Loglinear and Classification Tree Models of the Decision Paradigm of the Tuberculosis Diagnostic Committee

Caryl Rose E. Alfonte
University of the Philippines Diliman
cealfonte@yahoo.com

The TB Diagnostic Committees (TBDC) evaluate cases of pulmonary tuberculosis (PTB) symptomatics who are smear-negative, but whose chest x-rays show lesions suggestive of tuberculosis that may warrant anti-TB treatment. In a review of the 600 TBDC referral forms of new patients who consulted in Manila district health centers from 2006 to 2008, the demographic and clinical characteristics associated with a positive chest x-ray and eventually leading to a diagnosis of new active PTB are identified using loglinear models and classification trees.

Hemoptysis is the most important variable in differentiating a TBDC decision of active PTB from non-PTB. The final loglinear model signifies that history of alcoholic beverage drinking, hemoptysis, weight loss, and age less than 40 are individually associated with a TBDC diagnosis of active PTB. Sex is used as the root node variable in the classification tree in order to improve classification accuracy and explore sex-related factor differences. In females, hemoptysis, alcoholic beverage drinking history, and age are individually associated with the target variable; while interactions among predictors of a diagnosis of active PTB are more demonstrable in males. The TBDC may find the proposed models useful as a decision tool in making a diagnosis of active pulmonary tuberculosis in support of the chest radiograph.

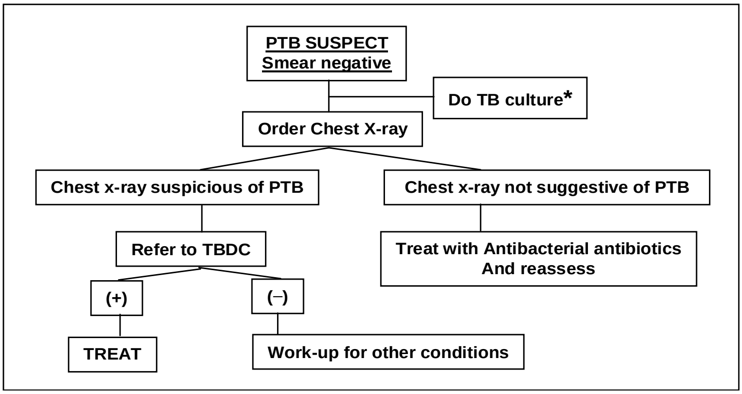
Keywords: *Classification tree, Loglinear model, Positive chest x-ray, Pulmonary tuberculosis, Sputum smear-negative, TB Diagnostic Committee*

1. Introduction

The Tuberculosis Diagnostic Committee (TBDC) evaluates cases of patients manifesting symptoms suggestive of pulmonary tuberculosis (PTB) but with negative sputum smears and suspicious chest radiograph results. The TBDC is composed of tuberculosis experts from the public and private sectors who represent various disciplines – a National Tuberculosis Program (NTP) coordinator, a radiologist, a clinician/internist often represented by a pulmonologist or an infectious disease specialist, and a NTP nurse who acts as the committee's secretary. The patients are referred by the health center physicians to the TBDC by sending all available chest x-ray films with the properly accomplished TBDC referral form. Upon review, the TBDC will deliberate and come up with a consensus based on the national guidelines (Figure 1) where patients requiring TB treatment shall avail of the directly observed treatment short-course or DOTS services provided by the health center (National Tuberculosis Control Program Manual of Procedures 2005).

In a setting where resources are crucial factors in health care decisions, the role of the TBDC is critical as it is the recommending authority in the option of which patient gets to be treated or not treated with antituberculosis drugs based on the patient's reported medical history, clinical findings and radiographic results. A quantitative characterization of the TBDC decision in terms of statistical models may give an objective description that would consequently facilitate in minimizing subjectivity in diagnosing a case or noncase of active pulmonary tuberculosis.

Figure 1. Diagnosis algorithm: Work-up of smear-negative PTB suspects. (Clinical practice guidelines for the diagnosis, treatment, prevention and control of tuberculosis in adult Filipinos: 2006 Update)



**TB culture Indications: Sputum TB culture with drug susceptibility testing is primarily recommended for those who are at risk for drug resistance and in smear-positive patients in all cases of – re-treatment, treatment failure, and in patients suspected to have one or multi-drug resistant TB (MDR-TB).*

The primary objective of the paper is to identify the demographic characteristics, medical history and physical examination findings that are associated with a positive chest x-ray and leading to a diagnosis of new active tuberculosis among sputum smear-negative patients referred to the TBDC.

2. Methodology

2.1 Study design, setting and subjects

This is a cross-sectional study of the TBDC referral forms of 600 patients who consulted in all 15 health centers of the first and second districts of Manila from 2006 to 2008. The inclusion criterion of the study is a patient who had been given the impression of being a new active TB case by the referring health center physician. A new patient is defined as one who had never had treatment for TB or, if with previous anti-TB medications, these were taken for less than four weeks. The reasons for exclusion of a subject upon review of his referral form are: a patient who had received any anti-TB drug for more than four weeks at any time and no specific TBDC recommendation (of prescribing antituberculosis treatment or no treatment) as in cases where the chest x-rays needed to be repeated due to poor quality films.

