. 737-746
[6] Gallagher C A 2000 An iterative approach to classification analysis
[7] Jing Cai 2017 Association between body mass index and Diabetes Mellitus in Tuberculosis patients in China: A community based cross-sectional study BMC Public Health 17 p. 228
[8] Kelestimur F 2004 The endocrinology of adrenal Tuberculosis: The effects of Tuberculosis on the hypothalamo-pituitary-adrenal axis and adrenocortical function Journal of Endocrinol Invest 27(4) pp. 380-386
[9] Kibirige D 2014 Endocrine dysfunction among adult patients with Tuberculosis: An African experience Indian Journal of Endocrinology and Metabolism 18(3)
[10] Perkumpulan Endokrinologi Indonesia (PERKENI) 2015 Konsensus pengelolaan dan pencegahan Diabetes Mellitus Tipe 2 di Indonesia 2015 (Indonesia: PB PERKENI)
[11] Syam M S 2013 Dukungan sosial penderita Tuberkulosis paru di wilayah kerja Puskesmas Ajangale Kebupaten Bone tahun 2013
[12] Viswanathan V 2012 Prevalence of diabetes and pre-diabetes and associated risk factors among Tuberculosis patients in India
[13] World Health Organization and the International Union Against Tuberculosis and Lung Disease 2011 Collaborative framework for care and control of Tuberculosis and diabetes (Geneva: WHO Press)
[14] Sayadi M, M Zibaeenezhad, S M T Ayatollahi 2017 Simple Prediction of Type 2 Diabetes Mellitus via Decision Tree Int Cardiovasc Res J, 11(2) pp. 71-76