Open Journal of Social Sciences, 2018, 6, 29-36
http://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

# Predicting the Relapse Category in Patients with Tuberculosis: A Chi-Square Automatic Interaction Detector (CHAID) Decision Tree Analysis

**Arnold Peralta Dela Cruz**

Nueva Ecija University of Science and Technology, Cabanatuan City, Philippines
Email: arnoldneust@gmail.com

**How to cite this paper:** Dela Cruz, A.P. (2018) Predicting the Relapse Category in Patients with Tuberculosis: A Chi-Square Automatic Interaction Detector (CHAID) Decision Tree Analysis. *Open Journal of Social Sciences*, **6**, 29-36.
https://doi.org/10.4236/jss.2018.612003

**Received:** November 9, 2018
**Accepted:** December 11, 2018
**Published:** December 14, 2018

## Abstract

Predicting the outcome of treatment among TB patients is a big concern of the Department of Health. Data mining in health care system can be used for decision making. The most widely used for data exploration is decision tree based on divide and conquer technique. The objectives of this article are to create a predictive data mining model for TB patient category to find the relapse treatment and to classify the factors influencing the relapse treatment to provide assistance, guidance, and appropriate warning to TB patients who are at risk. The dataset of TB patient records is verified and applied in CHAID classification tree algorithm using SPSS Statistics 17.0. The classification tree model identified the set of two statistically significant independent variables (DSSM Result, Age) as predictors of patient category.

## Keywords

Data Mining, CHAID Algorithm, Decision Tree, Relapse, Tuberculosis

## 1. Introduction

Philippine Tuberculosis (TB) is a foremost community health problem and remains a major cause of death and it is one of the nations with high TB incidence. "Philippines ranked ninth among the 22 high TB burdened countries" [1]. In 2015, 14,000 Filipinos died from tuberculosis and 4.8 million from this number are mostly poor [2]. TB is a treatable and preventable disease, yet many are still infected and are continuously suffering.

In TB treatment, one major problem is guaranteeing patients to pursue their treatment, together with medication and medical checkups till completion. Hence, there's a desire to boost the adherence and retention in care. Whereas there could also be several reasons for the lack of endurance and there could also be ways to boost completion of treatment programs by maintaining better contact between health workers and TB patients. Treatment results fill in as intermediary proportions of the nature of tuberculosis treatment provided by the health care system, and it is essential to assess the effectiveness of Directly Observed Therapy-Short course program in controlling the disease, and diminishing treatment failure, default and death [3]. TB patient relapse needs immediate retreatment, or fail following preliminary treatment success. "Results among patients getting a standard World Health Organization Category II retreatment routine are imperfect, resulting in increased risk of disease, transmission, and drug resistance" [4]. Tuberculosis category relapse is divided into two classes: firstly, a patient on whom the first onset has been treated, but the remaining mycobacterium tuberculosis restarts into a second onset of TB; and secondly, a patient with reinfection with new mycobacterium tuberculosis [5].

In this study, the Chi-Square Automatic Interaction Detection (CHAID) decision tree algorithm is employed to predict the patient category relapse in Cabanatuan City, Philippines. "CHAID applications focused on the field of medical and psychiatric research although it can be employed also in researches of different fields. The technique was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic" [6]. CHAID is utilized for prediction, for classification and for recognition of interconnection among variables. CHAID, because of its usefulness, was utilized in several studies. The authors in [7] use CHAID to form associations/structured relationships between factors in the classification of observations of the quality of housing eligibility in Kupang Regency while researchers in [8] utilized CHAID to dig up information on the potentials of local food availability of corn in regencies and cities in Java Island.

Other researches use CHAID to: explore the adverse effects of social networking sites on students' academic performance in secondary schools [9]; compare the quality of the information obtained on tourism market segmentation [10]; and determine the anemic status of infants as well as the risk factors in a representative downtown area of Beijing [11].

The reasons behind the selection of decision trees as the basis method of this study can be enumerated like: 1) CHAID decision tree model is understandable, easy, and interpretable model 2) and it is fast to build on the predictive methods, then it's highly appropriate and flexible for future changes of data as it has low training time [12]. Furthermore, for this kind of task, the decision tree technique has high prediction accuracy in many fields to make them preferable and trustable choices. With the support of this method, treatment category relapse patterns are exposed and a mostly accurate predictor is attained for treatment priority prediction.

