# CHAPTER 10

# SURVIVAL MODEL AND ESTIMATION FOR LUNG CANCER PATIENTS

Xingchen Yuan[a], Don Hong[b], and Yu Shyr[c,d]

[a]*Fermilab, Batavia, IL 60510, USA*
[b]*Department of Mathematical Sciences, Middle Tennessee State University,*
*P.O. Box 34, Murfreesboro, Tennessee 37132, USA*
*E-mail: dhong@mtsu.edu*
[c]*Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA*
[d]*Department of Statistics, National Cheng Kung University, Taiwan, ROC*

Lung cancer is the most frequently occuring fatal cancer in the United States. By assuming a form for the hazard function for a group of lung cancer patients for survival study, the covariates in the hazard function are estimated by the maximum likelihood estimation following the proportional hazards regression analysis. Although the proportional hazards model does not give an explicit baseline hazard function, the function can be estimated by fitting the data with non-linear least square technique. The survival model is then examined by a neural network simulation. The neural network learns the survival pattern from available hospital data and gives survival prediction for random covariate combinations. The simulation results support the covariate estimation in the survival model.

## 1. Introduction

Cancer develops when cells in a part of the body begin to grow out of control. It is the second most significant reason for US mortality. In 2001, cancer caused 553,768 deaths in the United States, accounting for 22.9% of all deaths in that year [13]. In the past fifty years, efforts have been made to reduce death rates for different diseases, but the death rate for cancer remains almost unchanged ([14], [15]). Among the various types of cancers, lung cancer is the most frequently occuring fatal cancer, for both men and women, in the United States. Each year there are about 170, 000 new cases of lung cancer in the U.S.A. and 150,000 deaths attributable to this disease. Men are affected somewhat more frequently (100,000 cases/year) than women (70,000 cases/year). Worldwide, there are 1 million new cases per year. Over the past 5 decades the number of yearly cases has increased, and the worldwide incidence may double to 2 million per year in the coming decade. The average patient is 60 years old, and only 1% of cases are under 40 years old. About 90% of patients have historically died from their disease.

Recently, there has been a great deal of interest in modeling survival data of cancer patients (see [2], [8], [12] for example). `Survival analysis` is concerned with studying the time between entry to a study and a subsequent event, such as death. In practice, after a lung cancer patient is hospitalized, a set of medical data regarding the patients' condition is recorded. This data set may include information such as the patient's survival time, the tumor's stage, the health grade, the disease free time, etc. With the data set, we wish to study how the patient's conditions might be associated with the survival pattern and also a lung cancer patient's survival chance, or a group of patients' survival distribution over time.

The goal of this study is to develop a survival model for relating the hospital data profile to censored survival data such as time to cancer death or recurrence. Censored survival times occur if the event of interest, i.e., the death, does not occur for a patient during the study period. Traditionally, there are two approaches to model the unknown survival distribution. One is to assume a classical parametric model such as normal, lognormal, gamma, Weibull, Pareto or beta, then use a histogram, kernel or other nonparametric estimate of the unknown density function. This method is straightforward but cannot reflect the contribution of patients' hospital conditions to the survival distribution. Another is the `proportional hazards model`, which was first proposed by D.R. Cox [1] in 1972 to investigate the effects of covariates on survival patterns, also known as Cox regression model [7]. The model permits having the patients' hospital conditions as a vector of covariates in the hazard function and can estimate the unknown parameters for the covariates by partial likelihood without putting a structure of baseline hazard. In this study, however, we propose a structure of the baseline hazard function, and estimate the parameters by the available censored survival data so that the explicit survival function is determined. This estimation is achieved by a least square fit for the cum hazard value computed by SPSS.

In a survey study, the design parameters for the survey are sometimes related to the hazard function but do not fit in the model. On some other occasions, the independence assumption of the covariates may be violated.

Sometimes correlations exist within each level of nesting. These could cause biases and affect variances of parameter estimation [10, 11]. Therefore, tests need to be done to evaluate the goodness of the estimated survival function. There are two popular ways to test the model. One is to use 1/2 or 2/3 of the time scale in the survival data to determine the parameters, and then use the whole data set to examine the model; another is to use the whole data set to set up the model, then using resample methods to check the model. Neural networks are increasingly being seen as an addition to the statistics toolkit which should be considered alongside both classical and modern statistical methods. It has been pointed out in [16] many different ways that classification networks have been used for survival data. In this study, due to the lack of patient data, we propose a neural network model to simulate the patients' survival pattern and use the neural network to generate a long list of "virtual data" to test the survival model.

The remainder of the paper is organized as follows: In Section 2, we give a description for the survival model. We first introduce the conception of hazard function and survival function as well as their relationship. We then outline the method of proportional hazard model and propose and justify the exponential form for baseline hazard function. In Section 3, we discuss the parameter estimation by statistical methods including maximum likelihood estimation (MLE) and non-linear least square estimation (LSE). We also introduce the idea and conception of the neural network and set up the proper neural network by MATLAB programs for testing. In Section 4, we present the computational result with actual patient data. Discussions and conclusions are given in Section 5.

## 2. Description of Model

### 2.1. *Survival Function and Hazard Function*

Following the notations in Actuarial Mathematics [4], we let $T$ be a nonnegative random variable representing the failure time of an individual in the population. Assume $T$ is distributed with the probability density function (pdf) $f(t)$, then the cumulative distribution function (cdf) is

$$F(t) = Pr[T \leq t] = \int_0^t f(z)dz \tag{2.1}$$

giving the probability that the event has duration $t$. The `survival function`, $S(t)$, is defined as the complement of the c.d.f., that is

$$S(t) = Pr[T < t] = 1 - F(t) = \int_t^\infty f(z)dz. \tag{2.2}$$

The survival function gives the probability of being alive at duration $t$. Naturally, when $t = 0$, $S(t) = 1$ and $t \to \infty$, $S(t) \to 0$.

