A Bayesian Neural Network for Censored Survival Data

by

Helen Wong

A thesis submitted in partial fulfillment of the requirements of Liverpool John Moores University for the Degree of Doctor of Philosophy

October 2001

This thesis is dedicated to my parents.

Fuk Sang, Wong and Lai Chun, Chan

學而不思則罔
思而不學則殆

孔子（論語　爲政）

" Learning without thought is labour lost; thought without learning is perilous."

(Confucius: Discourses and Dialogues, Book II)

Abstract

Medical statistics has important applications in cancer research, in particular through the analysis of censored survival data. Moreover, breast cancer is responsible for thousands of deaths each year in Britain, and is among the leading causes of mortality among women. This thesis is about the robust application of neural networks to survival analysis of breast cancer patients, taking advantage of its non-linearity and flexibility but providing an automatic mechanism to prevent over fitting of the data.

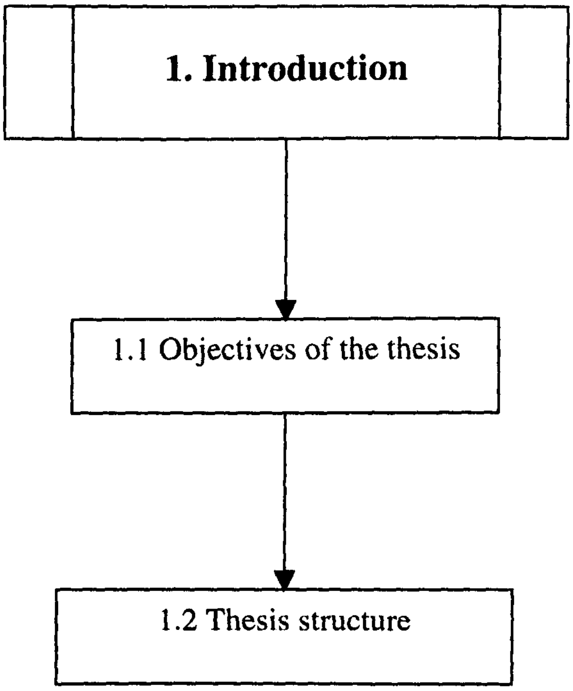
Censorship is a feature of survival data, which arises when the endpoint of interest cannot be observed for a particular individual. In this thesis, a Bayesian regularised neural network model that accommodates censorship is introduced, extending the "Partial Logistic Artificial Neural Network (PLANN) model". Within the neural network model, categorical data are treated differently from ordinal data and requires bias correction for the network prediction when the data distribution is heavily skewed. The network also uses the Automatic Relevant Determination (ARD) technique within the Bayesian regularisation framework, to perform a backward model elimination. The use of non-linear variable selection methods leads to the identification of pairwise interactions between covariates that may be implicitly modelled by the neural network or explicitly added to Cox regression, which is the most commonly used statistical modelling tool for survival analysis. Both methods were applied to the modelling of post-operative mortality with 5 years follow-up of two patients groups. The first group was used to design and compare the two methods, and comprises patients recruited between 1983-1989, The second group is used to validate the model's performance, and comprises patients recruited between 1990-1993. The two sets of data were divided into two cohorts each, according to the clinical separation criteria for low-and high-risk. The

missing data in the data sets were treated as a separate category. Performance estimation for the design data set was carried out through the use of v-fold cross validation. Patients were also divided into mortality risk groups using the log-rank test applied to a prognostic index, and the predicted survivorship for prognostic groups are assessed by the observed survivorship, which is described by the Kaplan-Meier survival estimation.

The robustness of Cox regression was explained by explicitly plotting the estimated hazard over time, showing that deviations from proportionality of the hazards are minor. The proposed extension of the PLANN model has successfully identified interaction terms that were added to the Cox regression model to improve prognostic group separation and attribute specificity.

Contents

Chapter 1


1. Introduction
1.1 Objectives of the thesis
1.2 Thesis structure

1 Introduction

This chapter gives an overview the issue in survival analysis that is this thesis, and the methodologies proposed in it, are intended to address. This is followed by a review of the structure of the remaining thesis chapters.

1.1 Objectives of the thesis

Cancer research is generally focused on improving patient survival rates, whether through early detection, development of new drugs, or improvements in therapy. However, surgery and adjuvant therapy carry with them significant side effects. Treatment and surgery are assigned following guidance based on standard clinical factor measurements, often without direct reference to accurate estimates of individual survivorship, Smith (2000). The main objective of this thesis is to predict the survivorship over time for individual patients through the use of Cox regression and neural network models, and to permit a clinical interpretation to be ascribed to the predictive results obtained from both of the models.

Survival data have special characteristics, for instance the data are not symmetrically distributed with a trend to be *positively skewed*, that is having a longer 'tail' to the right of time intervals. Also censorship is an inherent feature of survival data and arises when the end-point for particular individuals is not the event of interest, making the outcome beyond a fixed time point indeterminate. However, excluding these data from the model can introduce significant bias, Ravdin and Clark (1992) and Brown *et al* (1997), therefore these patients must remain in the study for the time they were observed. An example of censorship would be an early death from a cause unrelated to the breast cancer, sometimes called an intercurrent death.

The most widely used statistical modelling method for censored data is Cox regression, Cox (1972), which is based on the assumption that the hazards of different patient groups remain proportional to the baseline hazard over time. Some other well-known parametric statistical methods, are the Weibull model and accelerated failure time model, Collett (1994). Efron (1988) also proposed a flexible non-linear model using cubic spline and Bennett (1983) introduced log-logistic regression models for survival data, which require proportionality of the survival log-odds ratio, instead of the probability of death in a particular time interval that is the hazard ratio.

Artificial neural networks (ANN) are non-linear, semi-parametric models that have recently been considered as alternative methods for analysing survival data. Radvin *et al* (1992) proposed an extension of proportional hazard model using a standard MLP architecture with multiple output nodes to accommodate the censorship, where each output node represented a time interval. However, for a monthly study, this method requires many output nodes. Biganzoli *et al* (1996) introduced the Partial Logistic Artificial Neural Network model (PLANN), which is a straightforward Multi-Layer Perceptron, MLP, where censorship is encoded via the data structure. By assigning target values of zero and one to each patient record while observed alive, or when event of interest happened in that time interval, respectively, but omitting any target values after censorship. A patient will remain in the population at risk only while observed, but is removed from the study when the outcome for that time interval is not observed. This model has advantage of not requiring proportionality of the hazards overtime, and can implicitly model interactions between variables. However, neural networks are prone to over-fitting unless careful regularisation is applied. The Bayesian neural network approach (MacKay, 1992a,b) is commonly used to regularise binary classification problems without censorship including soft model selection through Automatic Relevance Determination (ARD) where the hyperparameters regularising the objective function suppress irrelevant variables. The magnitude of the hyperparameters thus provides a rank order that reflects the relative importance of the variables to the model predictions. Neural network models are often benchmarked with traditional statistical tools, Groves (1999), Radvin and Clark (1992), and in this study the regularised PLANN model is compared with Cox regression.

In this thesis, a longitudinal study is conducted where the modelling methodologies are developed using a data set with 1,616 records and tested with a further 1,653 records. They all comprise women patients admitted to Manchester Christie Hospital during 1983 to 1989, and 1990 to 1993, respectively, who were followed-up for at least 5 years after surgery. These two sets of data contain demographic information, clinical investigations, laboratory test results, post-surgery and treatment assignment, but do not include any genetic or life style information. Each of the data sets is divided into two cohorts on the basis of clinical staging, divided generically into low-and high-risk. Within these data sets, some of the variables contained large amounts of missing values. The attribute 'missing' was treated as a separate category, although investigations were also carried out predicting missing values using Nominal Logistic Regression.

Assigning patients into prognostic risk groups is of considerable importance in the management of breast cancer patients. A key objective of this thesis is to partition them into prognostic groups based upon their risk of mortality. The observed survival for particular patients groups is estimated non-parametrically by the Kaplan-Meier survival estimation (1958). In this thesis, we propose an extension of the PLANN model to include the estimation of hyperparameters within the Bayesian framework. The extended PLANN model is then applied to two monthly studies of mortality risk following breast cancer re-section, with follow-up to 60 months, for each of the two cohorts. The patients in each cohort are partitioned into prognostic groups using a prognostic risk indice derived from (i) proportional hazards model analysis and (ii) the Bayesian implementation of PLANN. The performances of the two approaches are compared for the design data using 3- and 5-fold cross validation for high-risk cohort and low-risk cohort, respectively, and the generality of the results thus obtained is validated using the later cohorts.

Forward step-wise variable selection was carried out using the proportional hazards model, and for the high-risk cohort additional variable selection was investigated also with ARD, using backward elimination. From a comparison of these two approaches to variable selection, specific interaction terms were identified that when integrated into the proportional hazards model, enhanced the differences in survivorship between the prognostic groups.

1.2 Thesis Structure

In the next chapter, chapter (2), details of the two sets of data are described, including characteristics of the explanatory variables, the distribution of missing data and mechanism, and the process of filling-in the missing data using Nominal Logistic Regression and results.

Chapter (3) summarises the literature review of the two modelling methods used to analysis the data, Cox regression and the Bayesian regularised neural networks.

The data analysis results using Cox regression are reported in chapter (4) using two approaches, predicting the event occurrence time and predicting the survivor function over time for identified mortality risk groups. The event occurrence time prediction for individuals is defined by the cross point of the threshold value and the estimated survival function over time, and presented with the Receiver Operating Characteristic (ROC) curve, Hanley (1989). This approach is considered to be sub-optimal. Then the data are divided into low- and high-risk cohorts according to the clinical staging criteria in the second approach. In each cohort, patients are divided into mortality risk groups according to the risk indexes by observing the indexes natural grouping behaviour and the log-rank test. The accuracy of the Cox survivorship prediction for each risk group is assessed by comparison with the observed survivor function, which is described by Kaplan-Meier estimate. Model selection in both of the approaches is implemented with the forward elimination procedure.

