Predicting TB Death Using Logistic Regression and Decision Tree on VA Data

Muhamad Rifki Taufik[1,a)], Apiradee Lim[1,b)], Phatrawan Tongkumchun[1,c)], Nurin Dureh[1,d)]

[1]*Department of Mathematics and Computer Science, Faculty of Science and Technology,*
Prince of Songkla University, Pattani, Thailand.

[a)]m.rifki.taufik@gmail.com
[b)]apiradee.s@psu.ac.th
[c)] phattrawan.t@psu.ac.th
[d)] dnurin@gmail.com

Abstract- Abstract. In 2015, 10.4 million people got infected with tuberculosis (TB) and 1.8 million died from the disease and over 95% of TB deaths occur in low- and middle-income countries. The aim of this study is to compare the prediction performance between statistical and machine learning method for TB deaths from verbal autopsy (VA) data in Thailand. A total of 9644 VA death record in 2005 was obtained from the Thai Ministry of Public Health which the sampling unit was a registered death of Thai citizen who was a permanent resident in Thailand. The remaining data was 9495 records after data management process include imputing missing value with 99 in age variable. Logistic regression and decision tree models had been employed for predicting people who died TB. The results showed that the predicting accuracy of logistic regression and decision tree were the same which was 98% and AUC of ROC curve indicated logistics has AUC 61.74 higher than decision tree which had AUC 59.28. In conclusion, the performance of prediction for both logistic and decision tree was the same with the small number of cases (195 of 9495).

Keywords: classification accuracy, comparison, ROC, verbal autopsy

INTRODUCTION

Tuberculosis (TB) is one of the top 10 causes of death worldwide. In 2015, 10.4 million people fell ill with TB and 1.8 million died from the disease and over 95% of TB deaths occur in low- and middle-income countries and TB has been declared as an infectious fatal disease mainly among the developing countries includes Thailand [1] . In Thailand, 161 of 100,000 people die due to TB [2]. This record makes Thailand has 2.5–4 times higher TB incidence than Europe and America regions. Later, spreading infection of TB must be controlled to decrease the mortality rate.

Mortality data are the only national statistics source which provides information consistently and continuously. Not only for historical data, but also mortality data is fundamental to evidence-based health policy, monitoring and evaluation [3]. Government is using death data to see the huge pattern of population that could give coverage news to decide important policy such as public health, family plan, etc. Even though death certificate has been developed in current purpose and content within qualified procedure, it does not guarantee the document has good accuracy [4]. They assume that mortality information is accurate.

Inaccuracy at mortality data often occurs in Death Registry (DR). Medical certifiers in countries heavily affected by human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS) can be pressured by family members not to mention AIDS on death certificates [5]. More often, errors are unintentional. Even in Western Europe, where medical certification of death is virtually universal, implausibly high proportions of deaths may be attributed to causes thought to be particularly prevalent [6]. The study determined that only 65 per cent of the observed underlying causes named on the death certificates fell into the groups that were defined as indicating good agreement [4].

In Thailand, 35% of all deaths occur in hospitals, and the cause of death is medically certified by attending physicians. About 15% of hospital deaths are registered with nonspecific diagnoses, despite the potential for greater accuracy using information available from medical records. Further, issues arising from transcription of diagnoses from Thai to English at registration create uncertainty about the accuracy of registration data even for specified causes of death. Procedures for death certification and coding of underlying causes of death need to be streamlined to improve reliability of registration data [7]. Ascertainment of cause for deaths that occur in the absence of medical attention is a significant problem in many countries, including Thailand, where more than 50% of such deaths are registered with ill-defined. Routine implementation of standardized, rigorous verbal autopsy methods is a potential solution [8].

Logistic regression was used and the model had acceptable accurate prediction [9]. Logistic regression with dichotomous outcome variables is especially useful in clinical research [10] but it has limitation. Logistic regression models use linear combinations of variables and, therefore, are not adept at modeling grossly nonlinear complex interactions [11]. Decision tree as one of the popular methods is machine learning does not require any assumptions but also it does not consider about the correlation like statistical analysis. Moreover, machine learning with feasible process and cost-effective is widely used in computer science and other includes public health [12]. Nevertheless, a study which discusses about TB mortality using machine learning never been done before in Thailand.

