Some Critical Remarks on 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.au, norliza@ic.utm.my

Abstract:- An important area of health care is the problem of detecting lung ailments such as tuberculosis, pneumonia and lung cancer for individuals making their initial visit to a medical institution. Clinical data involving symptoms and signs has been the centre of interest for diagnosing patients. Important decisions sometimes need to be made before the availability of the results of further tests. However, besides clinical data, the chest radiograph is also available. This study uses logistic regression models to critically reconsider the use of age, loss of appetite, loss of weight and cough for predicting respiratory disease type. 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

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 Tuberculosis (TB), Pneumonia (PNEU) and Lung Cancer (LC). Ignoring these three diseases could be fatal. In particular TB 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. A few other papers [4], [5], [6], [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, laboratory or 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, Kuala Lumpur, Malaysia aged between 15 to 82 years old with confirmed diagnosis of either 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. Patient records were obtained from the records office in the form of wallets containing the patient file, pathology results and radiology films. The wallets were then showed to a medical officer to confirm the diagnosis written in the file to ensure that the correct cases will be further analyzed.

The most common symptoms for all the three diseases considered in this study are coughing, loss of weight (LOW) and loss of appetite (LOA). After consultation with medical officers, coughing can be divided into four stages, from least chronic to most chronic, represented by increasing numbers. LOW and LOA on the other hand have only two stages, namely whether it is present or not. Table 1 shows the
numbers used to represent the states of a given symptom.

A sample of 40 patients for each disease and a group of 20 healthy individuals constitute the whole data set. For the healthy individuals, only the variable age is used.

3 The Logistic Model

The data is separated into three groups according to the disease they have plus the 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 the response variable (dependent variable) is either 0 or 1. For each group, 1 will denote the cases with a positive diagnosis of a disease while 0 denotes a negative diagnosis or in other words, the patient is healthy. The LRM may be expressed as follows:

\[ \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \]

or alternatively in the logit form,

\[ g(x) = \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x. \]

\(\pi(x)\) will be a value between 0 and 1 denoting the probability of the patient having a disease. S-Plus will provide the estimates of \(\beta_0\) and \(\beta_1\).

4 Maximum Likelihood Estimation

Each observation is an ordinary Bernoulli observation:

\[ \text{prob}(Y_i = 1) = \pi, \]
\[ \text{prob}(Y_i = 0) = 1 - \pi, \]

The likelihood function is,

\[ L(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}, \]

and,

\[ \frac{\partial \ln L(Y_i)}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 X_i) \]
\[ \sum_{i=1}^{n} X_i, \]
\[ \frac{\partial \ln L(Y_i)}{\partial \beta_1} = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 X_i) (X_i). \]

Equating both partial derivatives to zero and using numerical methods yields the maximum likelihood estimates of \(\beta_0\) and \(\beta_1\).

Although the univariate logistic model is discussed above, this study applied the multivariate version and details may be easily obtained from standard text book, [9].

5 Goodness of fit

Once the equation has been obtained, the goodness-of-fit of the model must be verified. Unlike the simple linear regression model with its coefficient of determination \((R^2)\), there is no one indicator for the goodness-of-fit for a LRM. Instead, there are three analogues of the \(R^2\), i.e. \(R_1^2\), \(R_2^2\) and \(R_3^2\), [10]:

\[ R_1^2 = \frac{1 - \log L(\hat{\beta})}{\log L_0}, \]
\[ R_2^2 = 1 - \frac{\hat{L}_0}{L(\hat{\beta})}, \]
\[ R_3^2 = R_2^2 / R_{\text{max}}^2, \]

where \(L(\hat{\beta})\) is the maximized likelihood for the model of interest, \(L_0\) is the maximized likelihood for the model that contains a constant term alone and \(R_{\text{max}}^2 = 1 - \{\hat{L}_0\}^{1/n}\).

6 Results

(a) The tuberculosis model

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

(i) \(Y = 1\) if patient is TB 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 TB 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 medical expert.

Based on \(R_1^2\), \(R_2^2\) and \(R_3^2\), the model

\[ g(x) = -4.1722 + 0.0511 \text{Age} + 0.0644 \text{Cough} + 14.4630 \text{LOW} \]

(M1) is selected. To investigate whether (M1) is robust, the leave-one-out method was used. In this method, one observation was removed and the coefficients of the model and the corresponding correlations were again recalculated. Then, the second observation was removed, but the first was returned to the data set. Once again, the relevant coefficients and correlations were recalculated. The leave-one-out method indicated two outliers. The adjusted model (M1) is as follows;

\[ g(x) = -24.5661 - (9.2781 \times \text{Age} + 1.8867 \times \text{Cough} + 49.1321 \text{LOW} \]

(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 \(p(TB) = \text{prob}(TB | PNEU)\) be the probability of misclassifying pneumonia patients as
being infected with TB. Using a test set of 40 pneumonia patients gives $p_p(TB) = 0.425$.