The alternative modelling method used in this thesis is Bayesian neural networks with the evidence approximation. Chapter (5) reports a preliminary study of Bayesian neural networks handling censored survival data, using the same two approaches as used earlier with Cox regression. Although neural networks give betters result than the Cox regression for event prediction, the result cannot be concluded to be significant. This chapter only reports the result for the low-risk cohort from the second approach, survivorship prediction for risk groups, using the PLANN model. Within these sections, two new modelling improvement techniques are introduced, baseline population assignment for categorical data and marginalising network output towards the averaged hazard of the data, that are necessary since the data are heavily skewed. The proportionality of hazards between risk groups is visualised simply by displaying the predicted hazard for each group over time.

A similar neural networks analysis is repeated for the high-risk cohort, which is summarised in chapter (6). Moreover, model selection using ARD is investigated. The selected models help to identify interactions between variables, which then can be explicitly represented in Cox regression models. The difference between the results by the neural networks and Cox regression for the low-risk cohort, lead to a further search for interaction terms. As a result, two pairs of interactions are identified, which apply separately to the highest and the lowest survival patients groups. However, these two interactions between variable pairs cannot be efficiently combined in a single Cox model, as they work against each other.

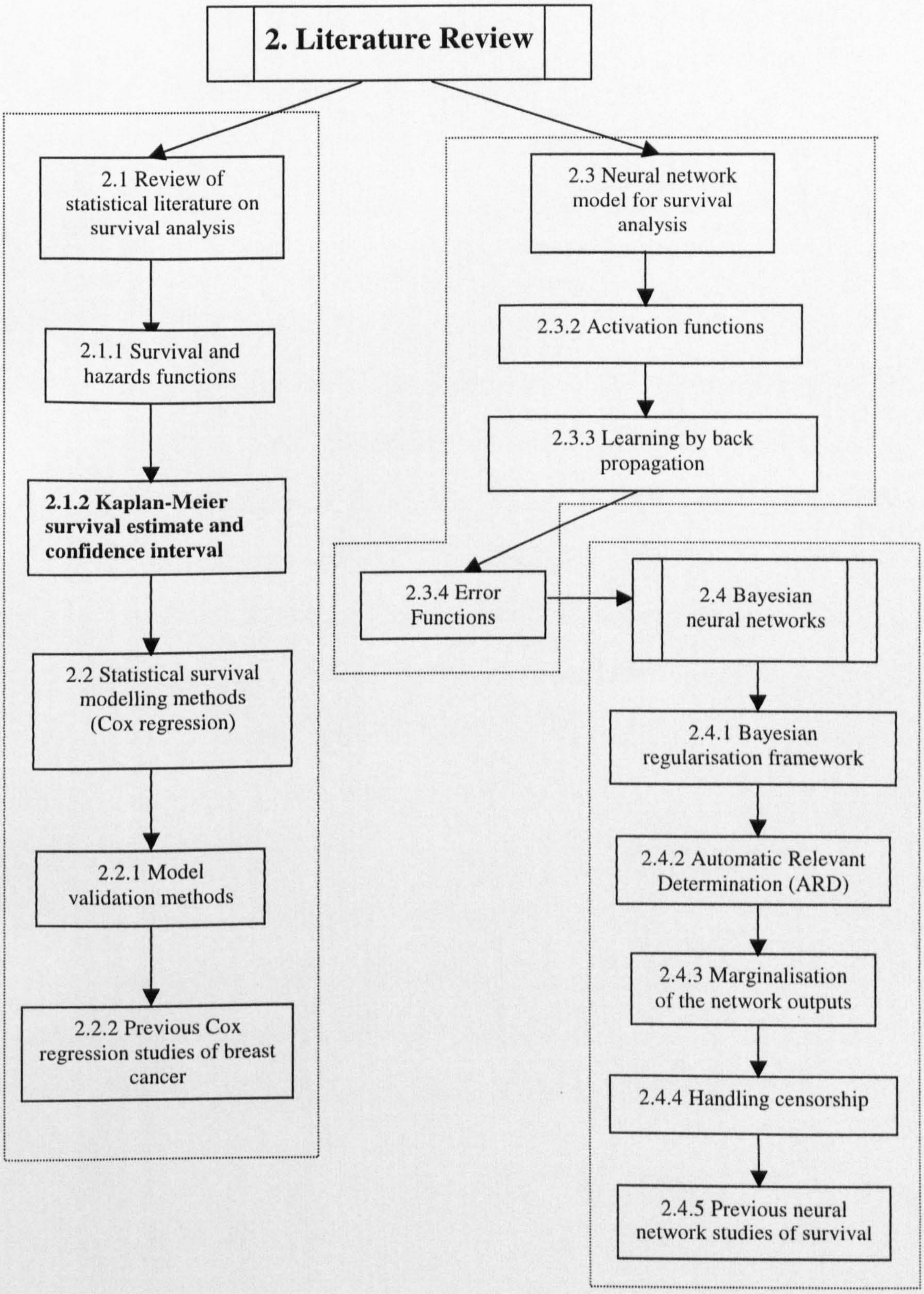
In the longitudinal study, the preferred Cox regression and PLANN models are tested with an independent data set, the results of which are shown for both cohorts in chapter

(7). The results show that the data distribution and the survivorship over time of two data sets are different.

The investigation of handling missing data methods is summarised in chapter (8). In this chapter, the results for the filled-in data using nominal logistic regression are reported. At this stage, the analysis is by Cox regression and variable interactions are not considered. The results using the previously defined models for each cohort are compared with the newly selected models, filling-in the missing data. The only difference between the predictions occurs in the high-risk cohort, since one of the data separation criteria contained large amounts of missing data.

Finally, the discussion of the results between two cohorts recruited over consecutive time periods and the comparison of two modelling methods are summarised in the conclusion, chapter (9).

Chapter 2


2. Literature Review
2.1 Review of statistical literature on survival analysis
2.1.1 Survival and hazards functions
2.1.2 Kaplan-Meier survival estimate and confidence interval
2.2 Statistical survival modelling methods (Cox regression)
2.2.1 Model validation methods
2.2.2 Previous Cox regression studies of breast cancer
2.3 Neural network model for survival analysis
2.3.2 Activation functions
2.3.3 Learning by back propagation
2.3.4 Error Functions
2.4 Bayesian neural networks
2.4.1 Bayesian regularisation framework
2.4.2 Automatic Relevant Determination (ARD)
2.4.3 Marginalisation of the network outputs
2.4.4 Handling censorship
2.4.5 Previous neural network studies of survival

2. Literature Review

2.1 Review of statistic literature on survival analysis

In many clinical studies, it is important to estimate the probability that set intervals of time occur before an event of interest, which may be death ascribed to a particular cause, recurrence of a disease or another prescribed event. The answer to these questions can be described with two functions, survivor function and hazard function, they are of central interest for analysing survival data. Survival data are not amenable to standard statistical procedures used in data analysis because of censorship and the unsymmetrical distributions of the data. The survival time of the data often appear to be positively skewed, that is, having a longer 'tail' to the right of the time intervals. The life-table and Kaplan-Meier methods (1958) are most commonly used for estimating the survival and hazard functions given an observed population. They are known as non-parametric, since they do not need a specific assumption to be made about the underlying distribution of the survival time or indeed any covariate dependencies. An other special feature of survival data is censorship, where the end point of an individual is not the event of interest, such as those who survived beyond the end of the study and those who are lost of follow-up, for instance due to death from an unrelated cause. The event of interest is usually either the death caused by a particular disease, or the recurrence of a disease.

The most commonly used modelling methods for survival analysis are the Cox Regression Model and the Weibull Model, Collett (1994). The Weibull Model was introduced in 1951 in the context of industrial reliability testing and depends on a particular form of probability distribution for the hazard function, hence it is referred to as a parametric model. Alternatively, Cox regression, to be described in section (2.2), has been used extensively for survival analysis for more than 20 years and is also known as *the Proportional Hazards Model.* This model has much flexibility and widespread applicability.

Another general family of survival models is given by the proportional odds model, also introduced by Cox (1972). It is a parametric method if the survival times for individuals are assumed to have a specific probability distribution, such as log-logistic distribution. One of the characteristics of the log-logistic proportional odds model is the involvement of time as an exponential variable.

2.1.1 Survivor and Hazard Function

Let t be the actual survival time of an individual, which can be regarded as a single nonnegative random variable, T. The hazard function $h(t)$ is the probability that an event happens between time t and $t+\delta t$ for that individual, conditional upon the individual having survived up to that time. This is defined as

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \le T < t + \delta t \mid t \le T)}{\delta t}.$$

The survivor function gives the probability that the individual survives longer than a particular time t, so that

$S(t) = P(T \ge t)$,

and also $S(t) = \exp\{-H(t)\}$,

where $H(t) = \int_0^t h(u)du$ is called the *cumulative hazard,* Collett (1994).

2.1.2 Kaplan-Meier Estimate Survival Function

The Kaplan-Meier estimate (1958), also known as product-limit estimate, is a non-parametric method capable of describing the survivor function for discrete censored survival data. Time is split into several time intervals, each includes at least one event case. The time intervals are not necessarily uniformly distributed. There is no interval starting at the censored time and the censored time interval falls between the death time intervals. There could be more than one individual observed to experience the event of interest at any particular event time as illustrated in figure (2.1), where C is the censored data and D represents the event cases.

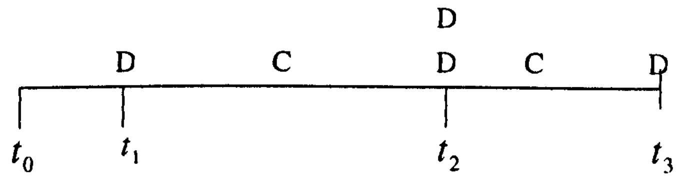


Figure (2.1): The structure of the event time and the relationship with the censored time of the Kaplan-Meier estimate.

Suppose there are *n* individuals observed with observed times $t_1, t_2, \ldots, t_n$. There are *r* death times in total, r ≤ n., so the ordered death times are $t_{(1)} < t_{(2)} < \ldots < t_{(j)}$, where j=1, 2, ..., r and d_j denotes the number of death at that time interval. The probability of an individual dying within that time interval is estimated by d_j / n_j and the corresponding estimated survival rate for that interval is $(n_{(j)} - d_{(j)}) / n_{(j)}$.