2.2 Study variables

In this study, the outcome variable is the diagnosis of new active PTB or "not PTB" by the TBDC among patients with negative sputum smears. The TBDC decision principally relates to their chest x-ray evaluation where the impression is negative or positive for PTB. A case of "no disease" in this study is represented by a negative result of the chest radiography leading to a TBDC diagnosis of inactive TB or other non-PTB lung disease where the recommendations for the patient are no anti-TB therapy, symptomatic treatment as needed and/or putting the patient under surveillance.

The independent variables taken into account as to their possible association with the TBDC diagnosis of a sputum smear-negative pulmonary tuberculosis are the demographic characteristics of the subjects such as age, sex, civil status, and job category. The variable dichotomies are male or female for sex, and single or not single (married/live-in and widow/separated) for civil status. Preliminary classification tree analysis was done to determine the dichotomous groupings of age (40 as cutoff) and job category (driver, student, or skilled worker against the other group of office/sales employee, service worker, professional, self-employed, Overseas Filipino worker (OFW), laborer/unskilled, or no occupation). The above groupings are used to construct loglinear models.

The medical history of the subjects is comprised of cough, sputum production, hemoptysis or coughing up blood, fever, weight loss, tiredness, chest/back pain, dyspnea or difficulty of breathing, exposure to a known PTB patient, any history of smoking, any history of alcoholic beverage drinking, co-existing disease, previous diagnosis or history of PTB, and intake of any anti-TB drugs. For the variable co-existing disease, the most recorded medical conditions of the subjects on the TBDC referral forms are bronchial asthma, diabetes mellitus, hypertension, heart disease, kidney disease, thyroid disease, and malignancy.

The pertinent physical examination findings based on patient records consist of emaciation (appearance of extreme wasting or thinness), pallor or paleness, and lung auscultation findings – namely, crackles or rales, decreased breath sounds, abnormal vocal fremitus, harsh breath sounds, rhonchi, and wheezes. The medical history and clinical features are categorized into the presence or absence of the condition. There are 26 independent variables under consideration in the study. To address the research question, loglinear models and classification trees are utilized to determine which of the 26 independent variables are associated, individually or in combination, with the TBDC diagnosis of sputum smear-negative tuberculosis.

Patient records were obtained upon approval of the Manila Health Department, the Division of Tuberculosis Control, the district offices and all the health centers of Manila Health Districts I and II. Prevention of redundant subject entries required the listing of patient identifiers during data collection. There is, however, no disclosure of identifiable patient information in any part of this research. Strict confidentiality in terms of data handling and record keeping is ascertained.

2.3 Univariate analysis

Univariate comparisons between smear-negative PTB and non-PTB patients as diagnosed by the TBDC are carried out using Student's t-test for the variable age, and the chi-square test for two independent samples for the rest of the predictors. Odds ratios (ORs) with 95% confidence intervals are calculated for each variable. Independent variables with chi-square p-values at or near 0.05 are prioritized for model building.

2.4 Loglinear model

The six variables which turned out significant in the univariate analyses are used to build loglinear models with TBDC diagnosis (T) as the variable of interest – these are any history of alcoholic beverage drinking (A), hemoptysis (H), job category (J), weight loss (W), any history of smoking (S), and age 40 (Y). Model building is done only on the seven variables since there was already overfitting (when the degrees of freedom already reaches zero and thus, no calculated likelihood ratio) upon fitting all four- and higher-factor terms. Likewise, the study objective to determine variables that would associate most with the TBDC diagnosis will be further underscored. Model selection by backward elimination is used by removing one term sequentially to determine which term would cause the least remarkable improvement in the model fit.

For a multidimensional contingency table with five variables, the expected number or frequency of cases in cell (i, j, k, l, m) is a function of all the variables in the model

$$\log \mu_{ijklm} = \lambda + \lambda_i^A + \lambda_j^H + \lambda_k^T + \lambda_l^W + \lambda_m^Y + \lambda_{ik}^{AT} + \lambda_{jk}^{HT} + \lambda_{kl}^{TW} + \lambda_{km}^{TY} + \lambda_{lm}^{WY}$$

where the log μ_{ijklm} is an additive function of the single-factor terms $(\lambda + \lambda_i^A + \lambda_j^H + \lambda_k^T + \lambda_l^W + \lambda_m^Y)$ denoting the main effects and the two-factor interactions $(\lambda_{ik}^{AT} + \lambda_{jk}^{HT} + \lambda_{kl}^{TW} + \lambda_{km}^{TY} + \lambda_{lm}^{WY})$ representing partial associations. The superscripts are simply the labels for the variables namely: alcoholic beverage drinking (A), hemoptysis (H), TBDC diagnosis (T), weight loss (W), and age 40 (Y).

