A Relook at Logistic Regression Methods for the Initial Detection of Lung Ailments Using Clinical Data and Chest Radiography

OMAR MOHD RIJAL*, MOHD. IQBAL*, ASHARI YUNUS**, NORLIZA MOHD. NOOR***
*Institute of Mathematical Science, Faculty of Science, University Malaysia
**Institute of Respiratory Medicine, Kuala Lumpur, Malaysia
***Dept. of Electrical Engineering, College of Science and Technology, Universiti Teknologi Malaysia
omarrija@um.edu.my, iqbal1510@gmail.com, ashdr64@yahoo.com, norliza@ic.utm.my

Abstract:- The problem of diagnosing patients with lung ailments such as Tuberculosis (PTB), Pneumonia (PNEU) and Lung Cancer (LC) when making their initial visit to a medical institution is the focus of this study. Clinical data involving symptoms and signs are used to make important decisions before the availability of the results of further tests. In practice, Logistic Regression Methods are frequently involved in this type of decision making. However, the problem of missing values when the numerical values of certain explanatory variables are not available persists in practical situations. In this paper a logistic regression model using four variables (age, cough, loss of weight (LOW) and loss of appetite (LOA)) are investigated for each of the three diseases. The main result of this study is that the probability of misclassifying the three disease type is large, and that good model fitting does not guarantee correct diagnosis. As a viable substitute, a graphical method of detection with an 85% chance of correct classification based on information extracted from the chest radiograph images is proposed.

Keyword:- Statistical detection, error probability, lung disease, clinical data, chest radiography, missing values

1 Introduction

The mortality rate due to lung diseases is only second to that of cardiovascular diseases. A lung disease, as defined by [1] is any disease or disorder where lung function is impaired. In this study three of the major lung diseases in Malaysia are considered, namely Pulmonary Tuberculosis (PTB), Pneumonia (PNEU) and Lung Cancer (LC). Ignoring these three diseases could be fatal. In particular PTB and PNEU are extremely infectious diseases but treatable if diagnosed early. LC is incurable, but with early detection, it is still possible to treat it. The similarities of these diseases are that early detection is essential. Detection of these diseases includes taking the medical history, physical examination and laboratory or radiography information.

The medical history of patients is extremely important in diagnosing a disease. A study done by [2] found that in 66 out of 80 cases, the medical history provided enough information to make an initial diagnosis of a specific disease entity which agreed with the one finally accepted. Another study [3], concludes that 76% of cases can be diagnosed correctly using the medical history. Several studies [4-7] also concluded that the medical history is the biggest component in making a medical diagnosis. Similarly, [8] which surveyed the perception of physicians instead of examining patients observed that doctors perceive the medical history of patients as having much higher value in diagnosis than either the physical examination or laboratory/radiography information.

However, in many cases the medical history of a patient may be incomplete or even totally absent as such decisions must still be made based on clinical data and chest radiograph.

2 Methods

Cases considered in this study comprise of patients of the Institute of Respiratory Medicine (IPR), Kuala Lumpur, Malaysia aged between 15 to 82 years old with confirmed diagnosis of Tuberculosis, Pneumonia or Lung Cancer between 2004 and 2007. Patients come from all over the country as the Institute is a referral hospital housing the country's most experienced respiratory experts. Every patient referred to IPR brings along all documents from the previous medical institution. In total, patient records in IPR is in the form of wallets containing the patient file, pathology results and radiology films. The wallets were then showed to a pulmonologist consultant to reconfirm the initial diagnosis written in the file to ensure that the correct cases will be further analyzed.

A close scrutiny of the IPR patient wallets indicated the existence of the missing-value problem. This study concentrates on 140 patients (40 PTB, 40 PNEU, 40 LC and 20 normal individuals) where only the four explanatory variables (age, cough, LOW and LOA) were available together with the response disease present or otherwise. After consultation with the pulmonologist, coughing was categorized into four states according to the degree of seriousness, specifically occasional coughing, intermittent coughing, acute coughing, persistent coughing and chronic coughing. LOW and LOA on the other hand have only two stages, namely whether it is present or not. Table 1 shows the numerical values used to represent the states of the symptoms. The data in this form is a natural candidate for Logistic Regression modelling [9-15]. A PTB model was developed using the 40 PTB patients and 20 normals. Similarly, a PNEU model and a LC model were derived. For every model, the constant term (β_0) and the coefficients ($\beta_1,...,\beta_p$) were estimated by the maximum likelihood method.

A stepwise regression method [12, 16] was used to build the model from a base model. The method used was the forward selection process, which involves starting with no variables in the model, trying out the variables one by one and including them if they are statistically significant. The indicators used for this purpose are the R_1^2, R_2^2 and R_3^2 [17-19]. The selected model (for given disease) was subjected to the leave-one-out method [20, 21] to determine the robustness of the model and to investigate presence of outliers.