The probability of survival to time t is

$$\hat{S}(t) = \prod_{j=1}^{k} (\frac{n_j - d_j}{n_j}),$$

for $t_{(k)} \leq t < t_{(k+1)}$, where k=1,2,...r, $t_{(k+1)}$ is taken to be ∞ and $S(0) = 1$. A plot of the Kaplan-Meier survival estimation is a step function, the estimated probability of survival is constant between adjacent death times and the curve is decreased over time due to the multiplication of probability of survival of each time interval. The graphical presentation of survival curve is used widely. An example of Kaplan-Meier curves is illustrated in section (2.1.2.2) using the breast cancer data and also a Kaplan-Meier survival plot of each variable of breast cancer data is displayed in appendix (I).

2.1.2.1 Standard Error and confidence interval of Kaplan-Meier estimate

The Kaplan-Meier survival estimation can be written as

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j ,$$

for k =1,2,...,r, where $\hat{p}_j = (n_j - d_j) / n_j$ is the estimated probability that an individual survives from the beginning of time j through that interval. Then the number of individuals who survive through the interval can be assumed to have a binomial distribution with parameters n_j and p_j, where p_j is the true probability of survival of that interval. The variance of a binomial random variable with parameters *n, p* is *np(1-p)*. Therefore, the variance for the observed number of survivors, $n_j - d_j$ is given by

$\mathrm{var}(n_j - d_j) = n_j p_j (1 - p_j)$.

The variance of $\hat{p}_j$ can be estimated by $\hat{p}_j(1-\hat{p}_j)/n_j$.

An approximation for the estimated standard error of the Kaplan-Meier estimate of the survivor function is given by

$$s.e.\{\hat{S}(t)\} \approx [\hat{S}(t)]\left\{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\right\}^{1/2},$$

for $t_{(k)} \le t < t_{(k+1)}$, which is also known as *Greenwood's formula*, Collett (1994), chapter 2.

Once the standard error of the estimated survivor function has been calculated, confidence intervals for the estimated survivor functions can also be found. The confidence interval is a range of values around the estimate, gives a percentage level that the true underlying survivor function is included within the interval. In general a 100(1-α)% confident interval for the estimated survival is given by

$$\hat{S}(t) \pm z_{\alpha/2} s.e.\{\hat{S}(t)\}.$$

The $\pm z_{\alpha/2}$ are the upper and lower 1-α/2 points of the standard Normal Distribution respectively, where $s.e.\{\hat{S}(t)\}$ is the standard error of the estimated survivor function given by *Greenwood's formula.*

2.1.2.2 Illustration the use of Kaplan-Meier curves

Table (2.1) displays the survival time and the status of 41 patients, status labelled with 1 represents the event of interest which is death due to breast cancer otherwise 0. Table (2.2) illustrates the necessary calculation needed to construct the Kaplan-Meier survival curve that displays in figure (2.2).

Subjects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Survival time in months	15	61	42	12	61	45	57	19	7	39	45	20	45	30	61	52	18
Status	1	0	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1

Subjects	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Survival time in months	57	28	32	17	26	27	61	23	44	61	27	44	52	37	8	47	61
Status	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0

Subjects	35	36	37	38	39	40	41
Survival time in months	61	27	61	14	61	7	24
Status	0	1	0	0	0	1	1

Table (2.1): An example of 41 subjects with their survival time in months and status labelling.

Survival time in months(t)	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
0	41	0	1.0000	1.0
7	41	2	0.9756	0.9512
8	39	1	0.9750	0.9268
12	38	1	0.9743	0.9024
15	37	2	0.9737	0.8774
17	35	1	0.9722	0.8523
18	34	1	0.9714	0.8272
19	33	1	0.9706	0.8022
20	32	1	0.9697	0.7771
23	31	1	0.9687	0.7520
24	30	1	0.9677	0.7270
26	29	1	0.9666	0.7019
27	28	3	0.9655	0.6267
28	25	1	0.9614	0.6016
30	24	1	0.9599	0.5766
32	23	1	0.9583	0.5515
37	22	1	0.9564	0.5264
42	21	2	0.9545	0.5001
44	19	2	0.9498	0.4475
45	17	3	0.9441	0.3685
47	14	1	0.9326	0.3422
52	13	2	0.9283	0.2895
57	11	2	0.9156	0.2369

Table (2.2): Kaplan-Meier estimate of the survivor function for the data from table (2.1).

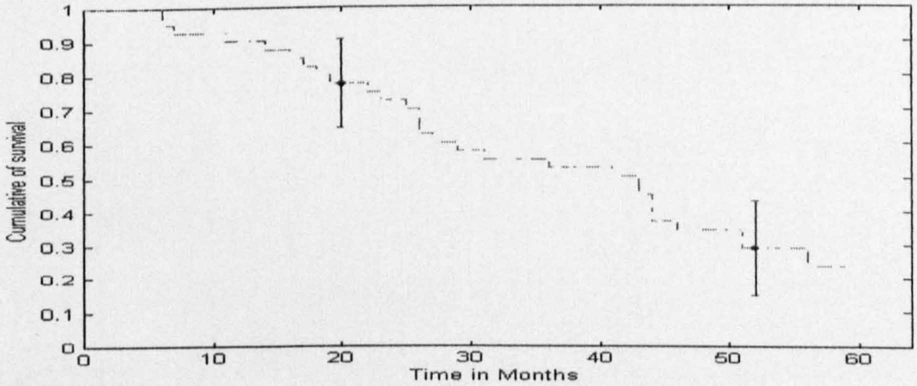


Figure (2.2): Graphically illustrates the Kaplan-Meier estimate of survivor function for the samples in table (2.1).

2.2 Statistical Modelling

In the analysis of survival data, the centre of interest is the probability of a specific event occurring at some time after the recruitment date for that individual. Cox (1972) proposed the *Proportional Hazard Model*, which is referred to as "Cox regression" in the following context. It is the most commonly used statistical modelling method for discrete censored survival data, in which the hazard function is modelled directly as a linear summation of attribute values. The Cox regression is referred to as a semi-parametric model, since it does not make direct assumptions about the underlying distribution of the hazards in different groups, except that the hazard for different patient groups remains proportional to that of a pre-selected baseline population. It allows a non-constant hazard rate to be modelled and involves determining which combination of potential explanatory variables corresponds to the form of the hazard function and also estimates the hazard function itself for an individual. Cox regression can be described as predictive, whereas Kaplan-Meier estimation is descriptive. From the relationship between the hazard function and the survivor function, described as above, an estimate of survivor function can be found. Let the $h_o(t)$ be the baseline hazard function at time t. The general proportional hazard model for the ith individual can be written as

$$h_i(t) = \exp(\beta_p x_{pi})h_0(t),$$

where x is the explanatory variables and p is the number of explanatory variables. The time dependence is described in the baseline population. The β is the unknown coefficients of the corresponding explanatory variables and can be estimated using the *method of maximum partial likelihood*, since the likelihood function does not make direct use of the actual censored and uncensored survival times. The maximum likelihood estimates of the β-parameters can be achieved by maximising the logarithm of the likelihood function, which is accomplished using the Newton-Raphson procedure, Collett (1994),which involves the derivative of the log-likelihood function.

The likelihood function over all death time for the Cox regression is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' x_{(j)})}{\sum_{l \in R(t_{(jP)})} \exp(\beta' x_l)},$$

where $R(t_{(j)})$ is the set of individuals who are alive and uncensored at a time just prior to $t_{(j)}$, called the population at risk, and $x_{(j)}$ is the vector of the explanatory variables for an individual who is observed to have died at the jth order death time.

2.2.1 Model validation method for Cox regression

After a model has been fitted to an observed data set, the adequacy of the fitted model needs to be examined. Residuals are one of the commonly used model checking procedures which are based on quantities for each individual. A number of residuals plots have been adopted in the analysis of survival data. e.g. Cox-Snell residuals, Martingale residuals and Deviance residuals.

The most widely used residuals for the Cox model are the Cox-Snell residuals. It is not similar to the residuals in linear regression analysis, however, since Cox-Snell residuals are not symmetrically distributed about zero, as they cannot be negative. Alternatively, Martingale residuals are derived from the modified Cox-Snell residuals and take values between -∞ and unity. Grambsch and Fleming (1990) give a comprehensive description of the Martingale approach to the analysis of survival data.

A major weakness of plots based on residuals is that there is no quantitative guideline on what constituents a good enough fit.

2.2.1.1 Residual calculations for Cox regression model

2.2.1.1.1 Cox-Snell Residual

The Cox-Snell residual is the most widely used residual in the analysis of survival data and is given by Cox and Snell (1968). For the *i*th individual, i = 1,2,...,n, it is given by

$$\mathbf{r}_{c_i} = \mathbf{exp}(\hat{\beta}'x_i)\hat{H}_0(t_i)$$

where $\hat{H}_0(t_i)$ is the estimated cumulative baseline hazard function at time t_i.

If the fitted model is correct, the Cox-Snell residuals have approximately a unit exponential distribution. Let $\mathbf{r}_{c_i}$ denote the Cox-Snell residuals and $\hat{\mathbf{S}}(\mathbf{r}_{c_i})$ the Kaplan-Meier estimate of the survivor function using the residuals $\mathbf{r}_{c_i}$. If the plot of log{-log$\hat{\mathbf{S}}(\mathbf{r}_{c_i})$} against log($\mathbf{r}_{c_i}$) is a straight line with unit slope and zero intercept, this indicates that the fitted survival model is correct.

2.2.1.1.2 Modified Cox-Snell residuals

Censored data leads to residuals that cannot be regarded on the same footing as residuals derived from uncensored data. The Cox-Snell residual needs to be modified taking into account the censorship , Collett (1994), chapter 5.