2.5 Classification tree

Classification trees are very flexible, conceptually simple, and can easily highlight interactions between the variables. Trees can also handle mixed variable types with consummate ease, and can cope with measurement vectors of different dimensionality from object to object, a property not possessed by all other classification methods (Breiman et al. 1984 and Hand 1997). Binary tree structured classifiers are created by repeated splits of subsets of the root node into two descendant subsets or nodes. Creating a classification tree is a series of determining the splits, deciding on a node to be terminal or proceed to splitting it, and assigning a class label to the terminal node. Mello et al. (2006), in a study in Rio de Janeiro, developed a prediction model for smear-negative pulmonary tuberculosis using logistic regression and classification and regression tree (CART) models where the patients' symptoms, physical signs and chest x-rays were the predictor variables.

2.6 Validation of loglinear and classification tree models

For both loglinear and classification tree analyses, model building is done with the training data (random 70% or 600 of the total 851 subjects) and then the model is validated using the test data (remaining 251 or 30% of the data). Two variables—job category and civil status—are with missing values. The training set has 12.67% (76 of 600) and 5.67% (34 of 600) missing values for job category and civil status, respectively. The test set has closely similar percentages at 13.55% (34 of 251) for job category and 5.58% (14 of 251) for civil status.

The univariate and loglinear analyses are conducted using the software SAS version 9.1. (Cary, NC: SAS Institute Inc). CART 6.2.0.160 PRO (San Diego, CA: Salford Systems) is used in the classification tree modeling.

3. Results and Discussion

3.1 Univariate analysis

There are 362 (60.33%) males and 238 (39.67%) females among the 600 study subjects. The mean age is 41.56 (± 17.87) where the age group 20–39 years old has the highest percentage of subjects at 40.83%, followed by the 40–59 age group at 30.67%. The rest of the subjects are 60 years old and above (19.17%) and below 20 years of age (9.33%).

Out of the 600 subjects, the TBDC gave a diagnosis of active PTB and recommended anti-TB treatment to 369 patients (61.50%); while 231 (38.50%) were negative for PTB. There is no significant difference (t-test p-value 0.1813) in the mean age of subjects diagnosed with PTB (age 40.79 ± 17.78) and those negative for PTB (age 42.80 ± 17.98). Utilizing chi-square test and odds ratios in Table 1, the resulting significant correlates at or near 0.05 level of a TBDC diagnosis of PTB are age below 40 (p-value = 0.0678), job categories of driver, student, or skilled worker (p-value = 0.0305), hemoptysis (p-value < 0.0001), weight loss (p-value = 0.0046), history of alcoholic beverage drinking (p-value = 0.0230), and history of smoking (p-value = 0.0243).

3.2 Loglinear modeling

Model fitting with the seven variables—TBDC diagnosis (T), any history of alcoholic beverage drinking (A), hemoptysis (H), job category (J), weight loss (W), any history of smoking (S), and age 40 (Y)—makes use of 524 observations from the training set since the variable job category (J) has missing values for 76 subjects. To maximize all 600 observations, modeling is likewise performed on the six variables (AHSTWY) and the five-factor term (AHTWY). Table 2 puts on view the goodness-of-fit tests and significant association terms for selected models under model groups (AHJSTWY), (AHSTWY), and (AHTWY).

Goodness-of-fit tests and residual analyses restricted the model selection to Model 8 ({AT}{HT}{TW}{TY}{WY}), Model 9 ({HT}{TW}{WY}{ATY}) and Model 10 ({HT}{TW}{WY}{AHY}{ATY}) under (AHTWY); mainly due to inclusion of significant terms relevant to the research objectives while exhibiting parsimony, and subsequently, good fit and least cell residuals. The chosen final loglinear model is Model {AT}{HT}{TW}{TY}{WY}, as shown in Table 2. Its simplicity and interpretability, as well as its consistently favorable performance for

Table 1. Demographic profile, medical history and clinical characteristics of the 600 study subjects