The PTB model estimates the probability of detecting Tuberculosis given age, cough, LOW and LOA. Similarly, the PNEU model estimates the probability of detecting Pneumonia. Henceforth, the ratio of the two types of probability can be compared.

3 The Logistic Model

The data is consist of 40 PTB cases, 40 PNEU cases, 40 LC cases, and 20 healthy individuals. The model chosen for the analysis is the logistic regression model (LRM) package available in S-Plus® by TIBCO Software Inc. The LRM provides that the response variable (dependent variable) is either 0 or 1. For each group with patient *i*, $Y_i = 1$ will denote the cases with a positive diagnosis of a disease while $Y_i = 0$ denotes a negative diagnosis or in other words, the patient is healthy. The LRM may be expressed as follows, let p be the number of variables to be considered:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}, \quad (1)$$

or alternatively in the logit form,

$$g(x) = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \quad (2)$$

The term $\pi(x)$ may be interpreted as the probability of patient *i* having a disease given $x_1, x_2, \cdots, x_p$. It will be a value between 0 and 1. The estimates of β_0, $\beta_1,...,\beta_p$ were obtained using S-Plus.

4 Maximum Likelihood Estimation

Each observation for the response variable Y_i is an ordinary Bernoulli [14,15] observation, hence,

$$Y_i \sim Bernoulli(\pi_i) \text{ with}$$
$$logit(\pi_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \quad i = 1, \ldots, n.$$

In particular,

$$prob(Y_i = 1) = \pi_i$$
$$prob(Y_i = 0) = 1 - \pi_i$$

The likelihood function is given as follows:

$$L(Y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

There will be $p+1$ likelihood equations that are obtained by differentiating the log-likelihood function with respect to the $p+1$ coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^{n} \left[y_i - \boldsymbol{\pi}(\mathbf{x}_i)\right] = 0$$

and

$$\sum_{i=1}^{n} x_{ij} \left[y_i - \boldsymbol{\pi}(\mathbf{x}_i)\right] = 0$$

for $j = 1, 2, \ldots, p$ [12].

5 Goodness of fit

Once the equation has been obtained, the goodness of-fit of the model must be verified. Readily available information from the software's output includes the deviance, $D = -2\log(\hat{L}_c / \hat{L}_f)$ where $\hat{L}_c$ is the likelihood of the current model and $\hat{L}_f$ is the likelihood of the full model. Normally, the deviance is used to summarize the goodness of fit of a model for grouped binary (binomial) data. Since the data considered is not grouped, the deviance cannot be used as a goodness of fit measure for binary data. Furthermore, the deviance is also unreliable as a measure of goodness of fit when the data is sparse (small sample) or not completely grouped. [13] Some other statistics that summarizes model adequacy are therefore considered.

Unlike the simple linear regression model with its coefficient of determination [12, 13]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

there is no one indicator for the goodness-of-fit for a LRM. Instead, there are three analogues of the R^2, in particular R_1^2, R_2^2 and R_3^2, [13, 17] where,

$$R_1^2 = 1 - \frac{\log L(\hat{\beta})}{\log L_0} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

applies the definition of R^2 directly by using Y_i in place of y_i and $\hat{\pi}_i$ in place of $\hat{\mu}_i$. The second analogue of R^2 is given by

$$R_2^2 = 1 - \left[\frac{\hat{L}_0}{L(\hat{\beta})}\right]^{2/n},$$

where

$$L(\hat{\beta}) = \sum_{i=1}^{n}\left[Y_i \log(\pi_i) + (1 - Y_i)\log(1 - \pi_i)\right]$$

is the maximized likelihood for the model of interest,

$$L_0 = n_0\left[\log\left(\frac{n_0}{n}\right)\right] + n_1\left[\log\left(\frac{n_1}{n}\right)\right]$$

is the maximized likelihood for the model that contains a constant term alone. The third R^2 analogue, R_3^2 is given below

$$R_3^2 = R_2^2 / R_{\max}^2$$

where $R_{\max}^2 = 1 - \{\hat{L}_0\}^{2/n}$. [18] concluded that R_2^2 is most satisfactory because it is a natural extension of R^2 both mathematically and conceptually. However, [19] concluded that R_3^2 has a few desirable properties that makes it suitable to represent the goodness-of-fit. Therefore, in this study all three analogues will be considered.

The forward stepwise regression method was applied for the selection of explanatory variables. In this method, the basic model consisting of one explanatory variable was considered first, followed by considering pairs of explanatory variables up to the stage when the R_1^2, R_2^2 and R_3^2 have the largest numerical values. S-Plus was used for all programming.

