HIV MODELLING IN A LABOUR FORCE

L.C. Masinga, D.S. Sherwell and C. Myburgh *

Abstract

A section of the workforce employed by a certain company has been found to be HIV positive. Demographic, social and clinical data has been gathered by medical practitioners at the company's Wellness Centre for each patient from the time they were put in a wellness programme to monitor their health and fitness for placement in various work categories with different levels of labout intensity. The medical staff would like to know how the collected data can be used to determine critical indicators that a given patient is fit to go back to work in the same job, and for how long he can stay on that job. CART analysis software is used to investigate the problem. Among the variables in the data, mass and viral load are found to rank high in importance as deteminants for making a decision on the fitness of the participants.

1 Introduction

The Problem presented to the Study Group concerned management of a population of workers (henceforth to be referred to as participants) who have been determined to be HIV positive, and were subsequently participating in a wellness programme at the organisation's Health Centre. Under the Wellness Programme the participants were required to visit the Health Centre regularly for assessment by medical practitioners.

The study group had access to a database of the participants, only identifiable by an index number. The data collected for each patient were of demographic, clinical and physical characteristics measured at different points in time by medical staff, over a period of up to five years, depending on when the participants were introduced to the wellness programme. Within the organisation the participants were assigned to different job categories

*School of Computational and Applied Mathematics, University of Witwatersrand, Private Bag 3, Wits 2050, South Africa. *emails: Londiwe.Masinga@cam.wits.ac.za, David.Sherwell@wits.ac.za, and colin@data.co.za*

of varying levels of labour intensity, according to their skills and demand. Depending on the condition of their health it was sometimes necessary to reassign the workers to other work categories, suitable for their health condition at the time.

The objective of this study was to analyse available participants' demographic and clinical information in order to determine indicators that could be used to predict whether or not a particular worker, currently participating in the wellness programme, will return to work. Further to this, to investigate whether the participant will be fit to return to the same job category and how long he can stay in that job. The above constituted two questions that needed answering.

The study group proposed the use of predictive modelling tools to answer some aspects of the questions posed. Classification and Regression Tree (CART) Analysis was identified as a possible technique to answer the question of what indicators would predict whether participants would return to work. This is a non-parametric decision-tree tool that has been widely used in the development of clinical decision rules. CART analysis can select from among a large number of explanatory variables those and their interactions that are most important in determining the outcome (response or target) variable to be explained. CART modelling software packages have the advantage of being user friendly for modelling without understanding the statistical details behind the tools, and they are readily available on the internet. Alongside the CART analysis model, the Logistic Regression Analysis model was proposed to answer the second question of how long a participant can be expected to return and stay in the same job. Another proposed approach was the use of life tables to compute the probabilities that participants would survive different time periods into the future. The life tables model requires the development of custom software to automate the laborious process.

The results obtained from some experiments on the available data are encouraging and seem to reveal the potential of using CART analysis software packages in answering the questions at hand.

2 Baseline Data

At the time of compiling, the database consisted of nearly 1500 records. The data recorded at the Wellness Centre is quite comprehensive. For each participant the following information variables, a mix of categorical, nominal and interval type, were recorded:

1. Residential site and type (RES) - The organisation runs business in several locations and houses the employees in different types of accommodation. This is a once off record.

2. Silicosis (SIL) and previous TB status (TB). This is a once off record.

3. Whether the patients have left (LFT) or are still with the Wellness Programme. This was the record of the company at the time of the compilation of the data for analysis. This is a once off record.

4. The WHO stage of the patients (WHO) at the time of presentation at the clinic. This is a once off record.

5. Record of the patients' mass (MAS) on visit dates. Since the employees were participating in the Wellness Programme they had regulator visits to the Wellness Centre for check ups and monitoring. This was measured regularly.

6. Record of $CD4^+$ count (CD4) and viral load (VIR) on visit dates . This was measured regularly.

7. The date when the patient started anti-retroviral treatment (ART). This would depend on the $CD4^+$ count and viral load. This is a once off record.

8. The fitness index of the patient on various dates. The participants were subjected to specific physical fitness tests. These are standard tests and were used, among other things, to determine whether the employees were still fit for the kind of job they were doing, for redeployment purposes if necessary. This measurement may change from time to time.

9. The job category of the patients at various dates. These would change from time to time, depending on their fitness test results.

3 Methodology

The available data was first cleaned and reorganised. The purpose of cleaning was to have a dataset with little missing data. After cleaning the dataset had about 500 cases.

CART Software

A trial version of The Salford Systems CART software[1] was downloaded from the internet and used to investigate its application to this study. The popularity of the CART procedures is based on the simplicity of basing them on 'Yes' and 'No' answers to explanatory variables.