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

(b) The pneumonia model

The experiment in (a) above was repeated but for 40 Pneumonia patients and the same 20 normal individuals. Table 3 gives the selected models.

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

\[ g(x) = 0.2659 - 0.0273 \text{Age} + 9.4144 \text{Cough} + 11.1150 \text{LOW} \]  

(M3)

Note that Table 3 shows much lower correlations compared to the TB model. The possible presence of outliers was again investigated, but only for model (M3)

The adjusted model (M3) is given as follows;

\[ g(x) = 2.0370 - 0.0714 \text{Age} + 9.9496 \text{Cough} - 7.5561 \text{LOW} \]  

(R2 = 0.6641, R2 = 0.5750 and R2 = 0.7939)

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

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

(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 4 gives the selected models. Again based on the correlation coefficients, the following model was accepted:

\[ g(x) = -15.1657 + 0.2080 \text{Age} + 13.4477 \text{Cough} + 12.0347 \text{LOA} \]  

(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;

\[ g(x) = -32.1918 + 0.4312 \text{Age} + 18.5982 \text{Cough} + 14.4852 \text{LOA} \]  

(R2 = 0.9450, R2 = 0.7019 and R2 = 0.9719)

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

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

7 Ratio of detection probabilities

Although using the correlations $R^2_1$, $R^2_2$ and $R^2_3$ suggest that (M2), (M4) and (M6) are the best models for detecting TB, 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 $\pi(TB)$ be the probability of detecting TB when using (M2), $\pi(PNEU)$ be the probability of detecting Pneumonia when using (M4) and $\pi(LC)$ be the probability of detecting LC when using (M6).

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

In general, Table 5 indicates the $\pi$-probabilities (in most cases) are very similar, hence the pair-wise comparisons does not help differentiate the diseases.

8 Discussion

The initial study of respiratory diseases usually begins with the use of clinical data, for example, age, cough, LOW and LOA, as indicators for initial screening. 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 additional 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 [11] and was shown to have 85% chance of correct detection (Table 6). 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.

9 Conclusion
The success of initial screening depends heavily on the selection of explanatory variables which may be obtained using the logistic model. Using correlations, this study suggests three possible models for the detection of TB, 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. This remark is further supported by studying ratios of the probability of detection $\pi(TB)$, $\pi(PNEU)$ and $\pi(LC)$. Instead of looking for other explanatory variables, as a viable substitute, the use of the proposed graphical method involving Andrews’ curves is strongly recommended.

Acknowledgements
This research was funded under an E-Science Fund from the Malaysian Ministry of Science, Technology, and Innovation, and Universiti Teknologi Malaysia.

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Original values</th>
<th>New values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Discrete</td>
<td>Unchanged</td>
</tr>
<tr>
<td>Cough on and off</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>more than 2 weeks</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>chronic</td>
<td>4</td>
</tr>
<tr>
<td>LOW / LOA</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: TB = 1, NL = 0, (TB = 40 cases, NL = 20 cases)

Regression Equation (TB@Normal = )

\[
\begin{align*}
R^2_1 & = -1.7405 + 0.0617\text{Age} \\
R^2_2 & = -2.3026 + 10.7118\text{Cough} \\
R^2_3 & = -16.1712 + 13.8686\text{LOW} \\
\end{align*}
\]

\[
\begin{align*}
R^2_1 & = -2.3026 - (1.5855 \times 10^{-14})\text{Cough} + 13.8686\text{LOW} \\
R^2_2 & = -4.1722 + 0.0511\text{Age} + 10.3586\text{Cough} \\
R^2_3 & = -4.1722 + 0.0511\text{Age} + 14.6419\text{LOW} \\
\end{align*}
\]

\[
\begin{align*}
R^2_1 & = -4.1722 + 0.0511\text{Age} + 0.0644\text{Cough} + 14.4630\text{LOW} \\
R^2_2 & = -4.1722 + 0.0511\text{Age} + 14.4630\text{LOW} \\
\end{align*}
\]

Table 3: PNEU = 1, NL = 0 (PNEU = 40 cases, NL = 20 cases)

Regression Equation (Pneumonia@Normal = )

\[
\begin{align*}
R^2_1 & = 1.4044 - 0.0146\text{Age} \\
R^2_2 & = -0.9163 + 9.7527\text{Cough} \\
R^2_3 & = 0.1398 + 9.4262\text{LOW} \\
\end{align*}
\]

\[
\begin{align*}
R^2_1 & = (-3.0130 \times e^{-16}) + 9.5659\text{LOA} \\
R^2_2 & = -1.0498 + 9.7699\text{Cough} + 10.0811\text{LOW} \\
R^2_3 & = -1.0498 + 9.7724\text{Cough} + 9.9024\text{LOA} \\
\end{align*}
\]

\[
\begin{align*}
R^2_1 & = -0.8401 - 0.0015\text{Age} + 9.7198\text{Cough} \\
R^2_2 & = 1.6580 - 0.0322\text{Age} + 9.6576\text{LOW} \\
R^2_3 & = 1.5565 - 0.0329\text{Age} + 9.7657\text{LOA} \\
\end{align*}
\]

\[
\begin{align*}
R^2_1 & = 0.2659 - 0.0273\text{Age} + 9.4119\text{Cough} + 10.9429\text{LOA} \\
\end{align*}
\]
Table 4: LC = 1, NL = 0 (LC = 40 cases, NL = 20 cases)