6 Robustness of LRM

The best model selected from the forward stepwise regression method needs to be tested for robust properties. The LRM is defined as robust if no individual patient's data can greatly alter the following statistics:

(i) the values of R_1^2, R_2^2 and R_3^2
(ii) the coefficients and constant terms for an individual model
(iii) predicted values, $\hat{\pi}(x)$, see (1) and (2)

The leave-one-out method removes one patient from the sample, recalculates the statistics stated in (i) and (ii), and then resubstitutes the data of the first patient into the recalculated model to obtain the new predicted value. If any of the statistics in (i), (ii) and (iii) changes significantly, then this patient is considered an outlier, and removed accordingly from the study.

If the first person is not considered an outlier, he is returned to the sample and the second patient is in turn removed and the above process is repeated.

7 Results

(a) The tuberculosis model

The logistic regression model was fitted to the data set involving 40 PTB cases and 20 normal individuals. The data for each case are;

(i) Y = 1 if patient is PTB case, 0 if normal
(ii) Patient's age
(iii) Indicator for cough
(iv) Indicator for loss of weight

The indicator for LOA was not considered for the PTB model because the values for LOW and LOA are identical for all 40 cases. Therefore, by including LOA in the estimation would yield identical values for both LOA and LOW. Table 2 calculates R_1^2, R_2^2 and R_3^2 for selected explanatory variables suggested by the pulmonologist.

Based on R_1^2, R_2^2 and R_3^2, the model

$$g(x) = -4.1722 + 0.0511\text{Age} + 0.0644\text{Cough} + 14.4630\text{LOW} \quad \text{(M1)}$$

is selected using the stepwise regression method. To investigate whether (M1) is robust, the leave-one-out method was used. The constant and the coefficients of the model is generally stable. This is also true with R_1^2, R_2^2 and R_3^2 and the predicted values. Further, the leave-one-out method also indicated two outliers (observations 14 and 15, see Table 3). After the removal of the two outliers, the model (M1) is adjusted and takes the following form;

$$g(x) = -24.5661 - (9.2781 \times e^{-17})\text{Age} + (1.8867 \times e^{-14})\text{Cough} + 49.1321\text{LOW} \quad \text{(M2)}$$

For (M2), R_1^2 = 0.9999, R_2^2 = 0.7242 and R_3^2 = 0.9999. To further investigate the validity of (M2), the error probability was calculated.

Let $p_1(PTB) = prob(PTB \mid PNEU)$ be the probability of misclassifying pneumonia patients as being infected with PTB. Using a test set of 40 pneumonia patients gives $p_1(PTB) = 0.425$.

Let $p_2(PTB) = prob(PTB \mid LC)$ be the probability of misclassifying lung cancer patients as being infected with PTB. Using a test set of 40 lung cancer patients gives $p_2(PTB) = 0.725$.

(b) The pneumonia model

The experiment in (a) above was repeated but for 40 Pneumonia patients and the same 20 normal individuals. Table 4 gives the selected models

Again based on the correlation coefficients, the following model was accepted;

$$g(x) = 0.2659 - 0.0273\text{Age} + 9.4144\text{Cough} + 11.1150\text{LOW} \quad \text{(M3)}$$

Note that Table 4 shows much lower correlations compared to the PTB model. The possible presence of outliers was again investigated, but only for model (M3)

The adjusted model (M4) is given as follows;

$$g(x) = 2.0370 - 0.0714\text{Age} + 9.9496\text{Cough} - 7.5561\text{LOW} \quad \text{(M4)}$$

with R_1^2 = 0.6641, R_2^2 = 0.5750 and R_3^2 = 0.7939

Let $p_1(PNEU) = prob(PNEU \mid PTB)$ be the probability of misclassifying PTB patients as being infected with pneumonia. Using a test set of 40 pneumonia patients gives $p_1(PNEU) = 0.975$.

Let $p_2(PNEU) = prob(PNEU \mid LC)$ be the probability of misclassifying lung cancer patients as being infected with pneumonia. Using a test set of 40 lung cancer patients gives $p_2(PNEU) = 0.85$.

(c) The lung cancer model

The experiment in (a) and (b) above was repeated but for 40 LC patients and the same 20 normal individuals. Table 5 gives the selected models. Again based on the correlation coefficients, the following model was accepted;

$$g(x) = -15.1657 + 0.2080\text{Age} + 13.4477\text{Cough} + 12.0347\text{LOA} \quad \text{(M5)}$$

The possible presence of outliers was again investigated, but only for model (M5)

The leave one out method detected a solitary outlier and the recalculated model is as follows;

$$g(x) = -32.1918 + 0.4312\text{Age} + 18.5982\text{Cough} + 14.4852\text{LOA} \quad \text{(M6)}$$

with R_1^2 = 0.9450, R_2^2 = 0.7019 and R_3^2 = 0.9719.