Predictor Variable	0 = No PTB / Inactive PTB (n=231)	1 = Active PTB (n=369)	Odds Ratio (95% CI)	Chi-square p-value
Demographic profile:				
Age*				
0 = Below 40 years old	105 (45.45)	196 (53.12)	0.74	0.0678*
1 = 40 years old and above	126 (54.55)	173 (46.88)	(0.53 – 1.02)	
Sex				
0 = Female	97 (41.99)	141 (38.21)	1.17	0.3571
1 = Male	134 (58.01)	228 (61.79)	(0.84 – 1.64)	
Civil status	(n=218)	(n=348)		
0 = Single	69 (31.65)	116 (33.33)	0.93	0.6780
1 = Married/Live-in, Widow/Separated	149 (68.35)	232 (66.67)	(0.65 – 1.33)	
Job Category*	(n=215)	(n=309)		
0 = No job, Office/ Sales/ Service worker, Professional, Self-employed, OFW, and Laborer/Unskilled	176 (81.86)	228 (73.79)	1.60 (1.04 – 2.46)	0.0305*
1 = Driver, Student, and Skilled worker	39 (18.14)	81 (26.21)		
Medical history:				
Cough				
0 = Absent	38 (16.45)	51 (13.82)	1.23	0.3780
1 = Present	193 (83.55)	318 (86.18)	(0.78 – 1.94)	
Sputum production				
0 = Absent	81 (35.06)	118 (31.98)	1.15	0.4346
1 = Present	150 (64.94)	251 (68.02)	(0.81 – 1.63)	
Hemoptysis*				
0 = Absent	207 (89.61)	274 (74.25)	2.99	<.0001*
1 = Present	24 (10.39)	95 (25.75)	(1.85 – 4.85)	
Fever				
0 = Absent	154 (66.67)	230 (62.33)	1.21	0.2816
1 = Present	77 (33.33)	139 (37.67)	(0.86 – 1.71)	
Weight loss*				
0 = Absent	118 (51.08)	145 (39.30)	1.61	0.0046*
1 = Present	113 (48.92)	224 (60.70)	(1.16 – 2.25)	
Tiredness				
0 = Absent	106 (45.89)	158 (42.82)	1.13	0.4612
1 = Present	125 (54.11)	211 (57.18)	(0.81 – 1.58)	
Chest pain/Back pain				
0 = Absent	76 (32.90)	121 (32.79)	1.00	0.9779
1 = Present	155 (67.10)	248 (67.21)	(0.71 – 1.43)	
Dyspnea				
0 = Absent	137 (59.31)	215 (58.27)	1.04	0.8009
1 = Present	94 (40.69)	154 (41.73)	(0.75 – 1.47)	

Predictor Variable	0 = No PTB / Inactive PTB (n=231)	1 = Active PTB (n=369)	Odds Ratio (95% CI)	Chi-square p-value
TB Exposure				
0 = Absent	174 (75.32)	284 (76.96)	0.91	0.6456
1 = Present	57 (24.68)	85 (23.04)	(0.62 – 1.34)	
Any History of Smoking*				
0 = Absent	157 (67.97)	217 (58.81)	1.49	0.0243*
1 = Present	74 (32.03)	152 (41.19)	(1.05 – 2.10)	
Any History of Alcoholic Beverage Drinking*				
0 = Absent	205 (88.74)	302 (81.84)	1.75	0.0230*
1 = Present	26 (11.26)	67 (18.16)	(1.08 – 2.85)	
Co-existing Disease				
0 = Absent	212 (91.77)	331 (89.70)	1.28	0.3994
1 = Present	19 (8.23)	38 (10.30)	(0.71 – 2.28)	
History of PTB				
0 = Absent	196 (84.85)	312 (84.55)	1.02	0.9221
1 = Present	35 (15.15)	57 (15.45)	(0.65 – 1.62)	
Intake of any anti-TB drugs				
0 = Absent	199 (86.15)	313 (84.82)	1.11	0.6557
1 = Present	32 (13.85)	56 (15.18)	(0.70 – 1.78)	
Physical examination (P.E.):				
Specific P.E. Findings:				
Emaciation				
0 = Absent	223 (96.54)	363 (98.37)	0.46	0.1469
1 = Present	8 (3.46)	6 (1.63)	(0.16 – 1.35)	
Pallor				
0 = Absent	219 (94.81)	357 (96.75)	0.61	0.2373
1 = Present	12 (5.19)	12 (3.25)	(0.27 – 1.39)	
Crackles or Rales				
0 = Absent	204 (88.31)	319 (86.45)	1.1843	0.5070
1 = Present	27 (11.69)	50 (13.55)	(0.72 – 1.95)	
Decreased Breath Sounds				
0 = Absent	203 (87.88)	311 (84.28)	1.35	0.2211
1 = Present	28 (12.12)	58 (15.72)	(0.83 – 2.20)	
Vocal Fremitus				
0 = Absent	225 (97.40)	366 (99.19)	0.3074	0.0802
1 = Present	6 (2.60)	3 (0.81)	(0.08 – 1.24)	
Harsh Breath Sounds				
0 = Absent	201 (87.01)	317 (85.91)	1.0991	0.7014
1 = Present	30 (12.99)	52 (14.09)	(0.68 – 1.78)	
Rhonchi				
0 = Absent	212 (91.77)	349 (94.58)	0.6394	0.1750
1 = Present	19 (8.23)	20 (5.42)	(0.33 – 1.23)	
Wheeze				
0 = Absent	221 (95.67)	356 (96.48)	0.8070	0.6168
1 = Present	10 (4.33)	13 (3.52)	(0.35 – 1.87)	

Table 2. Goodness-of-fit tests and significant parameters for selected loglinear models with variables TBDC decision[a] (T), any history of alcoholic beverage drinking[b] (A), hemoptysis[c] (H), job category[d] (J), any history of smoking[e] (S), weight loss[f] (W), and age[g] (Y).