The Cox-Snell residuals can be modified by the addition of a positive constant Δ. Therefore modified Cox-Snell residuals have the form

$$r'_{ci} = \begin{cases} r_{ci} & \text{for uncensored observations,} \\ r_{ci} + \Delta & \text{for censored observations,} \end{cases}$$

where r_{ci} is the Cox-Snell residual for the ith individual and it is suggested that Δ is taken to be unity, this leads to the modified Cox-Snell residuals

$$r'_{ci} = \begin{cases} r_{ci} & \text{for uncensored observations,} \\ r_{ci} + 1 & \text{for censored observations,} \end{cases}$$

The modified Cox-Snell residuals can be written as $r'_{ci} = 1 - \delta_i + r_{ci}$, where δ_i is a censoring indicator, which takes the value zero if the observed survival of the ith individual is censored and unity if it is uncensored.

2.2.1.1.3 Martingale residuals

The modified residuals r'_{Ci} have a mean of unity for uncensored observations and this can be relocated to have a mean of zero when an observation is uncensored. In addition, when multiplied by -1, this gives new residuals which are known as Martingale residuals as

$$r_{Mi} = \delta_i - r_{Ci}.$$

Fleming and Harrington (1991) gave a comprehensive account of the Martingale approach. Martingale residuals take values between $-\infty$ and unity, with the residuals for censored observations, where $\delta_i = 0$, being negative. However, the Martingale residuals are not symmetrically distributed about zero. Plots of the residuals against the survival time or the rank of the survival time can be used to detect departures from proportional hazards. Plots of the residuals against explanatory variables in or out of the model indicate whether the variables needs to be included or whether it is necessary to transform a variable that has already been included in the model. If the plot does not show any particular residuals that stand out from the rest, this confirms that the selected model is satisfactory.

2.2.1.2 Testing for time dependence of variables

Validating the model adequacy is important but the proportional hazard assumption itself also needs to be examined. If the hazards for the different patient categories were not proportional over time, the linear component of the model would become time-dependent. The time dependency can be tested by introducing time parameters into the model and checking the significance level for interactions between time and the covarites, Collett (1994), chapter 5.

According to the Cox proportional hazards model, the mortality hazard at a time t for the ith of n individuals in the study can be written as

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\} h_0(t),$$

where x_{ji} is the value of the jth explanatory variable and does not depend on time, x_j, j = 1,2,...,p, for the ith individual, i=1,2,...,n and $h_0(t)$ is the baseline hazard function. Modifying this model to fit the situation in which some of the explanatory variables are time dependent, the Cox regression model becomes

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}(t)\right\} h_0(t)$$

The relative hazard $h_i(t) / h_0(t)$ will therefore, also depends on time. This means that the model is no longer a proportional hazards model.

2.2.2 Previous studies of analysis using Cox regression

The Cox proportional hazards model can also be used to predict intervals of time in which death is likely to occur for individual patients. ROC curves may be used to display the accuracy of prediction with respect to different thresholds, Ohno-Machado (1997). Williams (1985) used Cox regression to predict the local or regional recurrence of breast cancer after a mastectomy operation. Gore *et al* (1984) predicted the year of death due to breast cancer by defining a threshold which crosses the estimated survival function and also discussed the non-proportionality of the hazard functions of the data. Magee *et al* (1996) investigated the prognostic factor for breast cancer recurrence after surgery and treatment, using Cox regression. A cubic-linear spline model is proposed by Efron (1988) which combins the characteristics of a cubic logistic model and a logistic regression model.

Kay (1977) Stablein *et al* (1981) and Gill and Schumacher (1987) and Pettitt and Daud (1990) highlighted the need to validate the proportional hazard assumption and the use of smoothed Schoenfeld (1982) residuals,. The stability of Cox regression can be tested by the use of bootstrap, Altman and Andersen, (1989). Alternatively, using the bootstrap resampling procedure for model selection in Cox regression, Sauerbrei and Schumacher (1992) and Lagakos (1980) proposed a graphical approach to evaluate the explanatory variables. Tibshirani (1982) demonstrated the powerful features of Cox regression, handling a large number of continuous and categorical prognostic variables, resembling the normal linear regression model to the analysis of survival data. Wei (1992) proposed the accelerated failure time model, this can be an alternative to Cox regression in survival analysis. Schoenfeld (1980), Andersen (1982) and Lin and Wei

(1991) tested the goodness of fit of Cox regression and also Arjas (1988) tested it using a graphical method.

Christensen (1987) demonstrated the use of prognostic indexes to separate patients into sub groups and rank the groups from high to low risk groups, providing a convenient way to visualise the survivorship of new patients. Prentice (1978) proposed the grouped data version of the Cox regression to handle large grouped survival data with many tied failure times. Prediction of breast cancer recurrence is another area that researchers are interested in, Magee *et al* 1996 and McCready *et al* (2000) demonstrated the use of Cox regression to identify the prognostic factors for breast cancer recurrence. Chen and Schnitt (1998) gave a detailed review of available literature on prognostic factors for patients with breast cancers 1cm and smaller and determined which of these prognostic factors might be of value for the identification of low risk patients with auxiliary node involvement and/or metastatic disease. Different regression models have been used in the analysis of breast cancer survival, Gore *et al* (1984), in which a few variable interaction pairs were found to be significant by these models and the departure from proportionality of hazards in breast cancer was confirmed.

Altman and Lyman (1998) pointed out that many studies are carried out in an effort to find the prognostic factors than explain the variation in prognosis of breast cancer patients. However the quality of these studies is often in doubt, since a good study design and analysis is less favourable for prognostic factor studies than for therapeutic trials, some guidelines are then proposed in this paper for conducting and evaluating prognostic factor studies to ensure the quality of research is improved. Henderson and Patek (1998) also highlighted that the newly discovered prognostic factors for early breast cancer are being used before this information has been properly utilised and little information actually helps in making a therapeutic decision in the management of individual patients.

2.3 Neural Network Model for Survival Analysis

Neural networks are adaptive non-linear models, and are commonly employed by computer scientists and engineers for classification and prediction problems. Some studies have applied neural networks to statistical problems with interesting results. They have been used in survival analysis to model "mortality" and "time to relapse" and claims have been made that they improve upon the accuracy of traditional statistical methods. Neural network models for survival extend the proportional hazards model to release the linearity and time dependence assumptions and they are usually based on the Multi-layer Perceptron network (MLP). Multi-layer networks having either threshold or sigmoid activation functions are generally called multi-layer perceptrons. The Bayesian neural networks has been proposed by MacKay (1992, 1994, 1995) using Bayes' theorem as a principled framework for regularisation of the MLP. This method included a number of important features to over-come over-fitting, also providing a mechanism to inhibit the influence of irrelevant input variables in the model, which as known as Automatic Relevance Determination (ARD).

2.3.1 Neural network Model

Neural networks is the generic title given to universal non-linear function approximation algorithms, characterised by a distributed structure with multiple non-linear processing units. Certain types of neural network structures simulate the associative memory function carried out by networks of neurons in the central nervous system and, historically, neural networks were used to help understand the principles of memory storage in biological nervous systems, as well as to build computational machines that can carry out complex tasks. The original model of a neuron was proposed by McCulloch and Pitts (1943) and consists of a simple threshold activation function. Figure(2.3) shows the structure of a MLP with three layers of nodes, namely the input, hidden and output layers. Input nodes in the input layer represented the explanatory variables. The hidden layer may have many nodes and there may be several such layers, depending on the complexity of the problem. One hidden layer is sufficient to provide a generic non-linear modelling capability, Bishop (1995). The final layer is the output layer which calculates the output of the network, and it too may consist of several nodes.

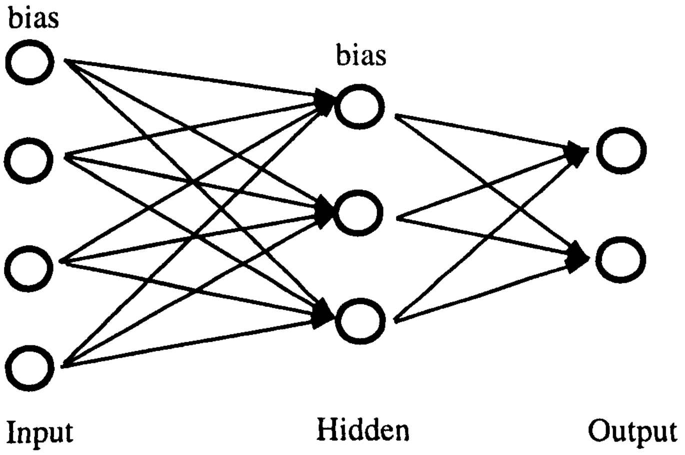


Figure (2.3): The structure of neural network model.

Feed-forward neural networks have one-way connections, from the input layer towards the output layer, with no feedback connections permitted. Each connection has an adjustable strength, called the connection weight. Each observation consists of a unique input signal and the corresponding desired response (target). The network is presented with the training sample and the network parameters, weights and bias, are modified so as to minimise a global objective function that is intended to match the network's response to the desired response, or target value. The training of the network is repeated until the network reaches a steady state, where the changes to the network weights are vanishingly small, or until pre-set value of the objective function is achieved, which is know as early stopping.

2.3.2 Activation functions

The universal approximation property of neural networks is contingent upon the use of non-linear activation in the hidden units. These functions take-in the signal received from the proceeding layer, which is a linear combination of the network activation there and outputs a non-linear function of this scalar variable.

2.3.2.1 Sigmoid Function

Sigmoid function is one of the most common form of activation used in the construction of artificial neural networks. It is a saturating, monotonic exponential function, given by

$$g(a) \equiv \frac{1}{1+\exp(-a)},$$

where a is the slope parameter, which is a linear sum of the weights and the output of previous layer. By varying the parameter a, sigmoid function of different slopes can be obtained. A sigmoid function assumes a continuous range of values from 0 to 1.

2.3.3 Learning by Error Back Propagation

The terminology of back propagation was used to describe the mechanism of optimise the network weights according to the value of network outputs and the desired target. Let there are *d* input units, *M* hidden units and *c* output units. The explanatory variables feed into the input layer, through the hidden layer to the output layer, the output of the *kth* output unit can be written as

$$y_k = g\left(\sum_{j=0}^{M} w_{kj} g\left(\sum_{i=0}^{d} w_{ji} x_i\right)\right),$$

where $g(\cdot)$ is transfer function and they both are sigmoid function when working with classification problem..