Let $p_1(LC) = prob(LC \mid PTB)$ be the probability of misclassifying PTB patients as being infected with LC. Using a test set of 40 PTB patients gives $p_1(LC) = 0.95$.

Let $p_2(LC) = prob(LC \mid PNEU)$ be the probability of misclassifying pneumonia patients as being infected with Lung Cancer. Using a test set of 40 pneumonia patients gives $p_2(LC) = 0.8$.

8 Ratio of detection probabilities

Although using the correlations R_1^2, R_2^2 and R_3^2 suggest that (M2), (M4) and (M6) are the best models for detecting PTB, PNEU and LC respectively, however the associated error probabilities are large. As such, it may be useful to consider the ratios of the $\pi(i)$ (probability of detection).

Let $\pi(PTB)$ be the probability of detecting PTB when using (M2), π(PNEU) be the probability of detecting Pneumonia when using (M4) and π(LC) be the probability of detecting LC when using (M6).

The relevant ratios are listed out in Table 6 using a test sample of 10 cases for each disease type. The ratio π(PNEU)/π(PTB) is given in the first column of Table 5. Since the test cases are confirmed PTB patients, hence the ratio should be less than one. Table 6 suggests the contrary implying that when a 'new' or 'unknown' patient is in fact infected with PTB, using M2 and M4 will not help in confirmation of disease-status.

In general, Table 6 indicates the π-probabilities (in most cases) are very similar, hence the pair-wise comparisons does not help differentiate the diseases.

9 Discussion

The initial study of respiratory diseases usually begins with the use of clinical data which are prone to missing-value problems. In this study only age, cough, LOW and LOA were available as indicators for initial screening for 140 patients. The logistic models used in this study appear to suggest that the frequently used explanatory variables age, cough, LOW and LOA cannot differentiate the three diseases confidently, hence the perpetual problem of selecting appropriate explanatory variables.

As such, we strongly suggest a graphical method based on the use of the Andrews Curve as a viable substitute to the use of logistic regression models for purposes of initial screening. This graphical method has been reported in [22] and was shown to have 85% chance of correct detection (Table 7). This section will now briefly describe the graphical method below. For a given chest X-ray, a region of interest (ROI) (Fig. 1) is selected from which a set of line profiles are chosen. Each line profile may be interpreted as a signal (Fig. 2) which in turn is subjected to the Daubechies 4 transformation. The average of these signals in the form of a vector of Daubechies coefficients represents the ROI. This average vector is then represented as an Andrews Curve. Given two patients, hence two average vectors, we will have two Andrews Curves. The vertical separation between two Andrews Curves is equivalent to the Euclidean distance between the two average vectors. Fig. 3 shows that for a given *t* value (along horizontal axis) three distinct clusters is clearly seen and henceforth the probability of classification for each disease type may be estimated.

10 Conclusion

The success of initial screening depends heavily on the selection of explanatory variables which may be applied in decision making using the logistic model. Using correlations, this study suggests three possible models for the detection of PTB, Pneumonia and Lung cancer. However, the error probabilities p_1 and p_2 are large suggesting that the selection of models should not be based on correlations alone. Ensuring robust logistic regression models did not help in differentiating the three diseases. This remark is further supported by studying ratios of the probability of detection π(PTB), π(PNEU) and π(LC). Instead of looking for other explanatory variables, as a viable substitute, the use of the proposed graphical method involving Andrews' curves is strongly recommended.

References

[1] American Lung Association. *Lung Disease Data: 2008*. New York: American Lung Association, 2008.

[2] Hampton J.R., Harrison M.J., Mitchell J.R., et al. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *British Medical Journal*, **2**:486-9, 1975.

[3] Peterson M.C., Holbrook J.H., Hales D.V., et al. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *The Western Journal of Medicine*, **156(2)**:163-5, 1992.

[4] Bordage G. Where are the history and the physical? *Canadian Medical Association Journal*. **152(10)**:1595-8, 1995.

[5] Crombie D.L. Diagnostic Process. *Journal of the College of General Practitioners*, **6(4)**:579-89, 1963.

[6] Roshan M. and Rao A.P. A study on relative contributions of the history, physical examination and investigations in making medical diagnosis. *Journal of the Association of Physicians of India*, **48(8)**:771-5, 2000.

[7] Reilly B.M. Physical examination in the care of medical inpatients: an observational study. *The Lancet*, **362**:1100-5, 2003.

[8] Rich E.C., Crowson T.W. and Harris I.B. The diagnostic value of the medical history. *Archives of Internal Medicine*, **147(11)**:1957-60, 1987.