A. Loglinear model group (AHJSTWY) n=524	*df*	Likelihood Ratio (G^2)	p-value	Chi-square	Significant Association at 0.05
1) {AS}{AT}{HT}{JT}{JY}{SW}{TW}{TY}	71	78.24	0.2598	83.18	AS, HT, JY, SW, TW, TY at 0.0564
2) {AS}{HT}{JY}{SW}{TW}{TY}	73	81.41	0.2339	85.90	AS, HT, JY, SW, TW, TY
B. Loglinear model group (AHSTWY) n=600	*df*	Likelihood Ratio (G^2)	p-value	Chi-square	Significant Association at 0.05
3) {AS}{AT}{HT}{SW}{SY}{TW}{TY}{WY}	39	39.84	0.4324	42.06	AS, HT, SW, SY, TW, TY, WY
4) {AS}{HT}{SW}{SY}{TW}{TY}{WY}	40	41.68	0.3974	44.53	AS, HT, SW, SY, TW, TY, WY
5) {AS}{HT}{SW}{TW}{TY}{WY}	41	49.15	0.1791	54.12	AS, HT, SW, TW, WY, TY at 0.0594
C. Loglinear model group (AHTWY) n=600	*df*	Likelihood Ratio (G^2)	p-value	Chi-square	Significant Association at 0.05
6) {AH}{AT}{HT}{TW}{TY}{WY}	18	22.25	0.2210	21.63	AH, HT, TW, TY, WY
7) {AH}{HT}{TW}{TY}{WY}	19	24.48	0.1785	23.79	AH, HT, TW, TY, WY
†8) {AT}{HT}{TW}{TY}{WY}	19	27.45	0.0947	28.76	AT, HT, TW, TY, WY
9) {HT}{TW}{WY}{ATY}	17	24.09	0.1170	25.36	AT, HT, TW, TY, WY
10) {HT}{TW}{WY}{AHY}{ATY}	14	15.17	0.3666	14.76	AY, HT, TW, TY, WY, AHY, ATY, AT at 0.0563

† Chosen final loglinear model
[a] where 0=Non-PTB and 1=Active PTB
[b] where 0=Absent and 1=Present
[c] where 0=Absent and 1=Present
[d] where 0=No job, Office/Sales employee, Service worker, Professional, Self-employed, Overseas worker, Laborer/Unskilled and 1=Driver, Student, Skilled worker;
[e] where 0=Absent and 1=Present
[f] where 0=Absent and 1=Present
[g] where 0=Below 40 years old and 1=40 years old & above

Table 3. Estimated odds ratios for Loglinear model {AT}{HT}{TW}{TY}{WY} with variables TBDC decision[a] (T), any history of alcoholic beverage drinking[b] (A), hemoptysis[c] (H), weight loss[d] (W), and age[e] (Y)

Association Term	Odds Ratio (95% CI)
AT	1.662 (1.021 - 2.704)
HT	2.829 (1.745 - 4.587)
TW	1.743 (1.24 - 2.451)
TY	0.672 (0.478 - 0.944)
WY	2.155 (1.544 - 3.007)

[a] *where 0=Non-PTB and 1=Active PTB*
[b, c, d] *where 0=Absent and 1=Present*
[e] *where 0=Below 40 years old and 1=40 years old & above*

both the design and test sets, render the model the best to represent and characterize our data.

The calculated fitted odds ratios for the final model {AT}{HT}{TW}{TY}{WY} are in Table 3. In the interpretation of the estimated odds ratios, among sputum smear-negative patients referred to the TBDC –

For AT: A diagnosis of active pulmonary tuberculosis is 1.66 (95% CI: 1.02, 2.70) times as likely for those with history of alcoholic beverage drinking as for those without alcoholic beverage drinking history.

For HT: A diagnosis of active PTB is 2.83 (95% CI: 1.74, 4.59) or almost 3 times as likely for those with symptom of hemoptysis as for those without hemoptysis.

The TW odds ratio of 1.74 (95% CI: 1.24, 2.45) means the odds of a diagnosis of active PTB are 1.74 times greater for those with weight loss than for those without symptom of weight loss.

For TY: With an odds ratio of 0.67 (95% CI: 0.48, 0.94), the interpretation of the reciprocal odds ratio is that the odds of a diagnosis of active PTB are 1.49 (95% CI: 1.0597, 2.09154) or about one-and-a-half times greater for those below 40 years of age than for those 40 years old and above.

For WY: Those who are 40 years old and above are 2.16 (95% CI: 1.54, 3.00) times as likely to have symptom of weight loss as for those below 40 years of age.

The interpretation of Model {AT}{HT}{TW}{TY}{WY} is the variables history of alcoholic beverage drinking (A), hemoptysis (H), weight loss (W), and age (Y) are individually associated with a TBDC diagnosis of active PTB (T); and there is an interaction between weight loss (W) and age (Y).

3.3 Classification tree

In the analysis using classification trees, age is entered as a continuous variable, while civil status and job category are multi-categorized into their natural groupings. Resulting from the initial tree fitting using all 26 predictors, six variables that are the leading root node splitters—hemoptysis (H), weight loss (W), history of alcoholic beverage drinking (A), job category (J), history of smoking (S), and age (Y)—are maintained to grow another set of trees. All possible combinations of the six predictors are forced into the tree via the root node and its left and the right child nodes to search for appropriate trees.