'Learning' is the term used to demote updating the network parameters, usually by minimising an objective function, E. Gradient descent is one of the simplest network optimisation procedures, starting with small random values w^0. The parameter w is updated at each step τ, using slope of the error by an amount

$$\Delta w^{(\tau)} = -\eta \nabla E^n |_{w(\tau)}.$$

The parameter η is called the learning rate and it is a gain parameter used to stabilise the learning process. If it is too large, the algorithm may overshoot the minimum, given by $\nabla E = 0$, leading to an increase in E and possibly into divergent oscillations, which may cause a complete breakdown in the algorithm. Or alternatively, the search proceeds extremely slowly which is computationally expensive. The learning rate is problem dependent and it can be adjusted manually to smooth out convergence. An alternative procedure is to use the method of scaled-conjugate gradients (SCG), MØllar (1993b), which adopts the principle of line search. SCG estimates the position of the minimum along a series of mutually orthogonal direction. It searches each direction in weight space in turn and adjusts the step length automatically along that direction. It is possible to choose the step size in the conjugate gradient algorithm without having to evaluate the Hessian matrix, which is computationally expensive.

Scaled conjugate gradient, is an alternative parameter optimisation algorithm, which reduces the number of evaluations of error function required for convergence, and avoids the need to specify the learning gain.

2.3.4 Error function

The error function measures the difference between the network outputs and the desired target values. For the classification problem, the cross-entropy error, Hopfield (1987) function is commonly used. For a particular class problem, let y be the posterior probability of $p(C_1 \mid x)$ belonging to the class. The posterior probability of not-belonging to the class is then $p(C_2 \mid x) = 1 - y$. Then target labeling t for the Class 1 is 1 and 0 for the class 2. Therefore, the probability of either target value is

$$p(t \mid x) = y^t (1-y)^{1-t},$$

which defaults to y if t=1, and (1-y) if t=0. For n independent classes, and the form of the error function is a penalised log-likelihood

$$E = -\sum_{k=1} \{t^k \ln y^k + (1-t^k)\ln(1-y^k)\}.$$

For a multi-layer networks, the error function is typically a highly non-linear function of the weights, in which many minima and saddlepoints may exist and their gradient in weight space is zero, $\nabla E = 0$. The minimum that gives the smallest value of the error function is called the global minimum, the other minima are called local minima. In order to find the minima for the error function, algorithms employ interactive search mechanisms through weigh space typically using gradient descent, of the form

$$w^{(\tau+1)} = w^{(\tau)} + \Delta w^{(\tau)},$$

for which the error function is guaranteed not to increase. The disadvantage for such algorithms is when they reach to a local minimum they may become trapped at saddlepoints, where the error function is flat, the algorithms may be stuck for an extensive period of time. In practice, different values of the initial weights lead to convergence to different local minima.

2.3.5 Early stopping and regularisation

When the algorithms involve a succession of steps and the values calculated within the algorithms are based of the previous values, then a stop point needs to be defined. Otherwise the network would be over trained, leading to data over-fitting (the networks fit the noise as well as the data). However, in reality, the best generalisation performance might be obtained at a local minimum, which is not the global minimum of the error function. Then the generalisation performance needs to be monitored as a function of time during the training, and the training is halted when the optimum generalisation performance is reached, *early stopping* is such a techniques. The error generally decreases as a function of the number of iteration during the course of training. However, the error with respect to the independent data (*validation set)*, often decreases during early training process, but then increases when the network is over trained. Training is stopped at the point when the smallest error is achieved with respect to the independent data, at which the network is expected to produce the best generalisation performance.

Alternatively, adding a penalty term to the error function, Ω, encourages smoother network mapping in the form of

$$\tilde{E} = E + \nu\Omega,$$

where E is an standard error function and the penalty term Ω is governed by the parameter ν the way it influences the form of the error function. When the network gives a good and smooth fit to the training data, it gives a small value to the combined expression for $\tilde{E}$ (although either of E or $\nu\Omega$ may be individually above their minimum possible value).

One of the simplest forms of the regulariser is called *weight decay*, which consists of the sum of the squares of all the adaptive weights in the network, not including biases

$$\Omega = \frac{1}{2}\sum_{i} w_i^2 .$$

Then the cross-entropy error function including the weight decay term is in the form of

$$E = -\sum_{k=1}\{t^k \ln y^k + (1-t^k)\ln(1-y^{k)}\} + \frac{v}{2}\sum_{i} w_i^2 .$$

This is called weight decay because in gradient descent it adds a term to the weight change, that is

$$\nabla\tilde{E} = \nabla E + v\frac{\partial\Omega}{\partial w}, \text{ where}$$

$$\frac{\partial\Omega}{\partial w} = -\eta w .$$

By itself, this term makes the error reduce exponentially to zero, hence the name of this form of regularisation.

2.4 Bayesian framework for network regularisation

The Bayesian framework was proposed by McKay (1992 a,b), in part to address the issue of regularisation. There are a number of important features offered by a Bayesian framework. (1) For regression problem, error bars or confidence intervals can be assigned to the estimated outputs. (2) The regularisation coefficients can be approximated analytically directly from the training data set. (3) Irrelevant input variables, are 'softly pruned' using the technique of *Automatic Relevance Determination* (ARD) (1994a, 1995), whereby, a separate regularization coefficient is given to each input node. If a particular coefficient is large, the corresponding weights will be forced towards zero, so that the corresponding input variable is of little influence upon the network output.

2.4.1 Distribution of the weights

Network training was originally described using maximum likelihood techniques, which minimise the negative log likelihood error function by attempting to find a single best set of values for the network weights. The Bayesian approach treats this differently, by considering a probability distribution function over the weight space, *p(w)*. Once the data *D* have been observed, this can then be transformed to posterior distribution *p(w|D)* by applying Bayes' theorem

$$p(w \mid D) = \frac{p(D \mid w)\,p(w)}{p(D)}.$$

The prior probability distribution for the weights was assumed to be Gaussian distribution.

$$E_W = \frac{1}{2} \| w \|^2 = \frac{1}{2} \sum_{i=1}^{W} w_i^2 \text{ , and}$$

$$p(w) = \frac{1}{\left(\frac{2\pi}{\alpha}\right)^{W/2}} \exp(-\frac{\alpha}{2} \| w \|^2),$$

where *W* is the number of the weights and biases in the network and the parameter α is the regularisation coefficient, called a hyperparameter, controlling the growth of the network weights.

The regularisation produces a penalised log-likelihood cost function, regularised using weight decay

$$E = -\{\sum_{n=1}^{N} [t_n \ln y_n + (1-t_n)\ln(1-y_n)] + \frac{\alpha}{2}\sum_{i=1}^{W} w_i^2\}.$$

2.4.2 Automatic relevance determination

We assumed the weight distribution as a single Gaussian distribution. But commonly the weights fall into a few distinct classes. Weights from different classes should be modelled with different prior by assuming a Gaussian prior for each class. Now each class has its own hyperparameter α_c. The error function of regularisation becomes

$$E = -\{\sum_{n=1}^{N} [t_n \ln y_n + (1-t_n)\ln(1-y_n)] + \frac{1}{2}\sum_{c=1}^{C} \alpha_c \sum_{i=1}^{W} w_{ci}^2\}$$

When presenting a large amount of input variables to the network and some of them are irrelevant to the network output. Any conventional neural network will fail to set the coefficients of these inputs to zero. As a consequence, a finite data will show random correlation between inputs and output.

This problem can be overcome by introducing multiple weight decay constants α_c, one for each input node. When an input variable corresponds to a large value of α_c, its value will be depressed towards zero, making it an irrelevant input. This helps to avoid causing significant overfitting.

2.4.3 Marginalisation

When making an assumption of a Gaussian distribution of the weights, there will be a contribution from the Gaussian noise to the network output distribution. For a classification problem, the logistic sigmoid function of the form

$$y = g(a) \equiv \frac{1}{1+\exp(-a)}$$

is chosen to be the activation function of the output layer since it allows the output to be interpreted as the probability $P(C_1 \mid x)$ of an input vector belonging to class x. As a consequence of the sigmoid activation function, the network output no longer can be approximated linearly by the network weights. Mackay (1992b) introduced a necessary modification to the network output, which is marginalisation.

He assumes the activation a in the sigmoid function is locally a linear function of the weights and since the posterior weight distribution is Gaussian, the distribution of a will be Gaussian. The mean and variance of this Gaussian distribution can be evaluated and gives

$$p(a \mid x, D) = \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{(a-a_{MP})^2}{2s^2} \right)$$

where a_{MP} is the most probable value of the activation, given by the usual combination of hidden node responses, and the variance s^2 is given by $s^2(x) = g^T A^{-1} g$, where A is the Hessian matrix and g is the gradient.

It follows that

$$P(C_1 \mid x, D) = \int g(a) p(a \mid x, D) da \,.$$

However, this integral is not analytically tractable, so Mackay (1992b) suggests the evidence approximation, which involves modulating the activation a_{MP} towards zero, corresponding to $P(C_1 \mid x, D) = 0.5$. Hence suggests

$$P(C_1 \mid x, D) \cong g\left(\frac{a_{MP}}{\sqrt{1+\pi s^2/8}}\right).$$

2.4.4 Neural Network Model Handling Censored Survival Data (PLANN)

Hazard function is assumed to be continuous in the proportional hazards model. However, in practice, the survival times are usually rounded to the nearest day, month or year, therefore tied survival times arise, of which the proportional hazard model is unable to handle. Therefore, there is a need for a discrete version of the proportional hazards model and it takes the form

$$\frac{h_i(t)}{1-h_i(t)} = \frac{h_0(t)}{1-h_0(t)} \exp(\beta x_i).$$

When the width of the discrete time intervals becomes zero, this model tends to the proportional hazard model and also assumes that the censoring has occurred after all the deaths at a given time, which resolves the ambiguity of which individuals should be included in the risk set at that death time.