[9] Kaur D. and Pulugurta H. Comparitive analysis of fuzzy decision trees and logistic regression methods for pavement treatment prediction. *WSEAS Transaction Information Science & Applications*, **6(5)**:979-90, 2008.

[10] Perez A.G., Rivas E.T., Echeverria F.R., et al. Logistic regression model for determining risks factor for hypertensive disorders in pregnancy. *WSEAS Transaction on Systems*, **5(9)**:2135-9, 2006.

[11] Sarlija N., Bensic M., and Zekic-Susac M. Logistic regression, survival analysis and neural networks in modeling customer credit scoring. *WSEAS Transactions on Business and Economics*, **3(3)**:156-62, 2006.

[12] Hosmer D.W. and Lemeshow S. *Applied Logistic Regression Analysis*, 2nd ed. New York: Wiley, 2000.

[13] Collett D. *Modelling binary data*, 2nd ed. Boca Raton, Fl.: Chapman & Hall/CRC, 2003.

[14] Agresti A. *Categorical Data Analysis*, New York: Wiley, 1990.

[15] Cox D.R. and Snell E.J. *The Analysis of Binary Data*, 2nd ed. London: Chapon & Hall, 1989.

[16] Menard S. *Applied Logistic Regression*, Thousand Oaks, CA: University Paper Series no. 07-105, 1995.

[17] Liao J.G and McGee D. Adjusted coefficients of determination for logistic regression. *The American Statistician*, **57(3)**:161-5, 2003

[18] Menard S. Coefficients of determination for multiple logistic regression analysis. *The American Statistician,* **54**:17-24, 2000.

[19] Nagelkerke, N.J.D. A note on a general definition of the coefficient of determination. *Biometrika,* **78:**691-2, 1991.

[20] Devroye L. P. and Wagner T. J. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, **25(2)**:202-7, 1979.

[21] Cawley G.C. and Talbot N.L.C. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition*, **36(11)**:2585-92, 2003.

[22] Rijal O.M., Noor N.M., Hussin A., et al. Using Statistical Features to Verify the Gold Standard for Pulmonary Tuberculosis Detection, *Inter. Journal of Comp. Assisted Radiology and Surgery*, 3(1):S426-428, 2008.

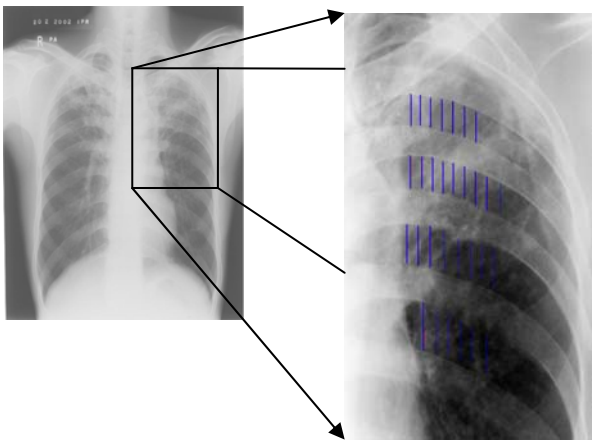

Fig.1: Line profiles taken on the region of interest (Source: Malaysian Institute of Respiratory Medicine).

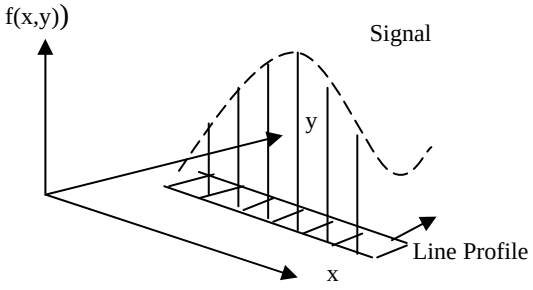

Fig. 2: A line profile: A two-dimensional light intensity function f(x, y), where x and y denotes spatial coordinates.

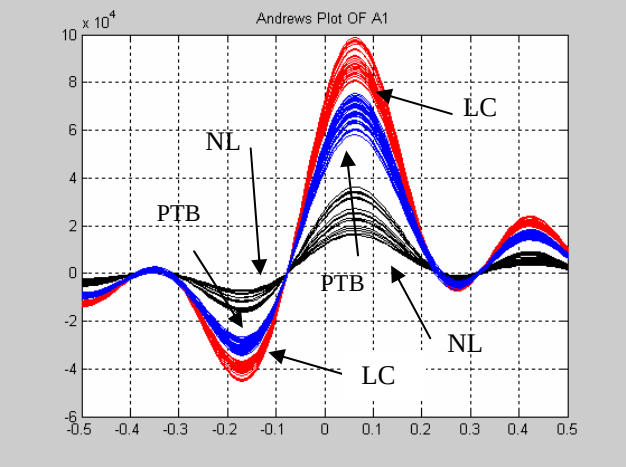

Fig. 3: Andrews curve of 90 average signals, for control group of 20 normal lung (NL), 40 pulmonary tuberculosis (PTB) patients and 30 lung cancer (LC) patients plot over the range of –0.5 to 0.5.