An alternative characterization of the distribution of $T$ is given by the `hazard function`. Sometimes it is also called the `force of mortality`, the mortality intensity function, or the failure rate. The hazard function is the probability that an individual will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the instantaneous risk of occurrence of dying at time $t$. The hazard function $h(t)$ can be estimated using the following equation:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr[t < T \leq t + \Delta t | T > t]}{\Delta t}. \tag{2.3}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $(t, t + \Delta t)$ given that it has not occurred before, and the denominator is the width of the interval. We obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval decreases to zero, we obtain an instantaneous rate of occurrence.

The conditional probability in the numerator may be written as the ratio of the joint probability that $T$ is in the interval $(t, t + \Delta t)$ and $T > t$ (which is, of course, the same as the probability that $t$ is in the interval), to the probability of the condition $T > t$. The former may be written as $f(t)\Delta t$ for small $\Delta t$, while the latter is $S(t)$ by definition. Dividing by $dt$ and passing to the limit gives the useful result

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = \frac{(1 - S(t))'}{S(t)} = -\frac{S'(t)}{S(t)}. \tag{2.4}$$

This equation suggests the relationship between the survival function and the hazard function. That is, the rate of occurrence of the event at duration $t$ equals the density of events at $t$ divided by the probability of surviving to that duration without experiencing the event. Furthermore, equation (2.4) suggests that

$$h(t) = -\frac{d}{dt} \log S(t), \tag{2.5}$$

then

$$\log S(t) = -\int_0^t h(z)\, dz + C. \tag{2.6}$$

Considering the boundary condition $S(0) = 1$ as we mentioned before, we have $C = 0$, and thus

$$S(t) = \exp\{-\int_0^t h(z)\, dz\}. \tag{2.7}$$

Combining (2.7) with (2.4), we obtain

$$f(t) = h(t)S(t) = h(t)\exp\{-\int_0^t h(z)\, dz\}. \tag{2.8}$$

A recent survey on dynamic mortality modeling in actuarial mathematics is given in [17].

### 2.2. *Cox Regression*

A `Cox model` is a well-recognized statistical technique for exploring the relationship between the survival of patient and a set of explanatory variables (see [1], [16] for example). We call these explanatory variables `covariates`.

Suppose that we have collected $n$ patients with lung cancer. For the $i$th patients, let $(t_i; \delta_i)$ be the observed phenotype, where $t_i$ is the failure time (in other words, when death occurs) when $\delta_i = 1$, and is the censoring time (e.g., last time known of a patient being cancer-free) when $\delta_i = 0$. Let $x_i = (x_{i1}, \cdots, x_{ip})$ be the vector of $p$ covariates for the $i$th sample taken from the $i$th patient. We assume that a general Cox model with the hazard function for the $i$th patient is modeled as

$$h(t|x_i) = h_0(t)\exp(f(x_i)), \tag{2.9}$$

where $h_0(t)$ is called the `baseline hazard function`. Although $f(x_i)$ may assume many formats, the most popular and also the simplest model for $f(x)$ is

$$f(x_i) = x_i \cdot \beta = x_{i_1}\beta_1 + \cdots + x_{i_p}\beta_p, \tag{2.10}$$

where $\beta$ is a column vector of coefficients. In this equation, it is assumed that the effects of the different covariates on survival are constant over time and are addictive in a particular scale. The Cox model makes no assumptions about the form of $h_0(t)$, but assumes the parametric form for the effect of the covariates (predictors) on the hazard. In this sense, the Cox model is a semi-parametric model.

The vector $\beta$ of parameters can be estimated by the partial likelihood method. Let the observed follow up time of the $i$th individual be $t_i$ with corresponding covariates $x_i$, $i = 1, .., n$. The conditional probability for the

$i$th individual failing at $t_i$ given that the individual is from the risk set $R(t_i)$ (i.e., $R(t_i) = \{j \,|t_j \geq t_i\}$) is [10]:

$$\frac{h_0(t)\exp(x_i\beta)}{\sum_{\ell\in R(t_i)} h_0(t_i)\exp(x_\ell\beta)}. \tag{2.11}$$

Assuming that there are $K$ failures. The partial likelihood function is then:

$$\prod_{i=1}^{K} \frac{\exp(x_i\beta)}{\sum_{\ell\in R(t_i)} h_0(t_i)\exp(x_\ell\beta)}. \tag{2.12}$$

Recalling the definition of $\delta_i$ at the beginning of this section, the partial likelihood function can be expressed as:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(x_i\beta)}{\sum_{j=1}^{n} y_j(t)\exp(x_j\beta)}\right]^{\delta_i}, \tag{2.13}$$

where $y_j(t) = 0$ when $t \leq t_j$, otherwise $y_j(t) = 1$. Equation (2.13) can be written in another way to remove the expression of $\delta_i$:

$$L(\beta) = \prod_{i\ uncensored} \left[\frac{\exp(x_i\beta)}{\sum_{j=1}^{n} y_j(t)\exp(x_j\beta)}\right]. \tag{2.14}$$

For a sample of size $n$, the log partial likelihood for expression (2.14) is

$$l(\beta) = \log L(\beta) = \prod_{i\ uncensored} \left\{x_i\beta - \log\left[\sum_{j=1}^{n} y_j(t)\exp(x_j\beta)\right]\right\}. \tag{2.15}$$

The maximum partial likelihood estimation of $\beta$ can be obtained as a solution to the equation

$$\frac{\partial l(\beta)}{\partial\beta} = 0,$$

and thus,

$$\sum_{i\ uncensored}^{n} x_i - \frac{\sum_{j=1}^{n} y_j(t)x_j\exp(x_j\beta)}{\sum_{j=1}^{n} y_j(t)\exp(x_j\beta)} = 0. \tag{2.16}$$

Cox and others have shown that this partial log-likelihood can be treated as an ordinary log-likelihood to derive valid (partial) MLE of $\beta$. Therefore, we can estimate hazard ratios and confidence intervals using maximum likelihood techniques whose principal will be discussed in the next section. To avoid the baseline hazard, estimates are based on the partial as opposed to the full likelihood.