The implementation of this model into a neural network model is straightforward. The input layer is the replication of the explanatory variables for all time intervals for individuals, in which the subject is observed and including the time as a covariate, since the value of βx_i in the algorithm does not change over time. The value of the time covariate is taken to be the mid-value of the time interval. Here, only one target variable is assigned to each individual, which is represented by the event indicator $d_{i.}$ This indicator only takes value of 1 or 0, 1 presented the event of interest happened on that subject at that time interval and 0 otherwise.

The output of the network is posterior probability of death at a given time and the estimated survival function over time for i^{th} individual is given as

$$S_i(t_j) = \prod_{k=1}^{j}(1 - y(t_j)),$$

where $y(t_j)$ is the network output at time j.

By taking the negative logarithm of the likelihood, we obtain

$$E = -\sum_{p=1}^{P}\sum_{i=1}^{n(p)} t_{pi} \ln y_{pi} + (1 - t_{pi})(1 - y_{pi}),$$

that is equivalent to the cross-entropy error function.

This means that the PLANN can be implemented with a standard neural network model without any modification to the neural network structure or the calculation algorithms. This is proposed by Biganzol (1996).

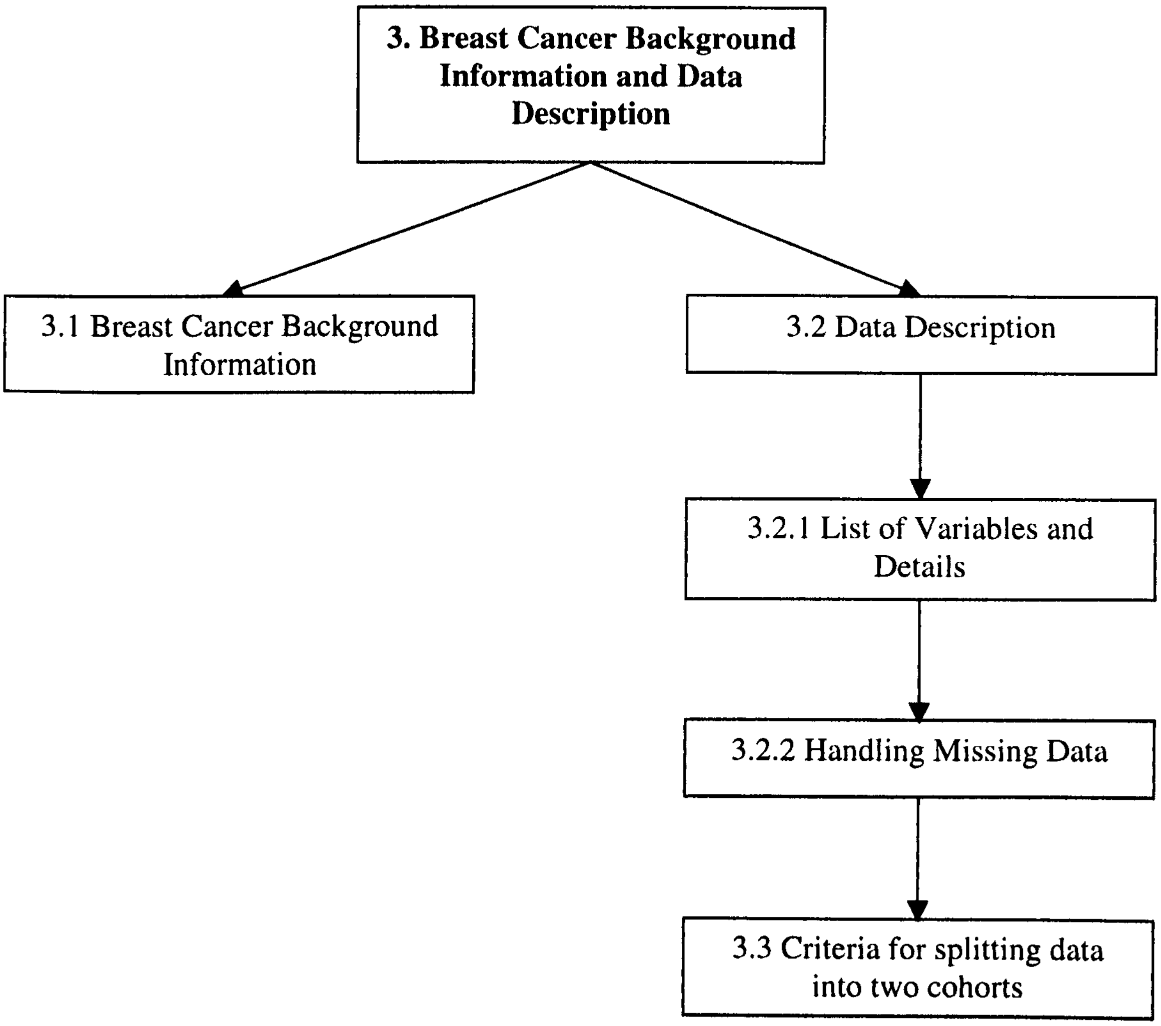
2.4.5 Previous Neural Networks studies of survival

Neural network models have been considered as an alternative tools of conventional statistical methods for survival analysis. At the early stage of development of neural network model for survival data censorship was ignored, Ohno-Machado et al (1995). Faraggi and Simon (1995) demonstrated a possible way to compare the traditional statistical methods with neural network model. Burke *et al* (1997) showed that the MLP predictions produced a better AUROC than simply assigning patients to the averaged survival of the patients in the same TNM stage. However, Brown *et al* (1997) and Radvin and Clark (1992) separately reported that excluding the censored data or treating them as missing will incur substantial bias in the estimation of survival.

Thereby, De Laurentiis *et al* (1994), Ohno-Machado et al (1995), Faraggi *et al* (1997) and Ripley *et al* (1998) alternatively proposed different techniques to handle censorship within the neural network models, which may require several output nodes to maintain the separation between the dependence on time and on the patient specific vector of covariates. A more efficient way to represent the time and using only a single output node is proposed by Radin *et al* (1992), De Laurentiis *et al* (1994) and Liestol *et al* (1994). Biganzoli *et al* (1998) gave a thoroughly description of the Partial Logistic Artificial Neural Network (PLANN) which is a non-linear extension of the discrete version of the proportional hazards model. This neural network model of survival has proved to be stable in monthly studies over a period of time after treatment and releases the proportionality of the hazards assumption and fitting non-linear effects, Laurentiis *et al* (1994), Biganzoli *et al*, (1998), Lisboa *et al*, (2000b).

In term of the interpretation of analysis results, Radvin *et al* (1992) and Christensen (1987) divided patients into three mortality risk groups, low, medium and high, according to their estimated survivorship. While Radvin *et al* (1992), Tarassenko *et al* (1996) and Ripley and Ripley (1998) used the neural network model to predict the recurrence of breast cancer. Groves et al (1999) tested the predictive power of Cox regression and the neural networks according to the area under the corresponding ROC curves by adding and removing factors from the model, which is an application of Acute Lymphoblasitc Leukaemia in children. Mariani *et al* (1997). The neural network model is also used to access prognostic factors for metachronous contralateral breast cancer in terms of model predictive ability Lariani et al (1997), in which variable interactions are also considered, and also Kappen (1993) investigated the prognostic factors for ovarian cancer using multiple neural network models and the Cox regression.

Chapter 3


3. Breast Cancer Background Information and Data Description
3.1 Breast Cancer Background Information
3.2 Data Description
3.2.1 List of Variables and Details
3.2.2 Handling Missing Data
3.3 Criteria for splitting data into two cohorts

3. Breast Cancer Background Information and Data Description

Within this chapter, a data set that is extensively used in this thesis is described in more detail. Two different methods of handling missing data are reported, treating of the missing data as a separate attribute and estimating the category using Nominal Logistic Regression. Also summarising the criteria that most likely to group the data into two subsets representing a low-risk and a high-risk cohort, which are investigated separately.

3.1 Breast Cancer Background Information

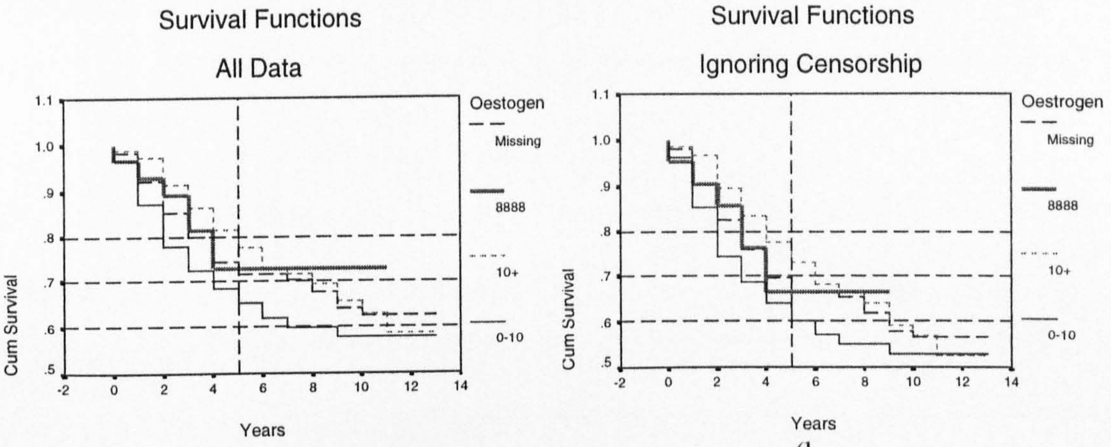
Like other type of cancers, the precise cause of breast cancer and the course of the disease are unknown. Moreover, while breast cancer is often perceived as a single disease, it is in fact a complex variety of diseases that can begin in different types of cells within the breast. It is the leading cause of death in women, whilst it is rarely found in men. Britain has one of the highest mortality rates for breast cancer in the world and 80% of cases occur in post-menopausal women, the UK Breast Cancer Awareness Campaign (1995) claimed. The mortality figures continue to decline due to public awareness of the disease and the development of better treatments, but presently there is still no way of curing the disease.

In general, patients are offered four types of treatments, namely surgery, chemotherapy, radiotherapy and hormone therapy. Treatments are usually tailored to the individual situation, either given alone or in any combination or even in a particular order.