This research aims to create a predictive data mining model for relapse category rate of TB by using Integrated Tuberculosis Information System (ITIS) data on reported cases and found variables influencing the relapse treatment of TB.

## 2. Methodology

The study is a quantitative research design that uses the statistical method to quantify and analyze the data to generalize results from a sample population. It was done in Cabanatuan City, Nueva Ecija, Philippines. The data came from dataset of TB patient records of Cabanatuan which were extracted from the database of Integrated Tuberculosis Information System (ITIS) of Cabanatuan City Health Office last 2017. This confirms the correctness and comparability of data, which are significant features in CHAID model. The collected data were coded and scrutinized using the Statistical Package for Social Sciences (SPSS Statistics 17.0). The study protocol of data collection and interview was approved by the Office of the City Mayor and the director of the City Health Office.

The study used data mining as a tool with CHAID classification tree as a technique to design the TB patient category relapse prediction model. According to [16], "data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends". Classification trees are broadly utilized in various fields such as botany, medicine, computer science, and psychology [13]. "These classification trees promptly give themselves to being presented graphically, assisting to make them easier to analyze than they would be if only a strict numerical interpretation were possible". "CHAID considered one of the classification tree algorithms is the name quantified to one version of the Automatic Interaction Detector that has been developed for categorical variables" [14]. Actually, CHAID is a system that halves a population into independent and particular portions. These portions called nodes are split in such a way that the disparity of the response variable is limited inside the portions and make the most of among the parts. The output of CHAID prediction model is presented in hierarchical tree-structured method, in which the root is the population, and the branches are the associating segments such that the variation of the response variable is limited within all the segments, and maximized among all the segments. "The important step in CHAID prediction model structure is selecting the significant features for classification and the purpose of feature selection techniques supports the reduction of computation time and increases the predictive accuracy of the model" [15].

## 3. Results and Discussions

### 3.1. CHAID Algorithm in Predicting Patient Category

The CHAID (Chi-Square Automatic Interaction Detection) algorithm is one of the most prevalent statistically based methods of supervised learning for decision tree development proposed by a statistician Kass in the late 1970's. The CHAID acronym denotes automatic and iteration technique of tree development based on Pearson's Chi-square statistic and corresponding p-value. The CHAID analysis builds a predictive model to help define how variables are best unified to describe the result in the specified dependent variable.

### 3.2. CHAID Analysis Application

In order to form a decision tree by means of CHAID algorithm, according to its nature, initially, a description of the used variables was achieved as follows: The variable, *Patient Category* is defined as a dependent variable. *Patient Category* is a nominal variable with two values (non-relapse, relapse), the creation model can be based on Chi-square splitting criterion.

As observed on **Table 1**, as to age, 404 patients were 51 and below and 106 were more than 51 years old. In terms of sex, the data revealed that there were more males than females. In their Bacterial status, 49% of the patients were under clinically-diagnosed TB while 51% belong to Bacteriologically-confirmed TB. Almost 100% of the Tuberculosis cases on the data were under pulmonary while in patients' category majority linked to non-relapse.

### 3.3. Modeling Results Analysis

The most significant independent variable in **Figure 1** is *DSSM Result* of TB patients. It has the most power in division of observations into groups and most strongly associated with the dependent variable. (Statistical significance of *DSSM Result* was determined using following values: *chi-square* = 111.459, *df* = 3, *p-value* = 0.000.) As the first discriminator, the *DSSM Result* splits the root node into four groups with 510 respondents presented as node 1, node 2, node 3, and node 4.

Most of the respondents (231) go to node 1 where in the value of *DSSM Result* are "2+" and "ODT". The 87 respondents belong to node 2 containing *DSSM Result* is equal to "1+". Node 3 has 130 respondents where *DSSM Result* is equal to "0", and the rest of the participants (62) belong to node 4 where *DSSM Result* is equal to "3+". For *DSSM Result*, within the first level of the tree, node 3 is parent node, while nodes 1, 2, and 4 are all terminal.