Table 1: Selected data available during the first consultation with a medical officer

Variable	Original values	New values
Age	Discrete	Unchanged
Cough	on and off (occasional)	1
	Intermittent coughing	2
	Acute coughing (more than two weeks)	3
	Persistent and chronic coughing	4
LOW / LOA	No	0
	Yes	1

Table 2: Stepwise regression for PTB (PTB = 40 cases, NL = 20 cases)

Estimated $g(x)$	R_1^2	R_2^2	R_3^2
-1.7405 + 0.0617Age	0.1204	0.1421	0.1974
-2.3026 + 10.7118Cough	0.8245	0.6499	0.9026
-16.1712 + 13.8686LOW	0.8245	0.6499	0.9026
-2.3026 - (1.5855×10^{-14})Cough + 13.8686LOW	0.8245	0.6499	0.9026
-4.1722 + 0.0511Age + 10.3586Cough	0.8314	0.6530	0.9069
-4.1722 + 0.0511Age + 14.6419LOW	0.8314	0.6530	0.9069
-4.1722 + 0.0511Age + 0.0644Cough + 14.4630LOW	0.8314	0.6530	0.9069

Table 3(a) Leave-one-out method for 40 PTB patients

Omit	Estimated $g(x)$	R_1^2	R_2^2	R_3^2	$\hat{Y}$
1	-4.1722 + 0.0511Age + 0.0661Cough + 14.4586LOW	0.8296	0.6544	0.9062	1.0000
2	-4.1722 + 0.0511Age + 0.0683Cough + 14.4495LOW	0.8296	0.6544	0.9062	1.0000
3	-4.1722 + 0.0511Age + 0.0461Cough + 14.5255LOW	0.8296	0.6544	0.9062	1.0000
4	-4.1722 + 0.0511Age + 0.1021Cough + 14.3962LOW	0.8296	0.6544	0.9062	1.0000
5	-4.1722 + 0.0511Age + 0.1022Cough + 14.3312LOW	0.8296	0.6544	0.9062	1.0000
6	-4.1722 + 0.0511Age + 0.0543Cough + 14.4976LOW	0.8296	0.6544	0.9062	1.0000
7	-4.1722 + 0.0511Age + 0.0644Cough + 14.4630LOW	0.8296	0.6544	0.9062	1.0000
8	-4.1722 + 0.0511Age + 0.0968Cough + 14.3502LOW	0.8296	0.6544	0.9062	1.0000
9	-4.1722 + 0.0511Age + 0.0727Cough + 14.4344LOW	0.8296	0.6544	0.9062	1.0000
10	-4.1722 + 0.0511Age + 0.0446Cough + 14.5309LOW	0.8296	0.6544	0.9062	1.0000
11	-4.1722 + 0.0511Age + 0.0600Cough + 14.4708LOW	0.8296	0.6544	0.9062	1.0000
12	-4.1722 + 0.0511Age + 0.0535Cough + 14.5004LOW	0.8296	0.6544	0.9062	1.0000
13	-4.1722 + 0.0511Age + 0.1947Cough + 14.1885LOW	0.8296	0.6544	0.9062	1.0000
14	465.4635 - 21.1455Age + 106.6397Cough + 826.3961LOW	-	-	-	-
15	-326.6535 + 5.9983Age - 4.5608Cough + 227.1048LOW	-	-	-	-
16	-4.1722 + 0.0511Age – 0.1219Cough + 14.8596LOW	0.8296	0.6544	0.9062	1.0000
17	-4.1722 + 0.0511Age + 0.0479Cough + 14.5196LOW	0.8296	0.6544	0.9062	1.0000
18	-4.1722 + 0.0511Age + 0.0436Cough + 14.5342LOW	0.8296	0.6544	0.9062	1.0000
19	-4.1722 + 0.0511Age + 0.0031Cough + 14.5701LOW	0.8296	0.6544	0.9062	1.0000
20	-4.1722 + 0.0511Age + 0.0609Cough + 14.4747LOW	0.8296	0.6544	0.9062	1.0000
21	-4.1722 + 0.0511Age + 0.0632Cough + 14.4669LOW	0.8296	0.6544	0.9062	1.0000
22	-4.1722 + 0.0511Age + 0.0991Cough + 14.4014LOW	0.8296	0.6544	0.9062	1.0000
23	-4.1722 + 0.0511Age + 0.0852Cough + 14.3908LOW	0.8296	0.6544	0.9062	1.0000
24	-4.1722 + 0.0511Age + 0.1096Cough + 14.3830LOW	0.8296	0.6544	0.9062	1.0000
25	-4.1722 + 0.0511Age + 0.0776Cough + 14.4173LOW	0.8296	0.6544	0.9062	1.0000
26	-4.1722 + 0.0511Age + 0.0467Cough + 14.5236LOW	0.8296	0.6544	0.9062	1.0000
27	-4.1722 + 0.0511Age + 0.0805Cough + 14.4345LOW	0.8296	0.6544	0.9062	1.0000
28	-4.1722 + 0.0511Age + 0.0446Cough + 14.5309LOW	0.8296	0.6544	0.9062	1.0000
29	-4.1722 + 0.0511Age + 0.0431Cough + 14.5005LOW	0.8296	0.6544	0.9062	1.0000
30	-4.1722 + 0.0511Age + 0.0896Cough + 14.4183LOW	0.8296	0.6544	0.9062	1.0000
31	-4.1722 + 0.0511Age + 0.0968Cough + 14.3502LOW	0.8296	0.6544	0.9062	1.0000
32	-4.1722 + 0.0511Age + 0.0432Cough + 14.5358LOW	0.8296	0.6544	0.9062	1.0000
33	-4.1722 + 0.0511Age + 0.0498Cough + 14.5130LOW	0.8296	0.6544	0.9062	1.0000
34	-4.1722 + 0.0511Age + 0.0079Cough + 14.5617LOW	0.8296	0.6544	0.9062	1.0000
35	-4.1722 + 0.0511Age + 0.0543Cough + 14.4975LOW	0.8296	0.6544	0.9062	1.0000
36	-4.1722 + 0.0511Age + 0.0599Cough + 14.4783LOW	0.8296	0.6544	0.9062	1.0000
37	-4.1722 + 0.0511Age + 0.0055Cough + 14.4946LOW	0.8296	0.6544	0.9062	1.0000
38	-4.1722 + 0.0511Age + 0.0896Cough + 14.3756LOW	0.8296	0.6544	0.9062	1.0000
39	-4.1722 + 0.0511Age + 0.0896Cough + 14.3756LOW	0.8296	0.6544	0.9062	1.0000
40	-4.1722 + 0.0511Age + 0.0462Cough + 14.4951LOW	0.8296	0.6544	0.9062	1.0000