METHODOLOGY

1. Data source

This study used secondary data from a 2005 VA survey, which assessed the causes of death based on a sample of 9,644 cases (3,316 in-hospital deaths and 6,328 outside-hospital deaths) from 28 districts in nine provinces [13]. The nine provinces selected were Bangkok and two provinces from each of the four regions in Thailand. The selected provinces were those whose numbers of reported deaths were above (one province) and below (one province) the median. Similarly, twenty-eight districts were selected from the provinces. This record provides 21 kinds of diseases but this study is only focus on TB case then records that die from TB coded as one (1) and others as zero (0). Other interesting valiables are age-gender group, Death Registry (DR), place of death (inside or outside hospital), and province where those variables were categorical variables and treated as determinants. All process used R program to manage and analyze the data.

2. Decision Tree

It is called decision tree because the technique of classification is constructed like a tree. This method is able to handle noisy data, continuous variables, variables with multiple (more than two) values, missing data, and classes with multiple values. The nature of the task for a decision-tree program is as follows: The dataset consists of a set of objects (also called observations), each of which belongs to a class. There is a set of attributes (also called variables) with each object having values for the attributes. The task is to use the attributes to find a decision tree that classifies the data appropriately. Since there is a large number of such a tree, the task is refined to find a small tree that classifies the data appropriately. If the set includes “noisy data" (variables whose values may not always be corrected) or if the attributes are not always sufficient to classify the data correctly, the tree should only include branches that are justified adequately.

Decision tree works by recursively picking the attribute that provides the best classification of the remaining subset. That is, the program uses the whole data set to find the attribute that best classifies the data. Then for each subset defined by the values of the selected attribute, the process is repeated with the remaining attributes. Then the tree is pruned to eliminate branches that are not justified adequately. Thus, the two basic functions of a decision-tree program are 1) selecting the best attribute to divide a set at each branch, and 2) deciding whether each branch is justified adequately. There are also multiple strategies for pruning the tree once it is generated. Normally a branch is pruned when the error introduced is within one standard error of the existing errors adjusted for the continuity correction [14]. First chosen variable by recursive partitioning is the most important variable, then go to the second variable and so on. The number of leaf could be controlled by using complexity parameter (cp) where smaller value of cp will give more leafs.

3. Logistic Regression

Logistic regression analyses will be used to identify the association between determinants and outcome variable. The multiple logistic regressions take form as

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum\nolimits_{i=1}^{k} \beta_i x_i$$

where p_i denotes the expected probabilities of the outcome for participant i, α is intercept, x_i through to x_k refers to the determinants variables, β_i through β_k as regression coefficients where k is number of predictors.

The difference of each factor variable will be compared to the overall percent of TB deaths by computing 95% confidence intervals. The adjusted proportion and the confidence intervals computed using weighted sum contrasts [15]. ROC curve will be created and used as a fundamental tool for diagnostic test evaluation.

RESULTS AND DISCUSSION

1. Decision tree

195 of 9495 records which consisted of 21 major causes of groups were TB case with 2.1%. Even the number of outcome pretty small, the data were treated as original data. Decision tree had many parameters and one of important parameters was complexity parameter (cp). Small value for cp was chose to achieve bigger number of leaves. 48 leaves were appeared with cp 0.0007 where the c-val relative error got higher and 13 predictors were presented.