<table>
<thead>
<tr>
<th>Regression Equation (Lung Cancer@Normal = )</th>
<th>$R_1^2$</th>
<th>$R_2^2$</th>
<th>$R_3^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-$10.9013 + 0.1870$Age</td>
<td>0.4200</td>
<td>0.4142</td>
<td>0.5752</td>
</tr>
<tr>
<td>-$12.040 + 9.9973$Cough</td>
<td>0.6322</td>
<td>0.5528</td>
<td>0.7678</td>
</tr>
<tr>
<td>-$0.5978 + 11.1639$LOW</td>
<td>0.4721</td>
<td>0.4517</td>
<td>0.6273</td>
</tr>
<tr>
<td>-$0.3567 + 10.9227$LOA</td>
<td>0.3968</td>
<td>0.3966</td>
<td>0.5508</td>
</tr>
<tr>
<td>-$1.8971 + 11.4988$Cough + $12.5514$LOW</td>
<td>0.8034</td>
<td>0.6404</td>
<td>0.8894</td>
</tr>
<tr>
<td>-$1.8971 + 11.4654$Cough + $12.6668$LOA</td>
<td>0.8034</td>
<td>0.6404</td>
<td>0.8894</td>
</tr>
<tr>
<td>-$17.7944 + 0.2572$Age + $12.3140$Cough</td>
<td>0.8806</td>
<td>0.6740</td>
<td>0.9362</td>
</tr>
<tr>
<td>-$10.2549 + 0.1572$Age + $11.3096$LOW</td>
<td>0.6608</td>
<td>0.5688</td>
<td>0.7900</td>
</tr>
<tr>
<td>-$9.6610 + 0.1529$Age + $9.8876$LOA</td>
<td>0.6005</td>
<td>0.5344</td>
<td>0.7422</td>
</tr>
<tr>
<td>-$15.1657 + 0.2080$Age + $13.5002$Cough + $11.8732$LOW</td>
<td>0.8314</td>
<td>0.6530</td>
<td>0.9069</td>
</tr>
<tr>
<td>-$15.1657 + 0.2080$Age + $13.4477$Cough + $12.0347$LOA</td>
<td>0.8681</td>
<td>0.6688</td>
<td>0.9289</td>
</tr>
</tbody>
</table>

Table 5: Ratio of detection probabilities

<table>
<thead>
<tr>
<th>TB test sample ($n_1 = 10$)</th>
<th>Pneu test sample ($n_2 = 10$)</th>
<th>LC test sample ($n_3 = 10$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi(Pneu)$ / $\pi(TB)$</td>
<td>$\pi(LC)$ / $\pi(TB)$</td>
<td>$\pi(Pneu)$ / $\pi(LC)$</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>1.0857</td>
<td>1.0857</td>
<td>0.9210</td>
</tr>
<tr>
<td>1.0857</td>
<td>1.0857</td>
<td>0.9531</td>
</tr>
<tr>
<td>1.0857</td>
<td>1.0857</td>
<td>0.1322</td>
</tr>
<tr>
<td>1.0857</td>
<td>1.0857</td>
<td>0.0790</td>
</tr>
<tr>
<td>1.0857</td>
<td>1.0857</td>
<td>0.9210</td>
</tr>
<tr>
<td>12.6651</td>
<td>12.6651</td>
<td>0.9210</td>
</tr>
<tr>
<td>1.0857</td>
<td>0.5439</td>
<td>0.0790</td>
</tr>
<tr>
<td>1.0857</td>
<td>0.9210</td>
<td>0.9210</td>
</tr>
<tr>
<td>1.0857</td>
<td>0.9210</td>
<td>0.9210</td>
</tr>
<tr>
<td>1.0857</td>
<td>0.9210</td>
<td>0.9210</td>
</tr>
</tbody>
</table>

Table 6: Discrimination result.

<table>
<thead>
<tr>
<th>CASES</th>
<th>CORRECT CLASSIFICATION CASES</th>
<th>MISCLASSIFICATION CASES</th>
<th>TOTAL CASES</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL</td>
<td>16</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>PTB</td>
<td>144</td>
<td>24</td>
<td>168</td>
</tr>
<tr>
<td>LC</td>
<td>23</td>
<td>3</td>
<td>26</td>
</tr>
</tbody>
</table>