International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Special Issue 14 (Jan-Feb 2015), PP. 03-05

DEVELOPING POST-OPERATIVE SURVIVAL PREDICTION MODELS FOR LUNG CANCER PATIENTS USING RECURSIVE PARTITIONING TECHNIQUES

Corey Stone
School of Business. Bond University, Gold Coast,
Queensland, Australia.

Abstract- The purpose of this report is to analyse the factors that may lead to a greater understanding of the post-operative survival rate in lung cancer patients who undergo lung resection surgery. This paper will explore the relationships between data collected prior to the operation, and whether the patient is still alive 12 months after the operation. Cutting edge Recursive Partitioning techniques such as Discriminant Analysis, Decision Tree Method, Random Forest, and Artificial Neural Networks will form the basis of the methodologies used for building a predictive model. A random sample of 224 patients who underwent surgery between the years of 2007 and 2011 in Wrawclaw, Poland will form the data set. Of the 224 sample of patients, 63 did not survive more than 12 months after the operation; the remaining 161 survived.

Artificial Neural Network and Random Forest intelligence methods are found to achieve the most accurate classification rates. These models can be used by medical professionals to provide an accurate survival outcome for lung cancer patients who elect to undergo lung resection surgery.

I. INTRODUCTION

Lung cancer is defined as the uncontrolled growth of malignant cells that start in one or both lungs, typically in the cells that line the air passages[1]. Classification of the cancer is dependent on which type of cell is affected: non-small cell lung cancer (NSCLC), small cell lung cancer, or mesothelioma, and is most commonly diagnosed with various screening tests.

Lung cancer is the second most commonly occurring cancer for both males and females, second only to prostate and breast cancers, and accounts for 1.5 million deaths each year [2]. It is estimated that approximately 85% of all cases are related to cigarette smoke inhalation [3].

As is the case with most cancers, the overall survival rate is significantly dependent on the stage of the tumour when diagnosed, with the highest terminal rate occurring when the tumour has become unresectable. Conversely, NSCLC patients have a significantly better prognosis with a 30% - 60% survival at five years [4].

Treatment for lung cancer is usually a joint decision made between the doctor and the patient, depending on a number of factors such as the patients overall health, the type and stage of the cancer, and the patients personal preferences. Treatment options typically include:

- **Surgery**- Resection of the carcinogenic tissue and a margin of healthy tissue
- **Chemotherapy**- Intravenous administration of drugs targeted to kill the cancer
- **Radiation Therapy**- High-powered energy beams, such as X-rays targeted at the cancer
- **Targeted drug therapy**- Targeting specific abnormalities in the cancerous cell

(Berge, 2014)

This paper will examine a cohort of 224 patients who elected to undergo surgical treatment for the removal of their lung cancer. A series of measurements (nominal and numerical) were collected prior to the operation and will be tested against the survival outcome (at 12 months) for statistical significance. Variables that do have a significant relationship with the survival outcome will be used to create a model for predicting the post-operative survival outcome of any given patient.

The benefits for being able to predict the post-operation survival outcome for a patient are as follows:

- Doctors will be able to provide any specific patient with an accurate survival prediction based on previous results
- Patients will be better emotionally prepared as they will have greater certainty surrounding their future
- Patients and Doctors can make more informed decisions on what course of treatment to follow depending on the likely outcome of an operation

II. METHODOLOGIES

There are two main data mining methods; unsupervised and supervised. The unsupervised method has no target variable and the algorithm therefor looks for patterns among all variables. Supervised models, conversely, have a pre-specified target and the algorithm finds the association between this target and the predictor variables.

Regression analysis is an example of a supervised model as it is concerned with measuring the relationship between a response (dependent) variable and one or more explanatory (independent) variables. A regression model is a type of probabilistic classification model that can assist in predicting or explaining a dependent variable as it relates to independent variables. It also provides an adequacy of the model by accessing 'goodness of fit' as measured by the 'r squared' (R^2).

Logistic regression has been widely used in the medical science field as a method for predicting various outcomes for patients depending on collected variables. Regression is of particular use to the medical field as it allows the analysis of dichotomous or binary outcomes with the ability to adjust for multiple predictors. This research uses logistic regression modelling to identify significant variables and uses them to create a predictive model.