Usually, the Cox proportional hazard regression model is a very useful tool to estimate the coefficients in a linear combination of covariates in survival analysis since both SAS PHREG procedure and SPSS Survival Package perform regression analysis of the survival data based on the proportional hazards model. However, because of the nature of proportional hazard regression, neither software packages give an explicit function expression for the baseline hazard function $h_0(t)$. In the next section, we will justify an explicit function of the baseline hazard function $h_0(t)$ and also estimate the parameters in $h_0(t)$ using non-linear least square technique based on the result obtained from the Cox regression for the survival function fitting the data set of lung cancer patients.

### 2.3. *Baseline Hazard for Lung Cancer Patients*

Like any cancer, the exact reason why one particular person is diagnosed lung cancer and another does not remains unknown. However, certain factors are strongly correlated with an increase in lung cancer, when groups of patients are studied. By rank, these factors are listed below [13]:

(i) Tobacco Smoking or exposure to smoke
(ii) Carcinogen Exposures
(iii) Radiation Exposure
(iv) Miscellaneous Risks Factors, including old scars in the lungs.

The first three factors involve an interaction between the individual and the environment. Presumably an individual is continuously exposed to and absorbs certain levels of smoke, radiation, or some kind of toxic material (like carcinogen) which then lead to lung cancer. Though a portion of the absorbed toxic materials is discharged from the body, the cumulative effect of retained toxins contributes to the individual's death [6].

For every given $\tau$ in $[0, t]$ and the infinitesimal time element $[\tau, \tau + d\tau]$, let the sum $\delta d\tau + o(d\tau)$ be the probability that a unit of toxic material is absorbed during $[\tau, \tau + d\tau]$ and the sum $\nu d\tau + o(d\tau)$ be the probability that a unit of toxic material in the body is discharged during $[\tau, \tau + d\tau]$. Assuming that $\delta$ and $\nu$ are independent of time, then the probability that an individual will absorb a unit of toxic material during $[\tau, \tau + d\tau]$ and will retain it in his/her body up to time $t$ is given by [6]

$$\delta d\tau \, \exp\{-(t-\tau)\nu\}. \tag{2.17}$$

Integrating (2.17) over all possible value of $\tau$ yields

$$\int_0^t \delta \exp\{-(t-\tau)\nu\} d\tau = \frac{\delta}{\nu}[1 - \exp\{-\nu t\}]. \tag{2.18}$$

The quantity in (2.18) is the expected amount of toxic material absorbed during the interval $[0, t]$ and present in the body at time $t$, which leads to a possible suggestion of a function format for the hazard for cancers caused through exposure to factors. Suppose the baseline hazard for lung cancer patients is proportional to the quantity in the following equation:

$$h_0(t) = \frac{a}{b}(1 - \exp(-bt)). \tag{2.19}$$

Defining the cumulative baseline hazard function, $H_0(t)$, by integrating $h_0(t)$ and applying boundary condition that $h_0(0) = 0$ yield:

$$H_0(t) = \int_0^t h_0(x)dx = \frac{a}{b}[x - \frac{1}{b}(1 - \exp(-bt))]. \tag{2.20}$$

## 3. Statistics Methods and Neural Network

### 3.1. *Maximum Likelihood Estimation*

Maximum likelihood estimation begins with writing a mathematical expression known as the likelihood function of the sample data. Roughly speaking, the likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the model's unknown parameters. The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimates, or MLE. Maximum likelihood estimation is a totally analytic maximization procedure. It applies to every form of censored or multi-censored data, and is even able to be used across several stress cells and estimate acceleration model parameters at the same time as life distribution parameters. Moreover, MLE and likelihood functions generally have very desirable large sample properties because they: (a) become unbiased minimum variance estimators as the sample size increases, (b) have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds, and (c) likelihood functions can be used to test hypotheses about models and parameters. Although it has many good attributes, MLE has an important drawback, that is, with a small number of failures (say, less than 30, and oftentimes, less than 50), MLE may be heavily biased and the large sample optimality properties do not apply.

If $X$ is a continuous random variable with *pdf*

$$f(x, \beta_1, \beta_2, \cdots, \beta_p), \tag{3.1}$$

where $\beta_1, \cdots, \beta_p$ are $p$ unknown constant parameters which need to be estimated. Denote $\beta^\tau = (\beta_1, \cdots, \beta_p)$. Conduct an experiment and obtain $N$ independent observations, $x_1, \cdots, x_N$, which correspond in the case of life data analysis to failure times. The likelihood function is given by

$$L = L(x_1, \cdots, x_N|\beta_1, \cdots, \beta_p) = \Pi_{i=1}^N f(x_i|\beta_1, \cdots, \beta_p). \tag{3.2}$$

The Logarithmic function is

$$l = \log L = \sum_{i=1}^N \log f(x_i|\beta_1, \cdots, \beta_p). \tag{3.3}$$

For the survival analysis, we assume (2.9) and (2.10). Then the *pdf* becomes

$$f(t_i|x_i) = h(t_i|x_i)S(t_i|x_i) = h_0(t_i)\exp\{x_i\beta - \int_0^{t_i} h_0(z)\exp(x_i\beta)\,dz\}. \tag{3.4}$$

The log-likelihood function $l(\beta)$ has the expression

$$\begin{aligned} l &= \sum_{i=1}^N \log f(t_i|x_i) = \sum_i [\log h_0(t_i) + (x_i\beta - \int_0^{t_i} h_0(z)\exp(x_i\beta)\,dz)] \\ &= N\log h_0(t_i) + h_0(t_i) + \sum_i x_i\beta - \sum_i \int_0^{t_i} h_0(z)\exp(x_i\beta)\,dz. \end{aligned} \tag{3.5}$$

When taking partial derivatives with respect to $\beta$ to maximize $l(\beta)$, the computation often becomes very difficult due to the presentation of $h_0(z)$ in the integration term. That is why a proportional hazard model is used in the Cox models so that the term $h_0(z)$ can be canceled out in MLE calculation.