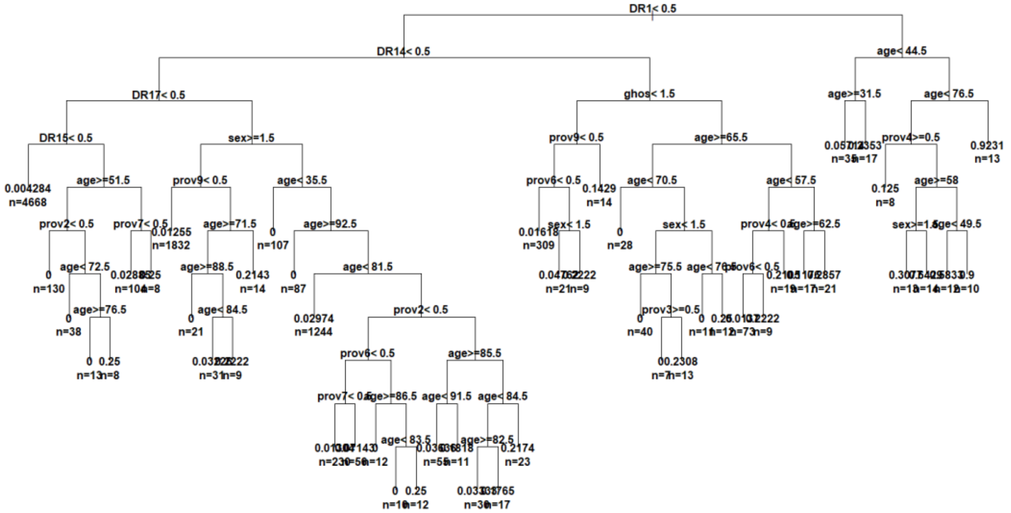

Figure 1. Tree with cp 0.0007

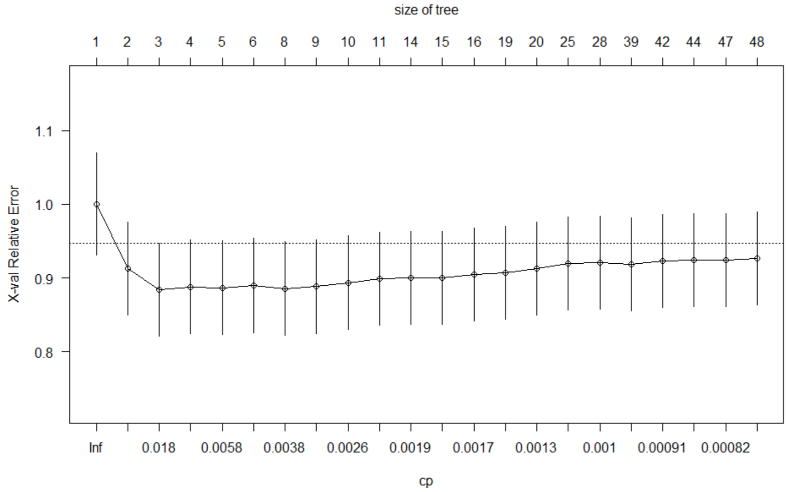

Figure 2 CP 0.0007 with relative error

To reduce relative error, cp must be changed to be bigger value where this process was called pruning the tree. However, the choosen value should confirm with model goodness-of-fit because by eliminating leaves, the number of predictors would be eliminated as well, then the cp must be chose properly. Cp with 0.0012 presented 20 leaves with 9 chosen predictors.