Artificial Neural Networks (ANN) are a form of artificial intelligence (AI) that can be traced back to early developments in 1943 by McCulloch and Pitts. ANN's have become the

preferred tool for many predictive data mining applications due to their power, flexibility and ease of use. The term neural network applies to a family of algorithms, characterised by a large parameter space and flexibility structure, which attempts to replicate the neural functions of the human brain.

According to Haykin, (1998) a neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use and resembles the brain in two respects:

- Knowledge is acquired by the network through a learning process
- Interneuron connection strengths known as synaptic weights are used to store the knowledge [5]

ANN's are organised in layers within the network, called the network architecture. The first layer of input neurons usually contains predictor variables, and the last layer contains the responses.

Each of the nodes in the middle hidden layer contain a function termed '*activation function'* which allows the network to model complex linear relationships between the input and output layers. After the network has trained itself, the knowledge gained is used to connect each input node to each hidden node by a connection weight. A weighted linear combination is then summed for each hidden node and then passed through an activation function: logistic or sigmoid function. The activation of the hidden node is then multiplied by a second set of connection weights and added to a bias weight. Finally, a logistic transformation of the weighted inputs is applied to determine the overall output of the network [6].

In this study, Multilayer perceptron (MLP) is used with all independent variables fed as 'standardised' covariates. The MLP network uses back-propagation to adjust the weights on each connection by assuming all of the connections are equally to blame for errors in the classification of data.

Random forests are a classification-based ensemble learning method that uses a multitude of decision trees in training to output an individual tree. When a random forest is given the task of classifying a new object from an input vector, the input vector is passed through each of the trees in the forest. Each tree that the vector passes through gives a classification and forms a voting system through the forest. The final output classification is a result of the majority vote by individual classification trees. Each individual tree is grown similarly to traditional decision tree classification; however each tree is grown to the largest extent possible without any pruning.

III. RESULTS

Binary logistic regression has been performed on the following 16 independent variables that were collected from each patient prior to their operation: Specific cancer Diagnosis, Forced vital capacity, FEV1, Zubrod scale, Pain before surgery, Haemoptysis before surgery, Dyspnoea before surgery, Cough before surger, Weakness before surgery, Size of the original tumour, Diabetes mellitus, MI up to 6 months, Peripheral arterial diseases, Smoking, Asthma, Age at surgery. Of the sixteen variables, four had a statistically significant relationship with the post-operative survival outcome in lung cancer patients at 12 months (12% significance): Original Diagnosis, Size of original tumour, Diabetes Mellitus, Smoking status.

The logistic regression model achieved a sensitivity rate of 95.1% by accurately classifying 154 out of 164 survivors, and a specificity rate of 71.0% by accurately classifying 44 out of 62 non-survivors. An overall classification rate of 88.40% was achieved. The strongest predictor of the post-operative survival rate was the original diagnosis, with an odds ratio of 15.82. This indicates that patients diagnosed with metastatic lung cancer are 15.1 times more likely to survive 12 months after the operation than those diagnosed with primary or secondary lung cancer.

The derived logistic regression equation for predicting the survival outcome of any given patient can therefore be stated as:

$$Z = -36.624 + 2.83\,(\text{Diagnosis}) + 2.268\,(\text{Size of the tumour}) + 1.283\,(\text{Diabetes Mellitus}) + 1.517\,(\text{Smoking status})$$

Where a Z score less than .50 indicates that the patient will survive and a score greater than .50 indicates that the patient will not survive 12 months after the operation.

Artificial neural network modelling achieved a higher overall classification with 90% accuracy. The holdout model achieved 98.2% sensitivity by correctly classifying 55 out of 56 patients who survived and 71.4% sensitivity by classifying 15 out of 21 patients who did not survive. The most important variables in predicting the post-operative survival of a patient, as illustrated in a normalized importance chart, were the size of the tumour and the diagnosis. The existence of a cough before surgery was considered the least important of all independent variables.