Recall (2.15), the MLE for $\hat{\beta}$ is $s(\hat{\beta}) = 0$, where the score function is

$$s(\beta) = \begin{pmatrix} \frac{\partial l(\beta)}{\partial \beta_1} \\ \cdots \\ \frac{\partial l(\beta)}{\partial \beta_p} \end{pmatrix}. \tag{3.6}$$

One of many nonlinear algorithms to compute this maximization is the Newton-Raphson iteration. The Newton-Raphson algorithm for computing $\hat{\beta}$ starts with an initial guess $\hat{\beta}^{(0)}$ and then iteratively determines $\hat{\beta}^{(m)}$ from the formula

$$\hat{\beta}^{(m)} = U^{-1}(\hat{\beta}^{(m-1)})s(\hat{\beta}^{(m-1)}), \tag{3.7}$$

where

$$U(\beta) = -N \cdot Hessian(\beta) = N \cdot \begin{pmatrix} \frac{\partial^2 l(\beta)}{\partial^2 \beta} & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ \frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 l(\beta)}{\partial^2 \beta_2} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_1} & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_2} & \cdots & \frac{\partial^2 l(\beta)}{\partial^2 \beta_p} \end{pmatrix}. \tag{3.8}$$

The Hessian matrix is positive definite, so it is strictly concave on $\beta$. However, the computation is obviously more complex. In practice, we use software to carry out this process for the MLE.

### 3.2. *Non-Linear Least Square Fit*

Least square regression (LSE) is a very popular and useful tool used in statistics and other fields. Suppose we want to find a relationship between a dependent (response) variable $Y$ and an independent (predictor) variable $X$, in which a statistical relation is

$$Y = g(X|\theta) + \epsilon, \tag{3.9}$$

where $\epsilon$ is the error, and $\theta$ is a vector of parameters to be estimated in function $g$. If $g$ assumes a non-linear format in terms of $X$, we are facing a non-linear regression. Suppose $X = (x_1, \cdots, x_m)^\tau$, $Y = (y_1, \cdots, y_m)^\tau$. We define

$$f_i(\theta) = y_i - \hat{y_i} = y_i - g(x_i|\theta) \tag{3.10}$$

The non-linear least square regression is to find $\hat{\theta}$ which minimizes $F(\hat{\theta})$, where $F(\theta)$ is defined as

$$F(\theta) = \frac{1}{2}\sum_{i=1}^{m}(f_i(\theta))^2 = \frac{1}{2}\|f(\theta)\|^2 = \frac{1}{2}f(\theta)^\tau f(\theta). \tag{3.11}$$

There are many non-linear algorithms for finding $\hat{\theta}$. These well-developed algorithms include the Gauss-Newton method, the Levenberg-Marquardt method, and Powell's Dog Leg method (see [7] for example). In this study, we use the Gauss-Newton method. It is based on the implementation of first derivatives of the components of the vector function. In special cases, it can give quadratic convergence as the Newton-method does for general optimization [8]. The Gauss-Newton method is based on a linear approximation to the components of $f$ (a linear model of f) in the neighborhood of $\theta$: For small $\|h\|$, we see from the Taylor expansion that

$$f(\theta + h) \approx \ell(\theta) := f(\theta)J(\theta)h, \tag{3.12}$$

where $J$ is the Jacob matrix. Inserting this to the definition for $F$, we obtain

$$\begin{aligned} F(\theta + h) \approx L(\theta) := \frac{1}{2}\ell(h)^t\ell(h) &= \frac{1}{2}f^t f + h^t J^t f + \frac{1}{2}h^t J^t Jh \\ &= F(\theta) + h^t J^t f + \frac{1}{2}h^t J^t Jh. \end{aligned} \tag{3.13}$$

The Gauss-Newton step $\hat{h}$ minimizes $L(h)$. In practice, the Gauss-Newton least square fitting the baseline hazard function can be achieved by using MATLAB software.

### 3.3. *Neural Network Testing*

In the Cox model, the main interest is usually about the parameter vector $\beta$. However, when one is interested in making predictions about the failure time for a given set of covariates, or when one assumes a parametric family for the baseline hazard function, just as what we have performed, then testing that $h_0$ is equal to a specified hazard rate function or evaluating how stable $h_0$ is for varying data source becomes important [12]. In the field survival analysis, there are two popular methods in order to test a model. One is to use 1/2 or 2/3 of the time scale in the survival data to determine the parameters and then use the whole data set to examine the model. In our study, however, to the short length of data (total of 66 rows, in which approximately two-thirds are censored) and the high data demand from MLE (refer to section 3.1), this solution is not feasible. Another way is to use the whole data set to set up the model and then use a resample method to check the model. This solution also has a problem on the principle by which we resample the original data. As we have known, MLE relies heavily on the given data set especially when the length of data is not exceptionally long. If we randomly resample the original data, the selected data for testing may be far from the “pattern” of the whole data set, e.g., having quite different mean and standard deviation.