3.2 Data Description

3.2.1 General information of the Data

The analysis techniques developed and reported in this thesis are applied to a data set consisting of 1,616 women breast cancer patients were referred to the Manchester Christie Hospital between 1983 to 1989. All patients were treated and underwent surgery with at least 5 years follow up and in some cases, as long as 13 years. Censorship is an important feature of survival data and cannot be ignored. However, Burke (1995) suggested that ignoring censorship would not significantly affect the survival of the study. Figure (3.1) displays the survival curves of the variable *oestrogen* including (left) and ignoring (right) censored data and concluded that the effect of ignoring censorship is that the calculation of survival is underestimated. Therefore, the event of interest in this thesis is 'death attributed to breast cancer'. All other causes of death and other loss of follow up were regarded as censorship. This is not always clear-cut, since death from unrelated cancers need to be identified and are not assigned to 'the event of interest'. However, in cases of heart attack, for instance it can be difficult to make a clear assignment as this may be related to systemic damage caused by prolonged chemotherapy. For instance, patients who are surviving beyond the time fame for the study are also censored. Since the scope of the study is a five years follow-up, all surviving patients are censored at five years if they survived more than 5 years. Eighteen categorical variables were collected, which can be summarised into 4 categories: 1) demographic information, 2) clinical investigations and 3) laboratory test results as well as 4) treatment received. No family history or genetic link was provided,



(a

(b

Table (3.1) shows a full listing of collected variables.

Figure (3.1): Demonstration of the effect of (a) including and (b) ignoring censorship by grouping data using variable *oestrogen*.

Variable	Categories	Labelling
1. Menopausal status	Pre-menopausal	1
	Peri-menopausal	2
	Post-menopausal	3

Variable	Categories	Labelling
2. Age Group	20 – 39	1
	40 - 59	2
	60+	3

Variable	Categories	Labelling
3. Predominant site (The position of tumour rested in the breast)	Upper Outer	1
	Lower Outer	2
	Upper Inner	3
	Lower Inner	4
	Subareolar	5
	Missing	9

Variable	Categories	Labelling
4. Side	Right	1
	Left	2

Table (3.1): List of variables assessed in each patient.

Variable	Categories	Labelling
5. Maximum Diameter of Tumour (Measured before tumour removal)	<2cm	1
	2-5cm	2
	5+cm	3
	Unknown	9

Variable	Categories	Labelling
6. Clinical stage Tumour (Measurement of tumour after removal)	T0 (No Tumour)	0
	T1 (Tumour less than 2 cm)	1
	T2 (2-5 cm)	2
	T3 (5+cm)	3
	T4 (any size but fixed on the rib cage)	4

Variable	Categories	Labelling
7. Clinical stage Nodes	N0 (cannot feel any node or nodes are negative)	0
	N1 (Tumour has been found under arm and the same side of breast)	1
	N2 (Fixed nodes)	2
	N3 (Nodes are further inside the body and cannot be removed)	3

Table (3.1): List of variables assessed in each patient, continues.

Variable	Categories	Labelling
8. Metastasis stage	M0 (negative)	0
	M1(positive)	1

Variable	Categories	Labelling
9. Clinical stage (also known as the Manchester Stage, it corresponds to different combination of TNM staging)	0	1
	1	2
	2	3
	3	4
	4	5

Variable	Categories	Labelling
10. Type of Surgery	none	1
	Incision Biopsy	2
	Excision Biopsy	3
	Simple Mastectomy	4
	Radical Mastectomy	5
	Wide Local Excision + Ancillary Clearance	6
	Radial Mast + Auxiliary Clearance	7
	Surgery after Neo Adjuvant Chemotherapy	8
	Missing	9

Table (3.1): List of variables assessed in each patient, continues.

Variable	Categories	Labelling
11. Adjuvant Radiotherapy	No	1
	Yes	2

Variable	Categories	Labelling
12. Adjuvant Treatment (Summarised different type of drugs, including chemotherapy and hormonetherapy)	none	0
	CMF	1
	MELPH	2
	TAM	3
	XRAM	4
	OOPH	5
	CYCLO	6
	TAM + CYC	7
	TAM + PRED	8
	ZOLADEX	9
	TAM + ZOL	10
	MEGACE	11
	ZOL + TAM + CMF	12
	NEO ADJ-PRE SURG	13
	CMF + TAM	14
	FAC	15
	Missing	9999

Table (3.1): List of variables assessed in each patient, continues.

Variable	Categories	Labelling
13. Histology	INF DUCT	1
	INF LOB / LOB IN SITU	2
	IN SITU / MIXED / MEDULLARY / UCOID / PAPILLARY / TUBULAR / OTHER MIXED IN SITU	3
	Missing	9

Variable	Categories	Labelling
14. Number of Nodes Involved (no. of nodes have been defined as tumour)	0	1
	1-3	2
	4+	3
	98 (too many to count)	4
	Missing	5

Variable	Categories	Labelling
15. Number of Nodes Removed (no. of nodes have been removed)	0 – 9	1
	10 –19	2
	20 +	3
	98 (too many to count)	4
	Missing	5

Table (3.1): List of variables assessed in each patient, continues.

Variable	Categories	Labelling
16. nodes ratio (number of nodes involved / number of nodes removed)	<=20 %	1
	20-30%	2
	30-60%	3
	60%+	4
	Missing	5

Variable	Categories	Labelling
17. Pathological Size	<2cm	1
	2-5cm	2
	5+ cm	3
	Missing	4

Variable	Categories	Labelling
18. Oestrogen Cytosol	0 – 10 (negative)	1
	10+ (Positive)	2
	8888 (Positive)	3
	Missing	4

Table (3.1): List of variables assessed in each patient, continues.

3.2.2 Missing data

Missing data are inevitable when collecting such a large scale cohort. In this data set, some records contain several missing variables, for example*: number of nodes involved* (968 missing), *oestrogen* (537 missing) and *pathological size* (452 missing). There are only 447 complete cases histories. In some clinical studies, incomplete data was discarded completely if the numbers were sufficiently small, Collett (1994). With this data set, the majority of missing data cannot be discarded and the cause of missing data is unknown. We do not know whether the data are missing at random, missing completely at random or missing but informative.

Two different methods of handling missing data are reported in this thesis. The first, missing data was gathered as a separate attribute, which is the simplest method to use. The second, missing data was estimated using nominal logistic regression, which is appropriate for categorical data. The process of filling in the missing data using nominal logistic regression included two parts. Firstly, using the chi-square test, to infer the relation of the complete variables and the incomplete variables. Then by determining the a subgroup of variables (predictor variables) from the compete variables, which is significantly related to the incomplete variables. Secondly, fitting the model (predictor variables) using the nominal logistic regression, which produces a set of log ratios of the possible categories with respect to the reference category of the incomplete variable. From these values, the category value of missing data can be determined. Altogether 4 incomplete variables were introduced to the nominal logistic regression and table (3.2) displays their determined predictor variables and the results are summarised in table (3.3).

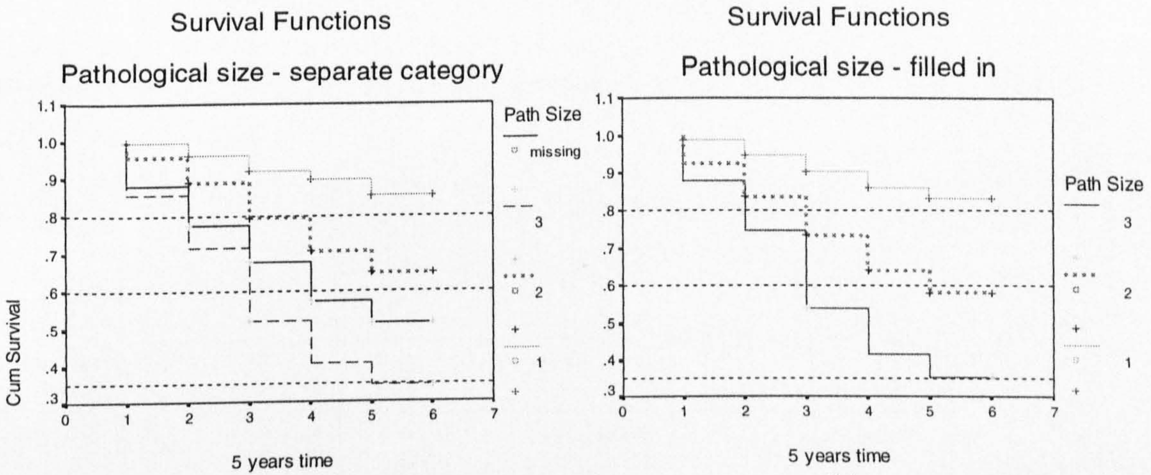
Incomplete variables	**Predictor variables**
Pathological size	*Tumour stage, Predominant site, Surgery, Histology, Adjuvant Treatment ,Node stage*
Number of nodes involved	*Adjuvant Radiotherapy, Manchester stage, Surgery, Adjuvant treatment, predominant site, Histology*
Number of nodes Removed	*Adjuvant Radiotherapy, Predominant site, Histology, Surgery, Metastasis stage*
Oestrogen	*Age group, Clinical stage, Histology*

Table (3.2): The incomplete variables and their predictor variables.

Variables	Estimated Category	Number of records
Pathological size	1(<2cms)	106
	2(2-5cms)	278
	3(5+cms)	0
Number of nodes involved	1(0)	623
	2(1-3)	200
	3(4+)	59
	4(98)	1
Number of nodes removed	1(0-9)	24
	2(10-19)	1
	3(20+)	0
	4(98)	0
Oestrogen	0-10	173
	10+	41
	8888	13

Table (3.3): Summarised the estimated values for each the incomplete variables.