A Decision Tree using random forest ensemble learning method was built with Classification and Regression Trees (CART) as its growing method. All independent variables were specified for CART, however only two were remaining after pruning was performed. The significant independent variables identified through the random forest technique were the size of the original tumour and the original diagnosis. A tree diagram was prepared to provide a graphic representation of the algorithm generated for classification of the dependent variable. The first decision node specifies that if the original size of the tumour is less than 11.5, there is a 90.4% probability that the patient will survive and a 9.6% probability that the patient will not-survive. If the size of the original tumour is greater than 11.5, the original diagnosis needs to be considered. Where the tumour is greater than 11.50 and the patient was diagnosed with primary or secondary lung cancer (Diagnosis 3 or 4), there is a 100% chance that the patient will not survive. Where the size of the tumour is greater than 11.50 and the diagnosis was metastatic lung cancer (Diagnosis 2), the patient has a 75% chance of surviving and a 25% chance of not surviving.

This classification tree achieved an 89.3% overall classification rate on the full data set.

IV. DISCUSSION AND CONCLUSION

Research relating to survival forecasting for lung cancer patients up until now has mainly focussed on a small number of independent variables (size of tumour and age) or variables that require invasive methods to obtain (genetic activity within cancerous tissue). This study has investigated the relationship between 16 independent variables, all of which are obtained

non-invasively, prior to the patients operation, and the 12 month survival outcome for the patient. Various statistical modelling techniques with underlying assumptions were used to analyse the data set and compared to prior research. In this study, the cutting edge recursive partitioning methods delivered superior modelling results to the traditional univariate and multivariate analytical models. The present findings are restricted to patients with a primary lung cancer, secondary lung cancer or metastatic lung cancer diagnosis that elect to undergo a lung resection as treatment.

Logistic regression, artificial neural networks, and random forest intelligence systems confirm that the most significant variables in predicting the survival outcome for a patient are: the original diagnosis, the size of the tumour, the existence of diabetes mellitus, and the patients smoking status. Medical professionals can use this information by feeding collected data into an algorithm and being able to provide an accurate survival prediction at 12 months for any given patient.

Significant changes could also result in the treatment method selected for lung cancer patients. The random forest intelligence system identified that patients within the examined sample who were diagnosed with a primary or secondary lung cancer with tumour size greater than 11.5 all deceased within 12 months of their operation. This suggests that for future patients with these characteristics, alternative forms of treatment (chemotherapy, targeted drug therapy, radiation therapy) are likely to lead to more optimal outcomes.

V. FUTURE WORK

Significant learnings could be achieved through the collection and analysis of data at hospitals from other countries around the world to ensure that findings discussed in this report are not specific to a single country.

Predictive accuracy for the survival of lung cancer patients could also be improved through collecting non-intrusive variables for patients who elect to undergo alternative forms of treatment; chemotherapy, radiation therapy, and targeted drug therapy.

Further improvements to this research include:

The development of a user friendly model whereby a medical professional can type in collected variables taken from a patient and provide a post-operative survival prediction.

The removal of variables that bear no significance on the patient outcome to be replaced with others such as; body mass index and pack years etc.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Professor Kuldeep Kumar for his continuous support, patience, and guidance in supervising my study and research. I could not have imagined a more knowledgeable and enthusiastic advisor to oversee the writing of this paper.

REFERENCES

[1] Cancer Council NSW. (2014, October 20). *Lung Cancer*. Retrieved from Councer Council Australia: www.cancercouncil.com.au/lung-cancer

[2] American Cancer Society. (2014). *What are the key statistics about lung cancer?* Retrieved from Cancer Org: www.cancer.org/cancer/lungcancer

[3] Lung cancer; new lung cancer study findings have been reported from A.K. ganti et al. (2012). *Obesity, Fitness & Wellness Week,*, 1234. Retrieved from http://search.proquest.com/docview/929276778?accountid=26503

[4] ountain, C. (1997). Revision in the International System for Staging Lung Cancer. *PubMed*, 1710-1717.

[5] Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation.* Upper Saddle River: Prentice Hall PTR.

[6] Tu, J. V. (1996). Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. *Elsevier Science*, 1225-1231.