In this study, we propose an artificial neural network testing model. First, we let the neural network “learn” the patients’ survival pattern from the given hospital data. We then use the neural network to generate a long list of “virtual data” and “simulate” the survival pattern to test our covariate estimation and baseline hazard estimation. By this process, we also show that the neural network has great potential as a research tool in survival analysis.

The conception of neural network came up as early as the middle of this century. A `Neural Network` (NN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. Simply speaking, it is software that is "trained" by having its examples of input and the corresponding desired output presented to it.

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been required to analyze.

The typical structure of neural network consists of a layer of $d$ (the dimension of the futures) input units, a layer of output units, and a variable number of hidden layers of units, as shown in Figure 1. Generally more layers result in higher accuracy, but also are more time-consuming on computation.

The construction of the NN for this study and test results will be shown in the next section.

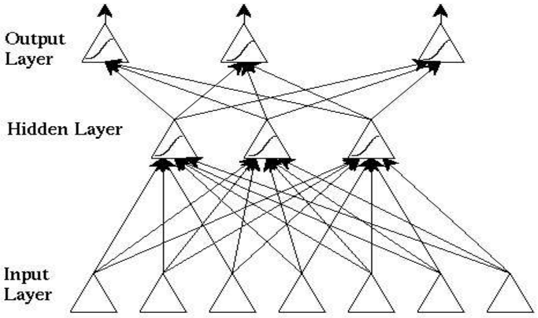

Fig. 1. Typical Structure of Neural Network.

## 4. Application to Lung Cancer Data

### 4.1. *Data Structure*

A data set records the survival times ($S\_INT$, in months) of the patients seen at Vanderbilt University School of Medicine Hospital. The data set also records patients' hospital condition including

$PT$: patient term, ranges from $T1$ to $T4$

$PN$: occurrence of lymph notes, a symptom of cancer invasion, ranges from $N0$ to $N2$

$STAGE$: pathological diagnosis of cancer and it is ordinal, ranges from $1A$ to $IV$

$DF\_INT$: disease free time, in months

$GRADE$: the fitness condition when patient in hospital, ranges from well to poor

$STATUS$: indicating if the patient is still alive (A) or deceased (D). If the $STATUS$ of a patient is "A" (alive), this row of data is censored.

In our study, we take $PT, PN, STAGE, DF\_INT$, and $GRADE$ six variables as covariates to be estimated. The original hospital data set records information for 66 patients and is listed in Appendix 1.

### 4.2. *Estimation for Covariates*

The proportional hazard regression to estimate $\beta$ is performed by SPSS. The results are shown in Appendix 2:

The Cox regression gives the mean and standard deviation for each covariate in given data. The $\beta$ is estimated at a certain significance level. For "patient term" and "grade," $\beta$ is positive, which means a higher value for these two variables will result in higher hazard or risk of death. For "disease free time," $\beta$ assumes a negative value. This means that the longer the patient is disease free, the less likely that he or she will die shortly, which is reasonable. The $\beta$ values for $PN$ and $STAGE$ are both near zero, which indicates that these two variables do not associated much with the hazard rate.

The Cox regression gives baseline cumulative hazard and overall cumulative hazard vs. survival time, at mean value of covariates. To estimate the hazard function, we fix the covariates at their mean values, then use least square regression to estimate the parameters $a$ and $b$ in (2.20), by fitting two columns of data in the survival table in Appendix 2.

### 4.3. *Estimation for Baseline Hazard Function*

Starting from the results of the Cox regression, let

$$\begin{aligned} X^\tau &= \text{SurvivalTime} = [1\ 2\ 3\ 4\ 5\ 6\ 8\ 9\ 11\ 16\ 17\ 18\ 33], \\ H^\tau &= \text{CumBaselineHazard} \\ &= [.006\ .010\ .022\ .029\ .037\ .054\ .065\ .089\ .129\ .163\ .303\ .377\ .991]. \end{aligned}$$

Following the Gauss-Newton least square estimation discussed by section 3.2, we find estimations for $a$ and $b$. The MATLAB computation results are summarized below.

```
FITTEDMODEL =
General model:
FITTEDMODEL(x) = a/b*(x-1/b*(1-exp(-b*x)))
Coefficients (with 95% confidence bounds):
a = 0.002185 (0.001524, 0.002845)
b = 0.01727 (-0.01574, 0.05029)
GOODNESS =
sse: 0.0129
rsquare: 0.9854
dfe: 11
adjrsquare: 0.9840
rmse: 0.0342
OUTPUT =
numobs: 13
numparam: 2
residuals: [13x1 double]
Jacobian: [13x2 double]
exitflag: 1
iterations: 7
funcCount: 22
firstorderopt: 1.4601e-004
algorithm: 'Gauss-Newton'
```

The estimated baseline hazard function is

$$h_0(t) = 0.1265(1 - \exp(-0.01727t)) \qquad (4.1)$$

Figure 2 shows the fit for the cumulative baseline hazard. Figure 3 plots the baseline hazard as a function of time.

### 4.4. *Survival Model Testing*

With the help of MATLAB command `newff`, a feed-forward backpropagation network is constructed to simulate the survival model. This network has a total of three layers: an input layer of dimension 6, a hidden layer of dimension 3, and an output layer of dimension 1. The unit of output layer may assume a value of “0” or “1”, representing “alive” and “dead” respectively. More hidden levels are proven not to improve NN performance. Since

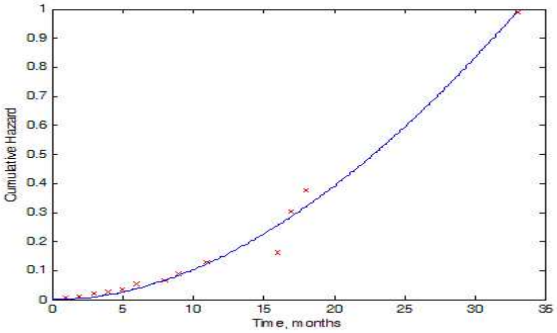

Fig. 2. Fit for Cumulative Hazard.