Tree Growing

Starting with the whole dataset (this could be learning data as is recommended in use of the techniques), the procedure involves searches of the whole set using specific, dichotomous criteria, associated with one input variable at each stage. Based on a criterion a split occurs according to whether the criterion is satisfied or not, to create two branches. In this way a parent node splits into two child nodes, which in turn become parent nodes for further splits, and the tree growth process continues on. The ability to define stopping rules is normally built into the CART software. These rules may lead to terminal nodes, which are mutually exclusive subgroups of the population that cannot be split any further. One such rule is based on a measure of the purity of the node.

Different trees can be created by selecting different combinations of input and output variables. Thus a given dataset cannot be used to grow a unique decision tree. Also, different interpretations can be obtained for different trees, hence answers to different investigative questions can be obtained. The analysis concludes by selecting an optimal tree, a step that applies cost measurement techniques. Such a choice may be a result of other processes including measurement of misclassification cost and tree pruning.

CART Output

The default output of the CART analysis procedures is displayed in the form of binary trees which indicate at each node: the splitting variable, the splitting criteria, number of cases that satisfied the specified criteria, as well as whether a particular node is a terminal node. The details of what is displayed may vary from one software package to another. Various forms of output on other aspects of the analysis preferred by the modeller can be obtained at the click of a button on the display screen. Among these are summary reports, splitting rules, hard copies, variable importance analysis, and others. Of particular interest would be the level of variable importance. This is a ranking that highlights any outstandingly important variables -

[1]CART is a registered trademark of California Statistical Software, Inc. and is exclusively licensed to Salford Systems

those that do most of the work in separating favourable from unfavourable records.

Several trees were generated using the available data. The response variable chosen to answer the first question was the job category. Some highlights of various degree of association between the variables were obtained. The information from these was used as input for the Logistic Regression model to model the second question of the project.

4 Logistic Regression Analysis

Logistic Regression Analysis (LRA) is a model-based analysis, largely used to estimate the 'average' effect of an independent variable on the probability of a success for the dependent variable, while accounting for other factors [1]. The dependent or response variable is normally assumed to be dichotomous, taking on values of 1 (for success or true) or 0 (for failure or false), with $P(Y = 1) = p$, and dependent on explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. LRA is based on a linear model for the natural logarithm of the odds in favour of $Y = 1$ [2]:

$$\ln\left[\frac{P(Y=1|\boldsymbol{x})}{1-P(Y=1|\boldsymbol{x})}\right] = \ln\left[\frac{p}{1-p}\right],$$

also written as

$$\begin{aligned}\text{logit}\{\Pr(Y=1|x_1,x_2,\ldots,x_n)\} &= \log\left\{\frac{\Pr(Y=1|\boldsymbol{x})}{1-\Pr(Y=1|\boldsymbol{x})}\right\}\\ &= \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n\\ &= \beta_0+\boldsymbol{x}'\boldsymbol{\beta}\end{aligned}$$

where β_0 is the intercept parameter and β_i are the weights of the change effect of the predictor variables. These parameters, also known as the logistic regression coefficients, must be estimated from the available data, using techniques such as the maximum likelihood principle. Fortunately, regression is incorporated in standard Statistical packages, and so the estimation is automatically done. Once the estimates $\hat{\beta}_i$, $i = 0, 1, \ldots, n$ have been found, a transformation of the above equations leads to

$$\frac{p}{1-p} = e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ldots+\hat{\beta}_nx_n}.$$

Finally,

$$p = P(Y=1|\boldsymbol{x}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ldots+\hat{\beta}_nx_n}}{1+e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ldots+\hat{\beta}_nx_n}}.$$

Various combinations of the predictor variables can be used, together with certain statistical procedures to measure the predictive effeciency of the model.

CART software does not require the user to know the above satistical details, but only to understand how to apply the tool and interpret the results.

LRA was proposed to answer the question of whether or not an individual will be in the same job category in k months time, given certain clinical and demographic data. In our case the response variable is 'same job category in k months time'. It will be true ($Y = 1$) with probability p. We wanted to find the probability that, $\Pr(Y = 1|\boldsymbol{x})$ for a given number k of months, where $\boldsymbol{x}$ is the clinical and demographic data.

5 Results

Given in Figure 1 below is part of a typical tree obtained from the Salford Systems' CART software on the cleaned dataset (some software may only handle a limited number of records in a database at a time). A description and interpretation of the tree is as follows:

Node 1 is the first parent node. At this node 472 records were searched for a criterion associated with the variable mass (MAS). A split created two branches, leading to two child nodes, Node 2 (containing 298 cases whose mass was less than a specified critical value) and Node 6 (containing 174 cases whose mass was greater than or equal to a specified crtical value). For the set with cases whose mass was less, the viral load was the variable used for the next split. Out of 298 cases, 203 had a viral load less than a specific level. The other 95 had a high viral load and the branch ended in a terminal node (Node 5). The next critical variable for the 203 cases whose viral load was not high was age. Of these 168 did not meet the weight criterion and for these the CD4 count (Node 4 - not shown on the diagram) was critical. There was one further split after this one on mass again. After this all the nodes on this branch were terminal (e.g. Nodes 5 and 8 on the diagram).