Table 3 (b) Leave-one-out method for 20 normals.

Omit	Estimated $g(x)$	R_1^2	R_2^2	R_3^2	$\hat{Y}$
41	-4.1757 + 0.0529Age + 0.0692Cough + 14.3861LOW	0.8298	0.6476	0.9052	0.1241
42	-3.9973 + 0.0474Age + 0.0547Cough + 14.4503LOW	0.8279	0.6467	0.9040	0.0567
43	-4.1995 + 0.0536Age + 0.0711Cough + 14.3794LOW	0.8300	0.6477	0.9053	0.1307
44	-4.1348 + 0.0517Age + 0.0659Cough + 14.3981LOW	0.8295	0.6474	0.9049	0.1124
45	-4.0334 + 0.0486Age + 0.0578Cough + 14.4324LOW	0.8285	0.6470	0.9043	0.0774
46	-4.0513 + 0.0492Age + 0.0593Cough + 14.4255LOW	0.8319	0.6554	0.9076	0.0848
47	-4.0083 + 0.0478Age + 0.0557Cough + 14.4437LOW	0.8313	0.6552	0.9072	0.0648
48	-4.0192 + 0.0481Age + 0.0567Cough + 14.4385LOW	0.8315	0.6552	0.9073	0.0707
49	-4.0134 + 0.0479Age + 0.0562Cough + 14.4412LOW	0.8314	0.6552	0.9073	0.0676
50	-4.0003 + 0.0475Age + 0.0550Cough + 14.4482LOW	0.8312	0.6551	0.9071	0.0592
51	-3.9948 + 0.0473Age + 0.0545Cough + 14.4522LOW	0.8310	0.6550	0.9071	0.0542
52	-4.0867 + 0.0503Age + 0.0621Cough + 14.4132LOW	0.8323	0.6556	0.9078	0.0975
53	-3.9930 + 0.0472Age + 0.0543Cough + 14.4539LOW	0.8310	0.6550	0.9070	0.0518
54	-4.2871 + 0.0562Age + 0.0781Cough + 14.3563LOW	0.8338	0.6563	0.9087	0.1542
55	-4.0040 + 0.0476Age + 0.0554Cough + 14.4461LOW	0.8313	0.6551	0.9072	0.0619
56	-4.3633 + 0.0584Age + 0.0843Cough + 14.3376LOW	0.8343	0.6565	0.9091	0.1736
57	-4.5172 + 0.0628Age + 0.0969Cough + 14.3035LOW	0.8352	0.6569	0.9096	0.2118
58	-4.1348 + 0.0517Age + 0.0659Cough + 14.3981LOW	0.8327	0.6558	0.9081	0.1124
59	-3.9973 + 0.0474Age + 0.0547Cough + 14.4503LOW	0.8311	0.6551	0.9071	0.0567
60	-4.0083 + 0.0478Age + 0.0557Cough + 14.4437LOW	0.8313	0.6552	0.9072	0.0648