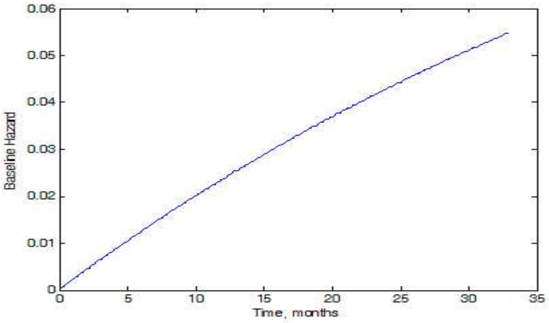

Fig. 3. Baseline Hazard as a Function of Time.

the output values assume only two possible values, we use `logsig` as the nonlinear transfer function between layers.

When having `traingda/learngdm` as the training/learning function, the NN reaches best performance, and the error rate for training set is 9%. The error rate is defined as the rate of false "alive-dead" judgment for all 66 training cases. The network performance is shown in Figure 4.

After the NN is set up, we generate a 1000 × 6 matrix to simulate 1000 patients' record. Each column of the matrix corresponds to a covariate, and each row stores a patient's information on $PT, PN, STAGE, S\_INT, D\_INT$, and $GRADE$. Then we use the trained NN to judge the $STATUS$ of the patient, as we "believe" the NN

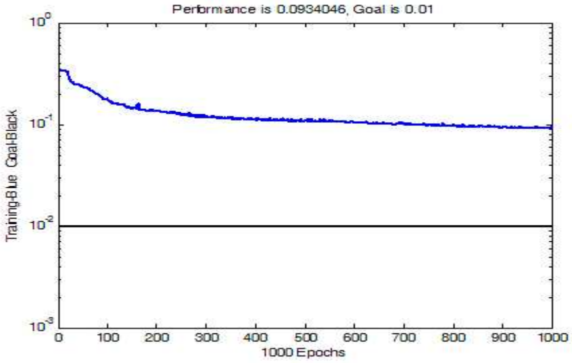

Fig. 4. Network Performance over Epochs.

has learned the “right” survival pattern of lung cancer patients.

At first, we generate the data for each column randomly and uniformly distributed in the domain. For example, the domain for *PN* column is the closed interval $[1, 4]$. All numbers are rounded to integers. After a Cox regression analysis, the computation cannot be converged. This result shows that randomly generated data is not acceptable. The covariates for lung cancer patients must be distributed with a certain pattern.

Recall the Cox regression results for original hospital data. The mean and standard deviation for each covariate are calculated. Respecting this result, another $1000 \times 6$ matrix is generated. For each column, the generated data assume normal distribution with a corresponding mean and standard deviation that are rounded to integers (disregarding that the rounding may shift the mean and deviations for each column).

After a Cox regression and a least square fit for the cumulative baseline hazard as we did before, the baseline hazard for the NN generated data is plotted as a function of time. It is compared to the baseline hazard function we found before for the original hospital data, as shown in Figure 5.

Further more, define the score function

$$s(x, \beta) = x^{\tau} \cdot \beta. \tag{4.2}$$

Then the hazard function changes to be

$$h(t|x_i) = h_0 \exp(s). \tag{4.3}$$

The score function determines the risk of death. The higher score, the more likely a patient will die (or will die sooner).

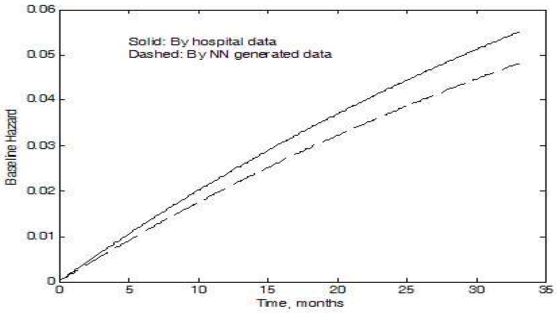

Fig. 5. Estimated Baseline Functions.

A scatter plot for "score vs. survival time" is shown in Figure 6. Notice that time assumes a negative value if it is censored (patient is still alive.)

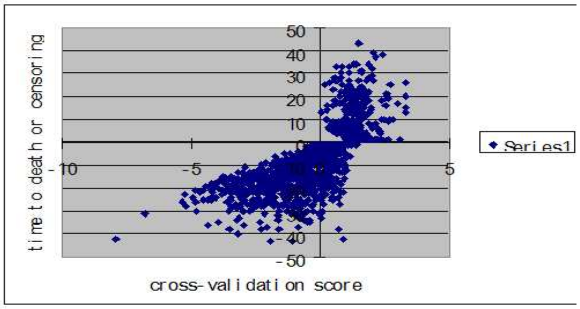

Fig. 6. Scores vs. Time to Death or Censoring.

Figure 6 shows that when a patient scores negative or very small value, he or she tends to survive; the lower the score is, the longer he or she will live. On the other hand, a high positive score means death. This proves that proportional hazard regression is a beneficial way to estimate $\beta$ coefficients.

*Final Remarks:* 1. In this study, we set up a survival model for lung cancer patients. This was achieved by three steps: using proportional hazard regression to estimate the coefficients for five covariates, using non-linear least square fit to estimate the exponential baseline hazard function, and using a neural network to exam the survival model. The analysis tools used in this research were SPSS, EXCEL and MATLAB.

2. MLE is a powerful statistical tool but it has its own limitation. When the data length is short, MLE might be heavily biased. In this study, there were data for 66 patients, but two thirds were censored and only one third is used in MLE. The shortage of data resulted in a unideal significance level of the estimation.