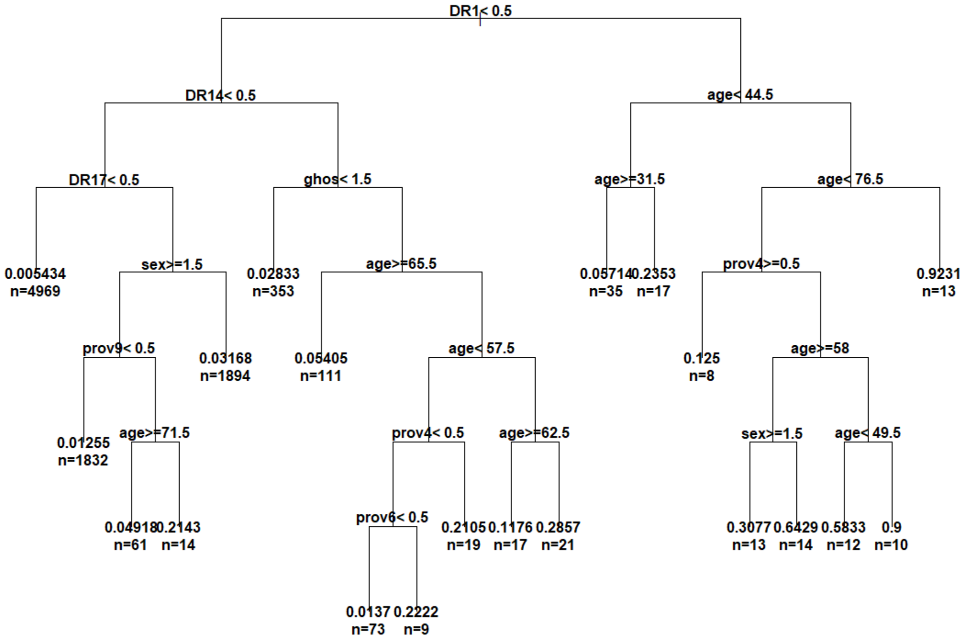

Figure 3. Tree with 0.0012

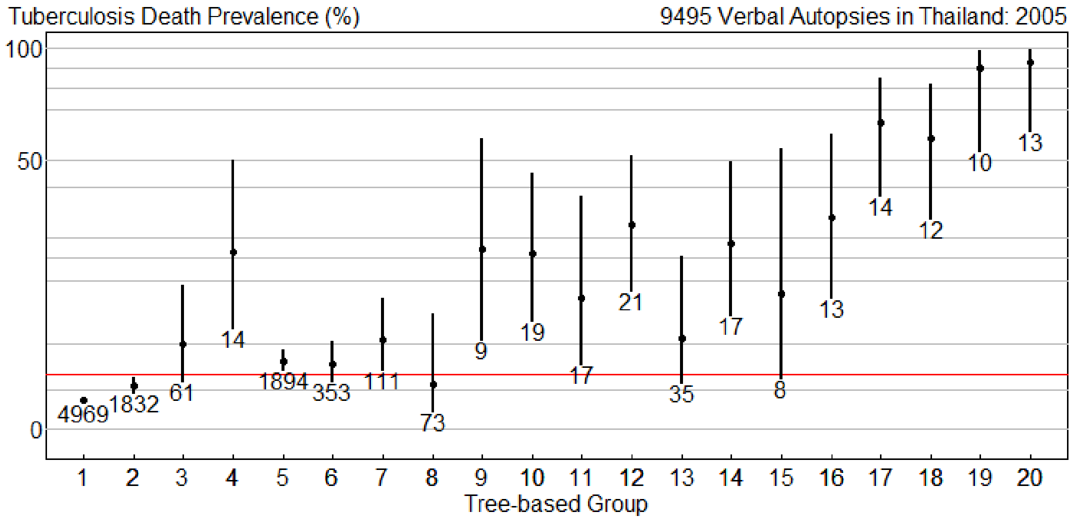

Figure 4 Confident interval of tree with cp 0.0012

Those 20 leaves represented groups classified by decision tree where each group depicted in confident interval plot in Figure 4. Some groups had large standard error that might be caused those group did not have big enough sample size. Groups with small sample size and narrow space of prevalence were merged to reduce the error and increase the number of sample size. Regrouping was showed in Figure 5.