Table 4: Stepwise regression for PNEU (PNEU = 40 cases, NL = 20 cases)

Estimated $g(x)$	R_1^2	R_2^2	R_3^2
1.4044 - 0.0146Age	0.3062	0.4767	0.6621
-0.9163 + 9.7527Cough	0.5614	0.5106	0.7092
0.1398 + 9.4262LOW	0.2223	0.2465	0.3423
$(-3.0130 \times e^{-16}) + 9.5659$LOA	0.2739	0.2945	0.4089
-1.0498 + 9.7699Cough + 10.0811LOW	0.5954	0.5314	0.7380
-1.0498 + 9.7724Cough + 9.9024LOA	0.5954	0.5314	0.7380
-0.8401 - 0.0015Age + 9.7198Cough	0.5614	0.5106	0.7092
1.6580 - 0.0322Age + 9.6576LOW	0.2916	0.3101	0.4307
1.5565 - 0.0329Age + 9.7657LOA	0.3018	0.3190	0.4431
0.2659 - 0.0273Age + 9.4144Cough + 11.1150LOW	0.6042	0.5366	0.7452
0.2659 - 0.0273Age + 9.4119Cough + 10.9429LOA	0.6042	0.5366	0.7452

Table 5: Stepwise regression for LC (LC = 40 cases, NL = 20 cases)

Estimated $g(x)$	R_1^2	R_2^2	R_3^2
-10.9013 + 0.1870Age	0.4200	0.4142	0.5752
-1.2040 + 9.9973Cough	0.6322	0.5528	0.7678
-0.5978 + 11.1639LOW	0.4721	0.4517	0.6273
-0.3567 + 10.9227LOA	0.3968	0.3966	0.5508
-1.8971 + 11.4988Cough + 12.5514LOW	0.8034	0.6404	0.8894
-1.8971 + 11.4654Cough + 12.6668LOA	0.8034	0.6404	0.8894
-17.7944 + 0.2572Age + 12.3140Cough	0.8806	0.6740	0.9362
-10.2549 + 0.1572Age + 11.3096LOW	0.6608	0.5688	0.7900
-9.6610 + 0.1529Age + 9.8876LOA	0.6005	0.5344	0.7422
-15.1657 + 0.2080Age + 13.5002Cough + 11.8732LOW	0.8314	0.6530	0.9069
-15.1657 + 0.2080Age + 13.4477Cough + 12.0347LOA	0.8681	0.6688	0.9289

Table 6: Ratio of detection probabilities

PTB test sample ($n_1 = 10$)		PNEU test sample ($n_2 = 10$)		LC test sample ($n_3 = 10$)	
$\frac{\pi(PNEU)}{\pi(PTB)}$	$\frac{\pi(LC)}{\pi(PTB)}$	$\frac{\pi(PTB)}{\pi(PNEU)}$	$\frac{\pi(LC)}{\pi(PNEU)}$	$\frac{\pi(PTB)}{\pi(LC)}$	$\frac{\pi(PNEU)}{\pi(LC)}$
1.0857	1.0857	0.9210	1	0.9210	0.9999
1.0857	1.0857	0.9531	1.0345	0.9210	1
1.0857	1.0857	0.1322	0	0.9210	0.9999
1.0857	1.0857	0.0790	1	0.9210	0.9999
1.0857	1.0857	0.9210	1	0.0790	1
1.0857	1.0857	0.9210	1	0.9210	1
12.6651	12.6651	0.9210	1	0.9210	0.9999
1.0857	0.5439	0.0790	1	0.9210	0.9999
1.0857	1.0857	0.0790	1	0.9210	0.9999
1.0857	1.0857	0.9210	1	0.9210	0.9999

Table 7: Discrimination result using Andrews' curve.

CASE TYPE	NO. OF CORRECTLY CLASSIFIED CASES	MISCLASSIFIED CASES	TOTAL CASES
NL	16	3	19
PTB	144	24	168
LC	23	3	26