3. Neural network simulation is a new idea for testing the model, especially when the original data set is short. Neural network application in survival analysis has promising prospects.

4. Although we assume a linear combination format in the score function, the five covariates are believed to be correlated with each other. A randomly generated covariate matrix may not result in a convergent Cox regression.

5. When the NN generated data assume the same mean and SD with the original data, they tend to have similar baseline hazard functions by LSE. This supports our assumption on the format of baseline function.

6. The score function provides a good indication for the risk of death. This supports the Cox regression for $\beta$ estimation.

7. In future work, we may do regression for longer hospital data for a more stable $\beta$ estimation and attempt to find out the correlation among the parameters, assuming a more accurate model for $f(x|\beta)$ in the hazard function and re-formulate the MLE in proportional hazards regression. This is quite complex work but truly worth to do. We may also explore more NN applications in survival analysis.

8. In survival analysis with long-term survivors, handling situations consisting of a proportion of subjects under study that may never experience the event of interest, one proposes to formulate the model as a mixture of long-term survivors (subjects that will never "fail") and susceptibles (subjects that will "fail" eventually). In [18], comparing (4.3), the hazard rate function is modeled as $h(t|x_i) = h_0(t)\exp(s)$ with $h_0(t) = \frac{pf_0(t)}{1-pF_0(t)}$ and $0 < p \leq 1$, here, $f(t)$ and $F(t)$ are defined in (2.1). Partial likelihood and full likelihood are then used to obtain the estimators of the coefficients of covariates and the proportion of long-term survivors.

## Acknowledgments

The authors would like to thank William Wu, Department of Biostatistics, Vanderbilt University for providing data sets for this study. This research was supported in part by Lung Cancer SPORE (P50 CA90949), Breast

Cancer SPORE (1P50 CA98131-01), GI (5P50 CA95103-02), and Cancer Center Support Grant (CCSG) (P30 CA68485) for Y. Shyr, and by NSF IGMS (#0408086 and #0552377) and MTSU REP for D. Hong.

## References

1. D.R. Cox, Regression models and life tables, *Journal of the Royal Statistical Society*, Series B, **34** (1972), 187-220.
2. N.L. Bowers et al, Actuarial Mathematics, Second Edition, Society of Actuaries, 1997.
3. S.J Walters, What Is a Cox Model, Hayward Medical Communications, volume 1, number 10, May 2001.
4. Hsi-Wen Liao, A simulation study of estimation in stratified proportional hazards model, In: NESUG 1998 Proceedings, pp. 118-125, Pittsburgh, PA, 1998.
5. Lung Cancer Transcripts, www.canceranswers.com.
6. C.L. Chiang and P.M. Conforti, A survival model and estimation of time to tumor, *Mathematical Biosciences*, **94** (1989), 1-29.
7. K. Madsen, H.B. Nielsen, and O. Tingleff, Methods for Non-Linear Least Squares Problems, 2nd Edition, Informatics and Mathematical Modeling, Technical University of Denmark, April 2004.
8. P.E. Frandsen, K. Jonasson, H.B. Nielsen, and O. Tingleff, Unconstrained Optimization, 3rd Edition, IMM, DTU, 2004.
9. P.K. Andersen and R.D. Gill, Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10** (1982), 1100-1120.
10. D.Y. Lin and L.J. Wei, The Robust inference for the Cox proportional hazards model, *Journal of American Statistician Association*, **84** (1989), 1074-1078.
11. D.A. Binder, Fitting Cox's proportional hazards models from survey data, *Biometrika*, **79** (1992), 139-147.
12. E.A. Pena, Smooth goodness-of-fit tests for the baseline hazard in Cox's proportional hazards model, *Journal of American Statistical Association*, **93** (1998), 673-692.
13. US Mortality Public Use Data Tape 2001, National Center for Health Statistics, Centers for Disease Control and Prevention, 2003.
14. 1950 Mortality Data – CDC/NCHS, NVSS, Mortality Revised.
15. 2001 Mortality Data – NVSR, Death Final Data 2001, Vol. 52, No. 3.
16. B.D. Ripley and R.M. Ripley, Neural networks as statistical methods in survival analysis, In: Artificial Neural Networks: Prospects for Medicine, (R. Dybowski and V. Grant Eds.), pp. 237-255, Cambridge University Press, 2001.
17. E. Pitacco, Survival models in a dynamic context: a survey, *Insurance: Mathematics and Economics*, **35** (2004), 279–298.
18. X. Zhao and X. Zhou, Proportional hazard models for survival data with long-term survivals, *Statistics and Probability Letters*, **76** (2006), 1685-1693.