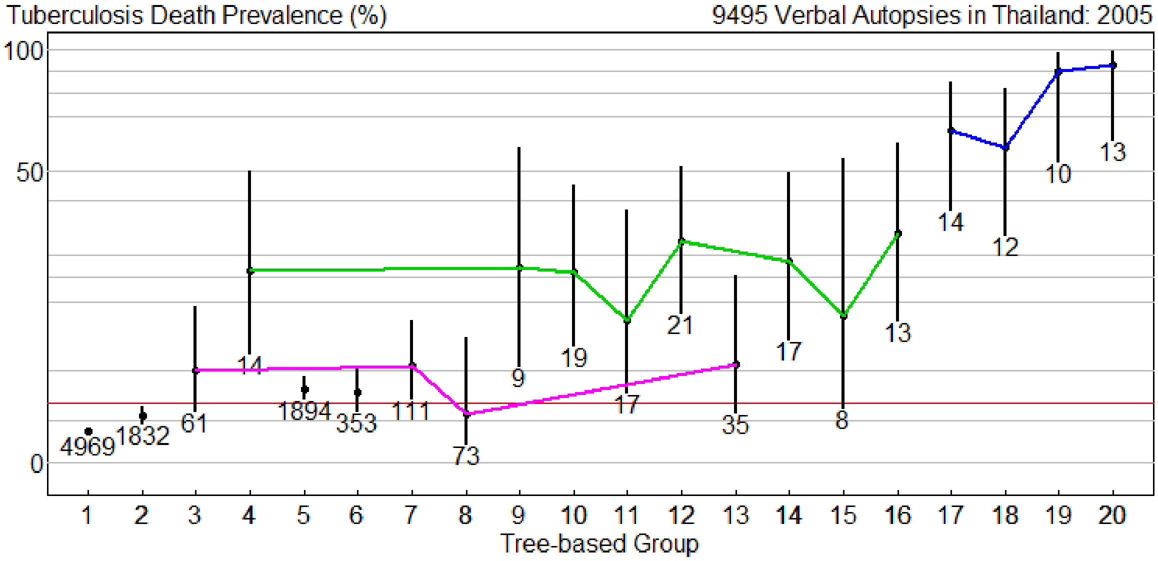

Figure 5 Regouping by mustering

Group 3, 7, 8, 13, group 4, 9, 10, 11, 12, 14, 15, 16 and group 17, 18, 19, 20 were united respectively. Figure 6 depicted confident interval plot after regrouping that 7 groups were remained and 6 groups were significantly difference and those groups had acceptable standard error

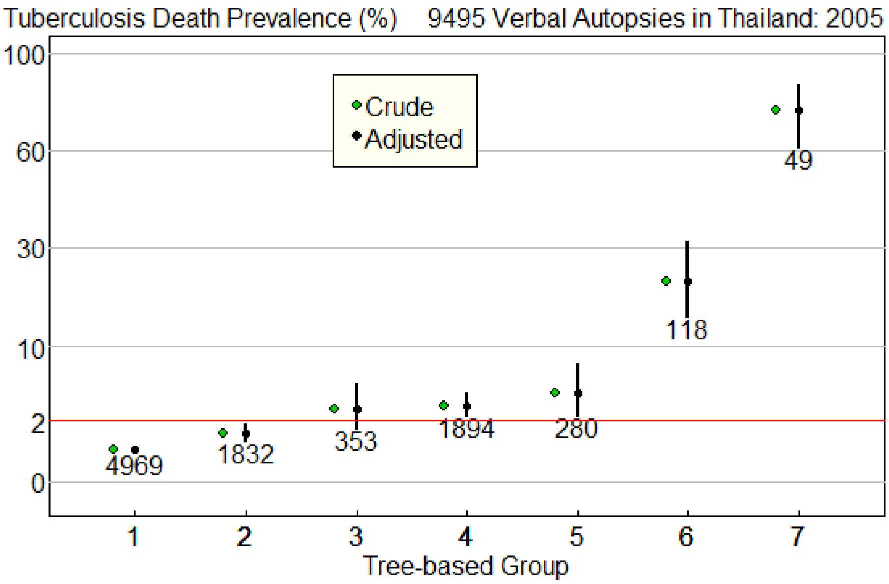

Figure 6 CI plot from tree based group

2. Logistic Regression

Province, DR-reported, location of death, and age-sex group were categorical variables and treated as determinant where death due to TB or others was treated as binary outcome. Figure 7 showed that in province variable only Songkhla province was significant, in age-sex group, male age 5-39 and 50-59 were significant, and in location of death and DR-reported, almost all observation who die outside hospital were significant, then DR-reported as Tuberculosis and other definition group who die inside or outside also significantly difference.