Sex is subsequently added to the six variables in order to improve the classification accuracy of the model tree and explore differences in the sex-related factors, even if it is not individually associated with the TBDC decision. In profiling subjects by sex for both training and test data, the chi-square test shows females have significantly higher percentages of nonsingles (p-value = 0.0353), with medical history of TB exposure (p-value = 0.05) and co-existing disease (p-value = 0.0429). A significant majority of males exhibit the characteristics: job=1 or driver, student, or skilled worker (p-value < 0.0001); and with histories of smoking (p-value < 0.0001) and alcoholic beverage drinking (p-value < 0.0001). Thus, differences in the splitter characteristics between male and female subjects may well be highlighted when sex is forced as the root node variable.

In a summarized evaluation of the four best performing trees in Table 4 from about 150 generated classification trees, results of the test data are likewise listed since they are considered as the fair estimates of the tree performance since the training sample is the one actually used in building the model. Majority of the most favorable criteria in both training and test data is observed for Tree 47. These are comparatively higher (or highest) values of prediction accuracy (64.14 – 64.47), ROC score (0.66 – 0.69), sensitivity (0.62 – 0.63), and specificity (0.65 – 0.69); together with lower misclassification rates (30.74 – 38.21) and relative cost (0.71).

Based on the aforementioned reasons, Tree 47 (with root node=sex, left node=weight loss, right node=alcoholic beverage drinking history) is considered as the most accurate classification tree for the study. The illustration of our selected classification tree is in Figure 2. Even if hemoptysis is suppressed from being in one of the top nodes, it persists to show strong relations with the target variable by appearing in both sides of the tree or for both sexes. Nodes ending with presence of hemoptysis (H=1) capture high percentages of the active PTB cases, namely, 82.7% in terminal node 1 and 80% in terminal node 11.

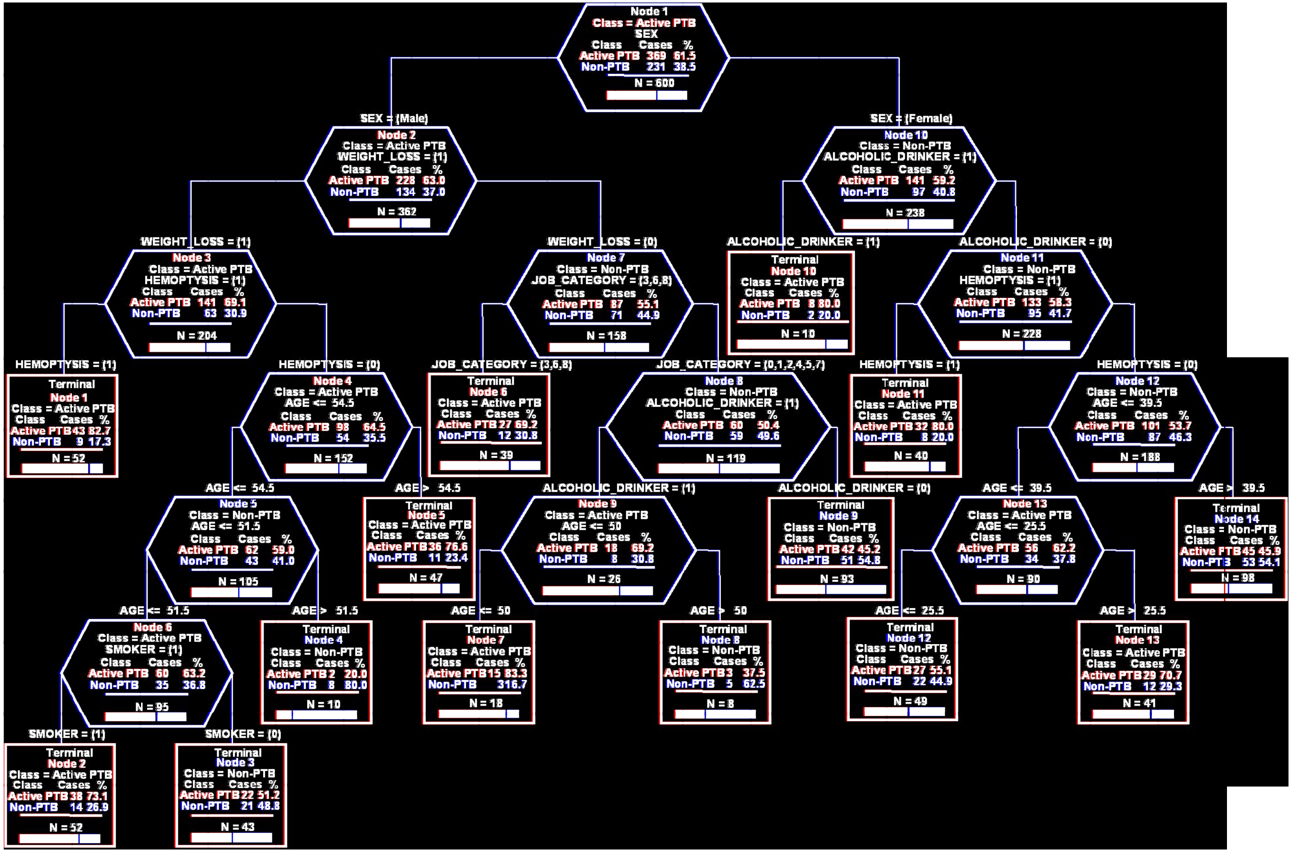
Table 4. **Evaluation criteria of candidate classification trees with TBDC decision (T) as the target variable and predictors sex, history of alcoholic beverage drinking (A), hemoptysis (H), job category (J), history of smoking (S), weight loss (W), and age (Y).**