## Appendix 1: Patients Data

| # | *PT* | *PN* | *STAGE* | *STAT* | *S_INT* | *DF_INT* | *GRADE* |
|---|---|---|---|---|---|---|---|
| 1 | T1 | N2 | IIIA | D | 11 | 5 | mod |
| 2 | T4 | N2 | IIIB | D | 11 | 9 | poor |
| 3 | T1 | N1 | IV | D | 17 | 0 | poor |
| 4 | T2 | N0 | IB | A | 24 | 24 | well-mod |
| 5 | T2 | N0 | IV | D | 9 | 0 | mod-poor |
| 6 | T2 | N2 | IIIA | A | 21 | 7 | well-mod |
| 7 | T4 | N0 | IV | D | 1 | 1 | poor |
| 8 | T1 | N0 | IA | A | 21 | 13 | well-mod |
| 9 | T3 | N0 | IIB | D | 2 | 0 | mod-poor |
| 10 | T2 | N0 | IB | A | 20 | 20 | mod |
| 11 | T1 | N0 | IA | D | 3 | 3 | mod |
| 12 | T2 | N0 | IB | A | 23 | 23 | poor |
| 13 | T1 | N0 | IA | D | 8 | 8 | mod-poor |
| 14 | T2 | N1 | IIB | A | 21 | 21 | mod |
| 15 | T2 | N0 | IB | A | 20 | 20 | mod |
| 16 | T2 | N0 | IB | D | 33 | 30 | mod-poor |
| 17 | T2 | N0 | IB | A | 18 | 18 | mod-poor |
| 18 | T2 | N2 | IIIA | D | 6 | 0 | poor |
| 19 | T2 | N2 | IIIA | D | 3 | 3 | mod-poor |
| 20 | T1 | N1 | IIA | D | 5 | 0 | poor |
| 21 | T2 | N2 | IIIA | A | 21 | 17 | poor |
| 22 | T2 | N0 | IB | A | 23 | 10 | mod-poor |
| 23 | T2 | N0 | IB | A | 26 | 26 | well-mod |
| 24 | T2 | N0 | IB | A | 26 | 26 | mod |
| 25 | T1 | N2 | IIIA | D | 18 | 0 | poor |
| 26 | T2 | N1 | IIB | A | 17 | 17 | mod-poor |
| 27 | T2 | N0 | IIB | A | 33 | 9 | mod |
| 28 | T2 | N0 | IB | D | 17 | 17 | mod |
| 29 | T2 | N0 | IIB | A | 42 | 42 | mod-poor |
| 30 | T2 | N0 | IIB | D | 16 | 5 | poor |
| 31 | T1 | N1 | IIA | D | 1 | 0 | poor |
| 32 | T2 | N0 | IB | D | 17 | 15 | poor |
| 33 | T2 | N2 | IIIA | D | 9 | 0 | poor |
| 34 | T2 | N2 | IIIA | D | 4 | 0 | mod-poor |

**Appendix 1: Patients Data (Cont.)**

| # | *PT* | *PN* | *STAGE* | *STAT* | *S_INT* | *DF_INT* | *GRADE* |
|---|---|---|---|---|---|---|---|
| 35 | T2 | N0 | IB | A | 2 | 1 | poor |
| 36 | T2 | N0 | IB | A | 5 | 1 | well-mod |
| 37 | T2 | N2 | IIIA | A | 6 | 6 | mod |
| 38 | T1 | N0 | IA | A | 1 | 1 | well |
| 39 | T1 | N0 | IA | A | 1 | 1 | mod |
| 40 | T1 | N0 | IA | A | 3 | 3 | mod-poor |
| 41 | T1 | N0 | IA | A | 1 | 1 | mod-poor |
| 42 | T1 | N0 | IA | A | 1 | 1 | well-mod |
| 43 | T3 | N0 | IIB | A | 1 | 1 | well |
| 44 | T1 | N0 | IA | A | 1 | 1 | poor |
| 45 | T2 | N0 | IB | A | 2 | 2 | poor |
| 46 | T2 | N0 | IB | A | 1 | 1 | well-mod |
| 47 | T2 | N0 | IB | A | 1 | 1 | mod |
| 48 | T1 | N0 | IA | A | 12 | 0 | mod-poor |
| 49 | T1 | N2 | IIIA | A | 6 | 4 | mod-poor |
| 50 | T2 | N0 | IB | A | 1 | 1 | mod |
| 51 | T2 | N0 | IB | A | 3 | 3 | poor |
| 52 | T3 | N0 | IIB | A | 10 | 4 | poor |
| 53 | T3 | N1 | IIIA | D | 6 | 6 | poor |
| 54 | T2 | N0 | IB | A | 1 | 0 | mod |
| 55 | T4 | N1 | IIIB | A | 2 | 0 | mod-poor |
| 56 | T2 | N0 | IB | A | 1 | 1 | mod |
| 57 | T2 | N0 | IB | A | 1 | 1 | mod-poor |
| 58 | T2 | N0 | IB | A | 5 | 4 | poor |
| 59 | T1 | N2 | IIIA | A | 1 | 1 | poor |
| 60 | T1 | N0 | IA | A | 1 | 1 | mod |
| 61 | T1 | N0 | IA | A | 7 | 7 | poor |
| 62 | T2 | N0 | IB | A | 2 | 2 | mod |
| 63 | T2 | N1 | IIB | A | 1 | 1 | mod |
| 64 | T2 | N2 | IIIA | A | 11 | 4 | poor |
| 65 | T1 | N0 | IA | A | 10 | 3 | poor |
| 66 | T1 | N0 | IA | A | 1 | 1 | poor |

## Appendix 2: Cox Regression Results

### Covariate Means

| | Mean |
|---|---|
| PT | 1.833 |
| PN | .515 |
| STAGE | 3.000 |
| $D_FREE$6.879 | |
| GRADE | 2.788 |

### Survival Table

| | | At mean of covariates | | |
|---|---|---|---|---|
| Time | Baseline Cum Hazard | Survival | SE | Cum Hazard |
| 1.00 | .006 | .983 | .012 | .017 |
| 2.00 | .010 | .971 | .018 | .029 |
| 3.00 | .022 | .939 | .030 | .062 |
| 4.00 | .029 | .922 | .036 | .082 |
| 5.00 | .037 | .903 | .042 | .102 |
| 6.00 | .054 | .860 | .053 | .151 |
| 8.00 | .065 | .835 | .061 | .181 |
| 9.00 | .089 | .780 | .073 | .248 |
| 11.00 | .129 | .698 | .091 | .359 |
| 16.00 | .163 | .635 | .105 | .455 |
| 17.00 | .303 | .431 | .123 | .842 |
| 18.00 | .377 | .350 | .121 | 1.050 |
| 33.00 | .991 | .064 | .115 | 2.755 |