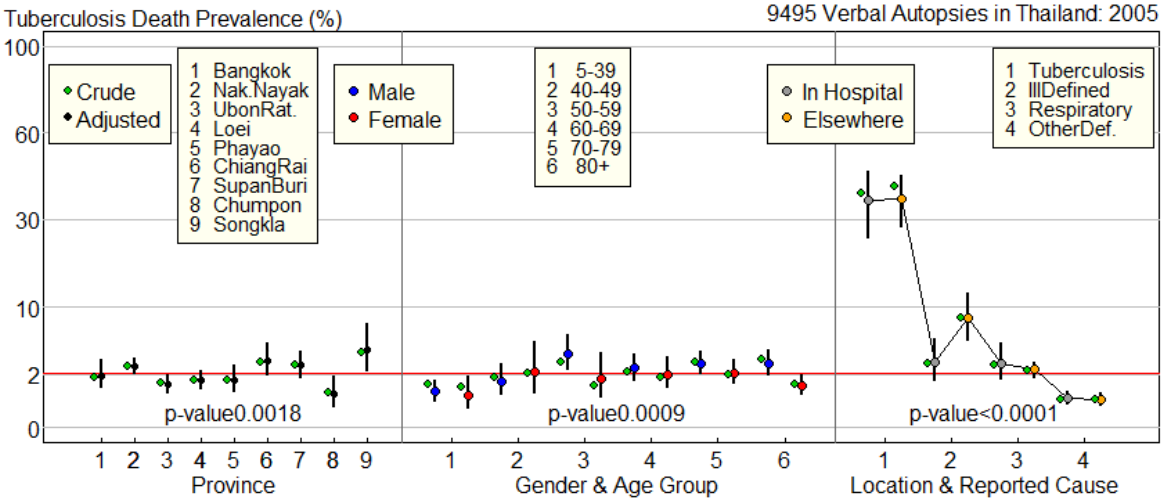

Figure 7 CI plot of logistic regression

3. Comparison

Data were split into training set and testing set 70:30 respectively where training set was applied to build both models and testing set was employed to examine for each model. Classification accuracies were not different for each model where 98.08% for logistic and 98.15% for rpart (recursive partitioning) or decision tree, respectively. Area under curve of ROC also identified both model had similar performance. This might be caused of small number of observation.