MODEL TREE	ROOT NODE	LEFT CHILD	RIGHT CHILD	Training Data (n=600)						Test Data (n=251)								
				Prediction Accuracy	Misclas- sification Rate for T=1	Misclas- sification Rate for T=0	ROC	Sensi- tivity	Speci- ficity	Prediction Accuracy	Misclas- sification Rate for T=1	Misclas- sification Rate for T=0	ROC	Sensi- tivity	Speci- ficity Nodes	Number of	Relative Cost	Variables in the Tree
23	W	H	S	65.67[b]	32.79	36.80	0.70	0.67	0.63	59.76[w]	37.41	44.23[w]	0.59[w]	0.63	0.56[w]	13	0.82[w]	AHJSWY
39	Sex	H	Y	62.67[w]	38.21[w]	35.93	0.65[w]	0.62[w]	0.64	62.15	39.46[w]	35.58	0.63	0.61[w]	0.64	9	0.75	HJSWY+Sex
†47	Sex	W	A	64.67	38.21[w]	30.74[b]	0.69	0.62[w]	0.69	64.14[b]	36.73	34.62[b]	0.66[b]	0.63	.065[b]	14	0.71[b]	AHJSWY+Sex
51	Sex	W	Y	65.33	31.71b	39.39[w]	0.76	0.68b	0.61[w]	63.35	34.69[b]	39.42	0.64	0.65[b]	0.61	12	0.74	AHJSWY+Sex
Minimum value				62.67	31.71	30.74	0.65	0.62	0.61	59.76	34.69	34.62	0.59	0.61	0.56	9	0.71	
Maximum value				68.67	38.21	39.39	0.70	0.68	.069	61.14	39.46	44.23	0.66	0.65	0.65	14	0.82	

† *Chosen final classification tree, most accurate tree based on the table values*
b *Best among column values*
w *Worst among column values*

Figure 2. Final classification tree with TBDC decision (T=1 for active PTB and T=0 for non-PTB) as the target variable

[Note on job_category: 0=No work, 1=driver, 2=service worker/vendor, 3=student, 4=office employee/sales worker, 5=professional/self-employed, 6=laborer/unskilled, 7=overseas worker, 8=skilled worker/others]



The left portion of the classification tree demonstrates that in males, a diagnosis of active PTB involves interactions among the variables history of alcoholic beverage drinking, hemoptysis, job category, history of smoking, weight loss, and age. Conversely in females, there is individual association of the TBDC diagnosis of active PTB with hemoptysis, alcoholic beverage drinking history, and age (as seen on the right portion of the tree).

In Table 4, the final classification tree has an overall prediction accuracy of 64% for both training and data sets, with misclassification rates of 30.74 to 38.21%. Sensitivity and specificity range from 62% to 69%. The ROC scores (0.69 for the training and 0.66 for the test data) of the chosen tree (Tree 47) suggest we would be able to tell which one the TBDC will diagnose as active PTB in 69% of all cases for the training data, and 66% for the test data. The relative cost of 0.714, the lowest among the four candidate trees, denotes the overall accuracy of our chosen tree which has an error rate that is 71.4% of that experienced without a model.

3.4 Validation of loglinear and classification tree models

The total records gathered for 851 subjects are randomly divided into 600 (70%) and 251 (30%), respectively, as the training and test data sets. In the validation, there is consistency in the results for both training and test data sets since the final models apply as well to the test data, which was not used in building the model. In this study, the strength of the classification tree analysis is in the handling of continuous and multi-categorical variables by determining cutoffs and partitions anywhere throughout the tree. Its ability to operate even with missing values for some variables speaks well of maximizing the data, instead of removing entries from the analysis. Interpreting the magnitude and direction of association between variables through the odds ratios is the forthright advantage of loglinear analysis.

4. Conclusions and Recommendations

4.1 Conclusions

The decision patterns of the TBDC in diagnosing an active PTB case among sputum smear-negative PTB suspects are based on positive chest radiographic findings – which significantly associate with symptoms of hemoptysis and weight loss, history of alcoholic beverage drinking, and the demographic factor age (less than 40 years old).

Hemoptysis or coughing up blood is the most important variable in differentiating a TBDC decision of active PTB from non-PTB.