The second level of the tree is variable *Age* which is statistically significant. Independent Variable *Age* is significant for splitting of node 3 (*Chi-square* = 144.870, *df* = 1, *p-value* = 0.001). In congruence to this, the following two groups of respondents are found: TB patients with *Age* of less than or equal to 51 (≤51) belong to node 5, while Age is greater than 51 (>51) belong to node 6. All the nodes are terminal in the final level of decision tree.

The four (4) terminal nodes in the formed tree structure are marked as 1, 2, 4 and 5 relates to *Non-Relapse*, however node 6 refers to *Relapse*. Actually, the lanes from the root to terminal nodes produce a set of rules for classification of TB patients into one of the defined categories of the variable *Patient Category*. This obviously specifies that the developed model and knowledge described in the decision tree can be formulated as if-then rules.

**Table 1.** Structure of variables used in CHAID analysis.

| Variable Name | Value (modalities) | Structure | | Type of Variable | |
|---|---|---|---|---|---|
| | | *Fi* | % | MS | IV/DV |
| Age | ≤51 | 404 | 79 | SV | IV |
| | >51 | 106 | 21 | | |
| Sex | Male | 343 | 67 | NV | IV |
| | Female | 167 | 33 | | |
| BacStatus (Bacteriogically Status) | Bacteriologically-confirmed TB | 258 | 51 | NV | IV |
| | Clinically-diagnosed TB | 252 | 49 | | |
| | Other Diagnostic Test (ODT) | 174 | 34 | | |
| DSSM Result (Direct Sputum Smear Microscopy) | 0 | 130 | 26 | NV | IV |
| | 1+ | 87 | 17 | | |
| | 2+ | 57 | 11 | | |
| | 3+ | 62 | 12 | | |
| Classification | Pulmonary | 507 | 99 | NV | IV |
| | Extra-Pulmonary | 3 | 1 | | |
| Patient Category | Non-Relapse | 455 | 89 | NV | DV |
| | Relapse | 55 | 11 | | |

*Legend*: MS is Measurement Scales; NV is Nominal Variable; SV is Scale Variable; *fi* is frequency; % is Percentage; IV is Independent Variable; DV is Dependent Variable.