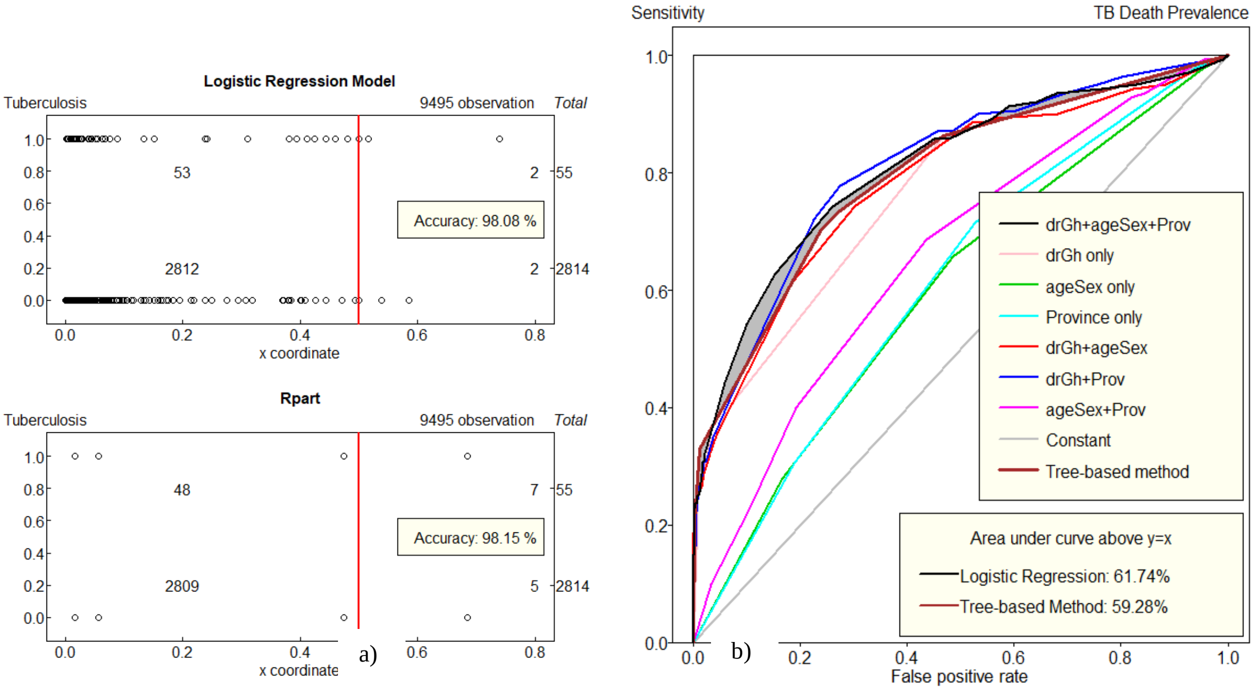

Figure 8a) classification accuracy of prediction; b) ROC curve of models

CONCLUSION

This study aimed to compare the prediction performance between statistical i.e. logistic regression and machine learning method i.e. decision tree for TB deaths from verbal autopsy (VA) data in Thailand. Decision tree was more able to classified the significant variables rather than logistic regression. However, result of prediction accuracy and AUC of ROC curve stated that both models were not different. This might be caused of the number of observation was too small (only 2.1%) where in calculating classification accuracy this small number was divided into training set and testing set that decreased the number of TB cases.

SUGGESTION

For further study, apparently the data that would be analyzed is big enough otherwise the data should be increased with proper method. The further study could attempt some additional model in machine learning such as random forest and neural network that would give more evidence in prediction performance.

ACKNOWLEDGMENT

I am grateful to Dr. Kanitta Bundhamcharoen as officer in Ministry of Public Health for allowing permission to obtain the data to do a study. Without funding support this work was not success, so also I would like to give my full respect to Thailand Educational Hub and graduate School Prince of Songkhla University for providing THE-AC scholarship and research grant.

REFERENCES

1. Mohajan, H. K. Tuberculosis is a Fatal Disease among Some Developing Countries of the World. *Am. J. Infect. Dis. Microbiol.* **3,** 18–31 (2015).
2. Namwat, C. Tuberculosis in Thailand TB : 2011 Global Estimates. *Power point* (2012).
3. Setel, P. W. *et al.* Sample registration of vital events with verbal autopsy: A renewed commitment to measuring and monitoring vital statistics. *Bull. World Health Organ.* **83,** 611–617 (2005).
4. Glasser, J. H. The quality and utility of death certificate data. *Am. J. Public Health* **71,** 231–233 (1981).
5. Timaeus, I. M. & Jasseh, M. Adult mortality in sub-Saharan Africa: evidence from Demographic and Health Surveys. *Demography* **41,** 757–772 (2004).
6. De Henauw, S., de Smet, P., Aelvoet, W., Kornitzer, M. & De Backer, G. Misclassification of coronary heart disease in mortality statistics. Evidence from the WHO-MONICA Ghent-Charleroi Study in Belgium. *J. Epidemiol. Community Health* **52,** 513–519 (1998).
7. Pattaraarchachai, J. *et al.* Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. *Popul. Health Metr.* **8,** 1–12 (2010).
8. Polprasert, W. *et al.* Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. *Popul. Health Metr.* **8,** 1–15 (2010).
9. Klinjun, N., Lim, A. & Bundhamcharoen, K. A logistic regression model for estimating transport accident deaths using verbal autopsy data. *Asia-Pacific J. Public Heal.* **27,** (2015).
10. Reed, P. & Wu, Y. Logistic regression for risk factor modelling in stuttering research. *J. Fluency Disord.* **38,** 88–101 (2013).
11. Eftekhar, B., Mohammad, K., Ardebili, H. E., Ghodsi, M. & Ketabchi, E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BioMed Cent. Med. Informatics Decis. Mak.* **5,** 1–8 (2005).
12. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55,** 78 (2012).
13. Rao, C. *et al.* Verifying causes of death in Thailand: rationale and methods for empirical investigation. *Popul. Health Metr.* **8,** 1–13 (2010).
14. Long, W. J., Griffith, J. L., Selker, H. P. & D'Agostino, R. B. A comparison of logistic regression to decision-tree induction in a medical domain. *Comput. Biomed. Res.* **26,** 74–97 (1993).
15. Tongkumchum, P. & Mcneil, D. Confidence intervals using contrasts for regression model. *Songklanakarin J. Sci. Technol.* **31,** 151–156 (2009).