Following the section of the tree with the longest path to a terminal node, the sequence of predictor variables are mass - viral load - age - CD4 - mass - terminal. Note that this is that branch of the tree with cases that do not meet the "greater-than-or-equal-to" criterion for each critical value of the split variables.

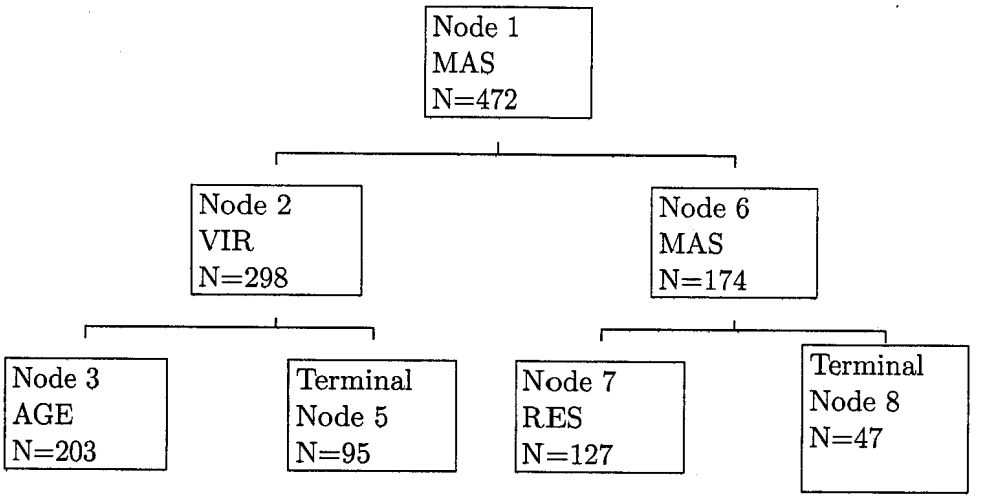

Figure 1: A typical CART tree

Typical of CART analysis software, The Salford Systems CART analysis software enables the user to look into the criteria and other determinant factors used in the splitting at each point. Moreover, it can also rank the variables in the order of their importance in influencing the classification. Shown below is a table of variable importance associated with the above tree.

Variable	Variable Importance
MAS	100.00
VIR	78.29
RES	37.47
WHO	19.16
$CD4^+$	12.97
AGE	11.42
SIL	4.2
TB	0.00

Table 1 Table of variable importance associated with the CART tree in Figure 1.

It can be deduced from Table 1 that mass is the most discriminating factor, followed closely by viral load, and residence. It is interesting to note that the type of residence (RES) is of higher importance than, say $CD4^+$ count, and that the history of TB is insignificant.

Before carrying out the Logistic Regression analysis, a simple correlation of the variables was performed. This led to some interesting findings:

- Mass was very significantly correlated with residing with one's spouse.
- Mass is very significantly correlated with $CD4^+$ level.
- TB is significantly correlated with WHO stage three.
- $CD4^+$ level is significantly correlated with residing with one's family/friends.

Using different values of k yielded the following results:
In single-person residence, $CD4^+$ and mass were all found to be significant predictors of being in the same job category after six months. For a twelve month forecast, $CD4^+$ count was highly significant while single-person residence and mass were marginally significant. For eighteen months only viral load turned up to be significant, and marginally so for twenty four months.

6 Conclusion and recommendations

This investigation was a quick statistical assessment of the 'return to labour category' problem. More thorough work is clearly needed, but the preliminary outcomes are as follows.

The CART analysis suggests that mass and viral load are the top two critical indicators for fitness to return to work. The regression model gives unsatisfactory results because of the high level of association between the explanatory variables. This requires taking fewer variables at a time and performing the regression on them.

Some results suggest that further application of the model would lead to interesting outcomes from the study which could be used for profiling and performing such management tasks as risk management.

It is recommended that if the investigation and analysis is to be pursued, the above software be purchased and used as a tool to assist in monitoring participants in the programme. However, in order to be able to use the data efficiently, reformating of the data is necessary. A lot of the cases were not included in the analysis because of un-synchronised data due to the fact that not all tests were performed during all visits to the Wellness Centre. This appears as missing data and is hence discarded. A strategy needs to be devised so that once a participant has been identified as needing to join

the programme, the data collected about the patient should be as complete as possible and should be recorded regularly.

Acknowledgements

We would like to thank the medical practioners and the company that employs them, whose names will be kept anonomous for confidentiality, for presenting the problem to the Study Group.

References

[1] Lemon, S.C., Roy, J., Clark, M.A., Friedman, P.D. and Rakowski, W. Classification and Regression Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. *Annals of Behavioral Medicine*, **26** (2003), 172–181 .

[2] http://online.sfsu.edu/~ efc/classes/biol710/logistic/logisticreg.htm

[3] Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. Classification and Regression Trees. Chapman and Hall, Wadsworth, Belmont, California (1984).