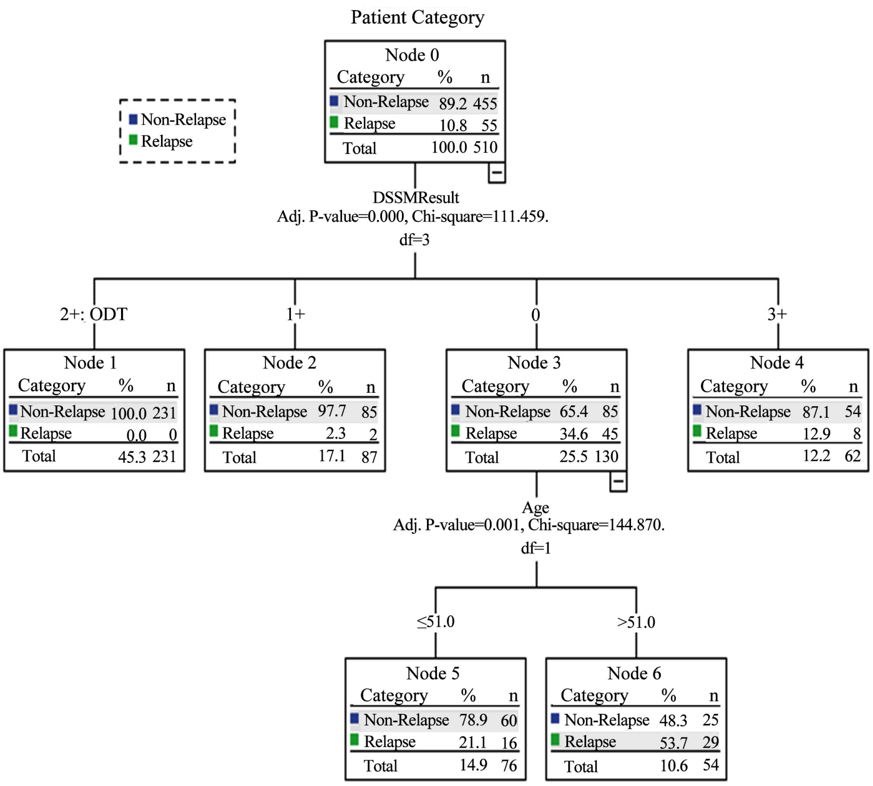

**Figure 1.** CHAID decision tree diagram.

### 3.4. CHAID Model Rule Sets

IF (DSSMResult = “2+”) OR (DSSMResult = “ODT”) THEN
Node = 1
Prediction = 1
Probability = 1.000
IF (DSSMResult = “1+”) THEN
Node = 2
Prediction = 1
Probability = 0.977
IF (DSSMResult = “3+”) THEN
Node = 4
Prediction = 1
Probability = 0.871
IF (DSSMResult = “0”) AND (AGE < = 51) THEN
Node = 5
Prediction = 1
Probability = 0.789
IF (DSSMResult = “0”) AND (AGE > 51) THEN
Node = 6
Prediction = 2
Probability = 0.537

Based on the rule set of CHAID algorithm using *Patient Category* as the dependent variable, prediction of node 1 is 1 referring to *Non-Relapse* category with a probability of 1.000; prediction of node 2 is 1 referring to *Non-Relapse* category with probability of 0.977; prediction node 4 is 1 referring to *Non-Relapse* category with probability of 0.871; prediction node 5 is 1 referring to *Non-Relapse* category with probability of 0.789 and; prediction node 6 is 2 referring to *Relapse* category with probability of 0.537. The nodes 1, 2, 4, and 5 have a prediction value of 1 referring to *Non-Relapse*, and node 6 has a prediction value of 2 referring to *Relapse*. The CHAID method shows that the variable *DSSMResult* is the best predictor in *Patient Category*. For the *DSSM result*, result “2+” is the significant predictor with 100% result for *Non-Relapse*.

For the growing stage, the next best predictors are the *DSSM Result* and *Age* with 78.9%, if the *DSSMResult* is “0” and *Age* is less than equal to 51 the category is *Non-Relapse*. For the last stage, 53.7% result for *Relapse* category, if the *DSSMResult* is “0” and *Age* is greater than 51 and this is considered a terminal node.

### 3.5. Classification Model Accuracy Assessment

Prediction risk was presented in Table 2 as a percentage of incorrectly classified observation. To be accurate, the findings suggest that, if the characteristics of a TB patient in terms of the two independent variables (*DSSMResult*, *Age*) are identified, the prediction risk that the TB patients will be incorrectly classified in

Table 2. Prediction risk.

| Estimate | |
|---|---|
| Re-substitution | Cross-validation |
| 0.100 | 0.100 |

Table 3. Classification matrix.

| Patient Category | | Predicted | | |
|---|---|---|---|---|
| | | Non-Relapse | Relapse | Percent Correct |
| Observed | Non-Relapse | 430 | 25 | 94.50% |
| | Relapse | 26 | 29 | 52.70% |
| | Overall Percentage | 89.40% | 10.60% | **90.00%** |

relations of TB patient category of 10%, and the risk of 10% in cross-validation for the test sample used.

Classification matrix was presented in Table 3 containing by categories of the empirical and modeled values, dependent variable, and predicted classifications. In line with the abovementioned, it can be specified that overall accuracy of the model is 90%. The model has accurately categorized 459 out of 510 TB patients in the observed sample. Observed by the categories of the dependent variable, significant differences in classification accuracy can be seen.

## 4. Conclusions

In this study, CHAID classification tree technique is used for prediction on the dataset of 510 TB patients to predict and analyze the patient category relapse. The CHAID prediction model was very convenient and useful to evaluate the coherence among variables that are utilized to predict the relapse in TB treatment category. A model was developed based on TB patient correlated input variables gathered from the ITIS database of city health office. The variables *DSSM Result*, and *Age* are the strongest indicators for the prediction of patient category relapse treatment. From the classification matrix, it is clear that 90% is the overall accuracy of the model, and only 10% in prediction risk.