Sex is used as the root node variable in order to improve the classification accuracy of the model tree and explore differences in the sex-related factors, even if it is not individually associated with the TBDC decision.

While the classification tree expectedly demonstrates interactions among the six predictors in relation to a diagnosis of active PTB by the TBDC; the individual association of hemoptysis, alcoholic beverage drinking history, and age with the target variable is strikingly evident among females.

The loglinear model {AT}{HT}{TW}{TY}{WY} signifies that history of alcoholic beverage drinking (A), hemoptysis (H), weight loss (W), and age (Y) are individually associated with a TBDC diagnosis of active PTB (T); and there is an interaction between weight loss (W) and age (Y).

The complementary use of loglinear models and classification trees improves the conduct of association and classification analyses. The classification tree proves to be simple, interpretable, and is able to explicitly characterize and differentiate variable interactions between sexes in the attempt to describe the TBDC decision of diagnosing active PTB cases.

4.2 Recommendations

Presenting the most likely scenario of the TB Diagnostic Committee findings, the TBDC may find the proposed models useful as a decision tool in making a diagnosis of active pulmonary tuberculosis in support of the chest radiograph. The research outcome would contribute in minimizing subjectivity in chest x-ray readings and would make the most of the available simple findings for decision-making in the diagnostic approach to sputum smear-negative PTB patients.

The following additional recommendations are formulated for both statistical and public health aspects of the research. The author recommends to:

1. Use loglinear models and classification trees congruently in the construction of association and classification rules. Classification trees work effectively in the preliminary analysis for determining cutoffs and dichotomizing continuous and multilevel categorical variables prior to loglinear analysis.
2. Further explore categorical data using multiclass target variables in fitting classification trees. For instance, determine the association of demographic and clinical findings with the dependent variable having for target classes – a normal chest x-ray reading, minimal PTB lesions, and extensive PTB lesions on chest radiography.
3. Since the study is purely descriptive in modeling the TBDC diagnosis in terms of the associated demographic and medical characteristics; the interest then is to validate how well the TBDC findings measure up to sputum culture, the gold standard in the diagnosis of PTB. The author eagerly recommends observational and evaluation studies on the procedures and conduct of the TBDC deliberations.
4. The value of a good quality chest x-ray film to an accurate and prompt diagnosis of PTB is unquestionable. The author strongly proposes the formulation and implementation of strict measures to guarantee and monitor competent personnel

and well-equipped diagnostic centers to produce quality chest x-ray films of good visualization and with the recommended size of at least 11 by 14 inches. Efficient diagnosis leads to efficient TB control. Humanely and economically, patients with PTB will be spared from futile costs (caused by others' ineptitude) for undergoing repeated chest x-rays due to poor quality and inappropriately sized films.

REFERENCES

National Tuberculosis Control Program Manual of Procedures, 4th ed., 2005, Manila: Department of Health.

Clinical Practice Guidelines for the Diagnosis, Treatment, Prevention and Control of Tuberculosis in Adult Filipinos: 2006 Update. Manila: Philippine Society for Microbiology and Infectious Diseases (PSMID), Philippine College of Chest Physicians (PCCP), Philippine Coalition against Tuberculosis (PhilCAT), Philippine College of Physicians (PCP), Philippine College of Radiology (PCR), Philippine Academy of Family Physicians (PAFP), Philippine College of Occupational Medicine (PCOM), and the Department of Health (DOH).

International Standards for Tuberculosis Care [ISTC] (2006). Hague: Tuberculosis Coalition for Technical Assistance.

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: John Wiley and Sons, Inc.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Chong, Y.F. (2004), Combining Loglinear Model with Classification and Regression Tree (CART): An Application to Birth Data, *Computational Statistics and Data Analysis*, 45(4): 865-874.

Hand, D.J. (1997), *Construction and Assessment of Classification Rules*, West Sussex: John Wiley and Sons, Ltd.

Kanaya, A.M., Glidden, D.V. and Chambers, H.F. (2001), Identifying Pulmonary Tuberculosis in Patients with Negative Sputum Smear Results, *Chest* 120: 349-355.

Knoke, D. and Burke, P.J. (1980), Loglinear Models, *Sage University Paper Series on Quantitative Applications in the Social Sciences* 07-020, Newbury Park, CA: Sage Publications.

Mello, F., Bastos, L., Soares, S., Rezende, V., Conde, M., Chaisson, R., Kritski A., Ruffino-Netto, A. and Werneck, G. (2006), Predicting Smear-negative Pulmonary Tuberculosis with Classification Trees and Logistic Regression: A Cross-sectional Study, *BMC Public Health* 6: 43.

Rudas, T. (1998), Odds Ratios in the Analysis of Contingency Tables, *Sage University Paper series on Quantitative Applications in the Social Sciences* 07-119, Thousand Oaks, CA: Sage Publications.

Steinberg, D. and Golovnya, M. (2006), *CART 6.0 User's Manual*, San Diego, CA: Salford Systems.