As a future work, the author is planning to create models with a three-year period of dataset to attain more precise results, and engage additional techniques from the dataset.

## Acknowledgements

The author is thankful for the support offered to him by the Provincial Health Office of Nueva Ecija and Cabanatuan City Health Office.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

[1] 2010-2015 Philippine Plan of Action to Control Tuberculosis (PhilPACT). http://www.nationalplanningcycles.org/sites/default/files/country_docs/Philippines/philippine_tb_plan_2010-2015_draft.pdf

[2] Diaz, J. and Crisostomo, S. (2017) TB Still a Major Cause of Death among Filipinos. Press Reader, The Philippine Star.

[3] Getnet, F., Sileshi, H., Seifu, W., Yirga, S. and Alemu, A.S. (2017) Do Retreatment Tuberculosis Patients Need Special Treatment Response Follow-Up beyond the Standard Regimen? Finding of Five-Year Retrospective Study in Pastoralist Setting. *BMC Infectious Diseases,* **17**, 762. https://doi.org/10.1186/s12879-017-2882-y

[4] Dooley, K.E., *et al.* (2011) Risk Factors for Tuberculosis Treatment Failure, Default, or Relapse and Outcomes of Retreatment in Morocco. *BMC Public Health*, **11**, 140. https://doi.org/10.1186/1471-2458-11-140

[5] Heldal, E., Döcker, H., Caugant, D.A., *et al.* (2000) Pulmonary Tuberculosis in Norwegian Patients. The Role of Reactivation, Re-Infection and Primary Infection Assessed by Previous Mass Screening Data and Restriction Fragment Length Polymorphism Analysis. *International Journal of Tuberculosis and Lung Disease*, **4**, 300e7.

[6] Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, **29**, 119-127. https://doi.org/10.2307/2986296

[7] Atti, A. and Dodo, D. (2018) Chi-Square Automatic Interaction Detection (Chaid) Analysis for Home Quality Status Segmentation. *American Journal of Engineering Research*, **7**, 183-188.

[8] Susanti, Y., Zukhronah, E., Pratiwi, H., Respatiwulan and Sulistijowati, H. (2017) Analysis of Chi-Square Automatic Interaction Detection (CHAID) and Classification and Regression Tree (CRT) for Classification of Corn Production. *Journal of Physics: Conference Series*, **909**, Article ID: 012041. https://doi.org/10.1088/1742-6596/909/1/012041

[9] Onoja, A.A., Babasola, O.L. and Ojiambo, V. (2018) Chi-Square Automatic Interaction Detection Modelling of the effects of Social Media Networks on Students Academic Performance. *Journal of Statistics and Mathematical Sciences*, **4**, 32-39.

[10] Perez, F. and Cejas, M. (2016) CHAID Algorithm as an Appropriate Analytical Method for Tourism Market Segmentation. *Journal of Destination Marketing & Management*, **5**, 275-282. https://doi.org/10.1016/j.jdmm.2016.01.006

[11] Ye, F., Chen, Z.H., Chen, J., Liu, F., Zhang, Y., Fan, Q. and Wang, L. (2016) Chi-Squared Automatic Interaction Detection Decision Tree Analysis of Risk Factors for Infant Anemia in Beijing, China. *Chinese Medical Journal*, **129**, 1193-1199. https://doi.org/10.4103/0366-6999.181955

[12] Tang, Z. and MacLennan, J. (2005) Data Mining with Sql Server 2005. John Wiley & Sons, New York.

[13] Camdeviren, H.A., Yazici, A.C., Akkus, Z., Bugday, R. and Sungur, M.A. (2007) Comparison of Logistic Regression Model and Classification Tree: An Application to Postpartum Depression Data. *Expert Systems with Applications*, **32**, 987-994. https://doi.org/10.1016/j.eswa.2006.02.022

[14] Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistic*, **29**, 119-127. https://doi.org/10.2307/2986296

[15] Witten, I.H. and Frank, E. (2005) Data Mining—Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann Publisher, San Francisco, CA.

[16] Rouse, M. (n.d.). Data Mining. https://searchsqlserver.techtarget.com/definition/data-mining