Filling-in the missing data allows the whole data set to be used for data analysis. A separate category was used for missing data and as consequence none reduced the degrees of freedom. If an inappropriate method were used that introduces significant bias to the prediction, the analysis would also be inaccurate. So far, there is no definite solution available for the categorical missing data; therefore the missing data in this data set needed to be handled carefully. Figure (3.2) displays the survival curves of *pathological size* where the missing data are treated as a separate category and filled in using the nominal logistic regression, respectively. As a result, more than 60% of the missing records were estimated to belong to category 2 and the rest were assigned to category 1, which explained the substantial changes that happened to the survival curve of category 2. If the data is missing at random, the survival curves should not show substantial changes after being filled-in. However, the substantial difference between the two plots in figure (3.2) suggests the missing mechanism may be informative. Therefore, the development of survival analysis techniques in this thesis were based on the use of a separate category for the missing data, which minimizes the bias introduced to the analysis if the filling in method turns out to be inappropriate.



(a) (b)

Figure 3.2: Showing the survival curves of *pathological size*. (a): Treating the missing data as separate category and (b): Filling in the missing data using nominal logistic regression.

3.3 Splitting data into low and high-risk cohort

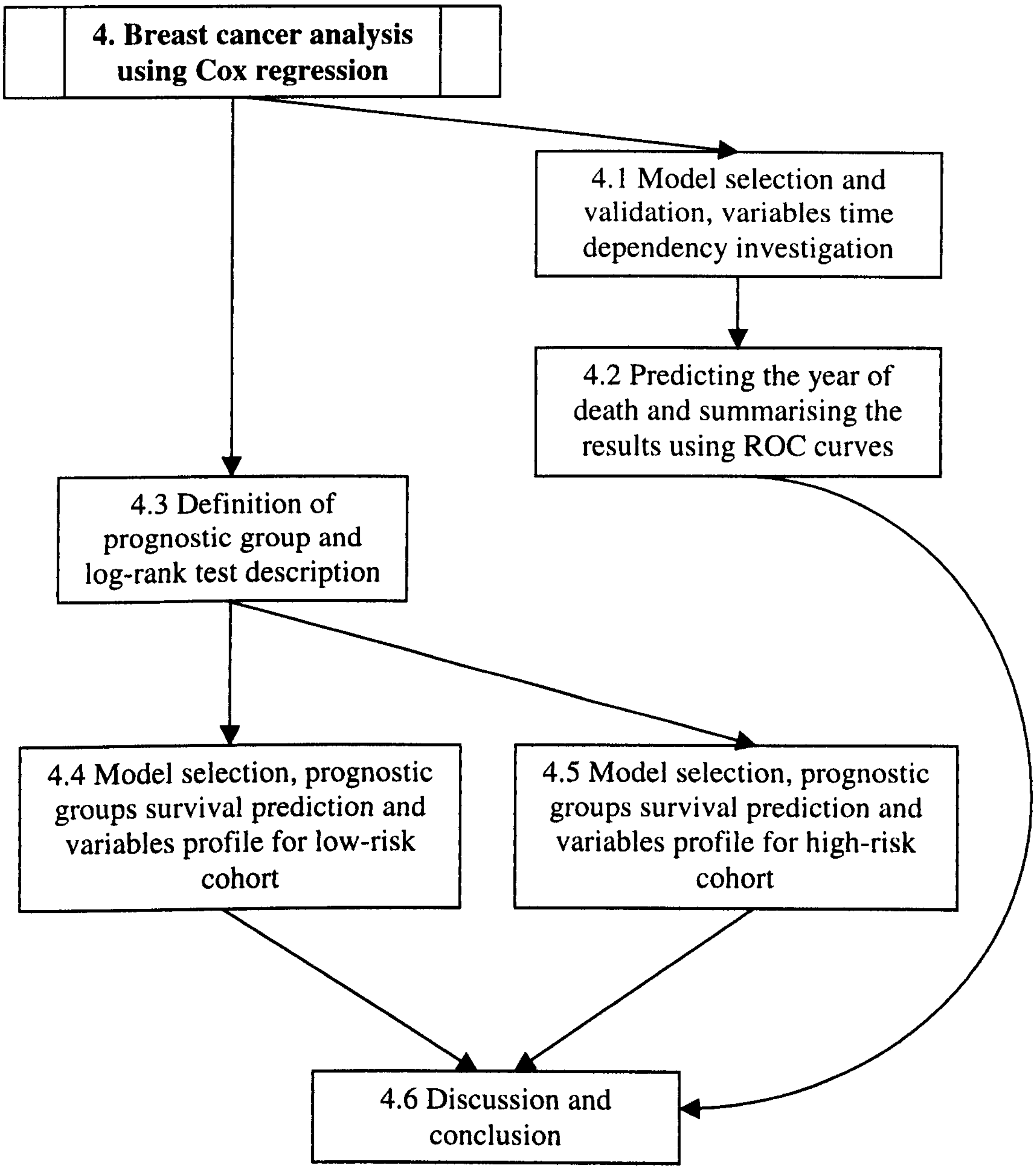
The data analysis in this thesis was based on using the entire data set to predict the year of death for individuals and splitting the data into two parts, low-risk and high-risk cohorts, which allows precise analysis to be conducted in each cohort. In each cohort, an estimated survival function over a fix time period was calculated for each individual, thus grouping the patients into prognostic groups in mortality risk order.

Variables	**Attributes Value**	**Attributes Value**
Metastasis	0	
Tumour stage	1	2
Pathological size	<2cms	2-5cms
Node stage	0	1

Table (3.4): List of variables that contributed to the low-risk cohort separation criteria and their values using clinical staging methods.

The low-risk cohort separate criteria are summarised in table (3.4). The patients in the low-risk cohort are at the early stages of the disease. The rest of the records were regarded as the high-risk cohort. Therefore the numbers of subjects in the low-risk and high-risk cohort are 917 and 633 records, respectively. A total of 66 records were discarded owing to the tumour stage being assigned a value of 0, which appears to indicate that no tumour is present. The low-risk cohort comprises the majority of patients.

Chapter 4


4. Breast cancer analysis using Cox regression
4.1 Model selection and validation, variables time dependency investigation
4.2 Predicting the year of death and summarising the results using ROC curves
4.3 Definition of prognostic group and log-rank test description
4.4 Model selection, prognostic groups survival prediction and variables profile for low-risk cohort
4.5 Model selection, prognostic groups survival prediction and variables profile for high-risk cohort
4.6 Discussion and conclusion

4 Breast cancer survival analysis using Cox regression

This chapter investigates two possible survival analysis approaches. After modelling the expected survival function for individual patients, there are two ways to interpret the results. One is to predict the likelihood of the patient surviving in fixed time intervals, the other is to group patients according to prognostic risk. In this chapter, these two approaches are compared.

The second part of this chapter involves partitioning the data into two groups: a low-risk cohort and high-risk cohort. For each cohort, prognostic groups are identified by means of a ranked mortality risk score of individuals, hence predicting the survivorship over 5 years or 60 months for each group. The survival prediction based upon Cox regression is compared with the observed survivorship which is described by the Kaplan-Meier survival estimate.

4.1 Cox Regression analysis of the whole data set

4.1.1 Model selection

One of the important applications of Cox regression (1972) is to identify variables that may be of prognostic importance. The approach adopted here for the choice of variable to be included in the model is the forward selection stepwise procedure, which was applied to the 1,616 records and the analyis is based on a yearly basis over 5 years and on a monthly basis over 60 months. Variables were added to the model one at a time and assessed as to whether they significantly made improvements to the goodness of fit value to decide which variable to include in the model, Collett (1994). Subjects who survived more than 5 years were viewed as being censored at year 6. A total of 8 variables were selected from the original 18 variables. All of the variables were converted to categorical format. For those variables contained large amount of missing data, the missing data was treated as a separate category; otherwise, the records were removed when the missing data of the variable is significantly small. Therefore, out of the 1,616 records, 120 cases were removed, in which 115 records of missing data and 5 cases of non-positive survival times, leaving 1,496 cases for analysis. From these 1,496 records, 503 patients died of breast cancer 5 years after surgery and are thus regarded as 'event cases', with the remaining of 993 records being viewed as censored data. The threshold of p-value for the acceptance of a variable is ≤ 0.05 and $p > 0.1$ for removal which is the default setting of Statistical Package for the Social Sciences (SPSS). At each model selection stage, there may be more than one variable made significant to the test statistic, only the most significant variable was selected to be included in the model. Table (4.1) summarises the variables entering the model together with the closest alternative variables at each stage.

Finally, eight explanatory variables were selected, namely, *pathological size, node stage, histology, surgery, age group, number of nodes involved* and *oestrogen.*

Variables in the model	**Close alternatives** **f(p value less than 0.05)**
Diameter	Manchester Stage, Pathological Size, Predominant Site, Age Group, Histology, No. of nodes involved, Stage T, Stage M, Node Stage, Surgery, Oestrogen
Diameter + Manchester Stage	Oestrogen, Age Group, Histology, Menopausal Status, No. of nodes involved, Pathological Size, Node Stage, Tumour Stage, Surgery
Diameter + Manchester Stage + Oestrogen	No. of nodes Involved, Age Group, Histology, Pathological Size, Node Stage, Tumour Stage, Surgery
Diameter + Manchester Stage + Oestrogen + No. of nodes involved	Pathological Size, Surgery, Age Group, Histology, Node Stage, Tumour Stage
Diameter + Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size	Age Group, Histology, Node Stage, Tumour Stage, Surgery
Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size	(Diameter is removed)
Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size + Surgery	Histology, Age Group, Node Stage
Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size + Surgery + Histology	Age Group, Node Stage, Tumour Stage
Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size + Surgery + Histology + Age Group	Node Stage, Tumour Stage
Manchester Stage + Oestrogen + No. of	Tumour Stage

nodes involved + Pathological Size + Surgery + Histology + Age Group + Node Stage	
Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size + Surgery + Histology + Age Group + Node Stage + Tumour Stage	
Manchester Stage + Oestrogen + No. of nodes involved + Pathological Size + Surgery + Histology + Age Group + Node Stage	(*Tumour Stage* is removed)

Table (4.1): Cox regression model selection of the breast cancer data.

4.1.2 Model validation

The Cox-Snell residual calculation is one of the most commonly used methods for model validation, Collett (1994). Figure (4.1) displays the modified Cox-Snell residuals plot of the entire data. The graph of the residuals is shown an approximately a straight line with unit slope and zero intercept, indicating no real evidence against the fitted model being adequate.

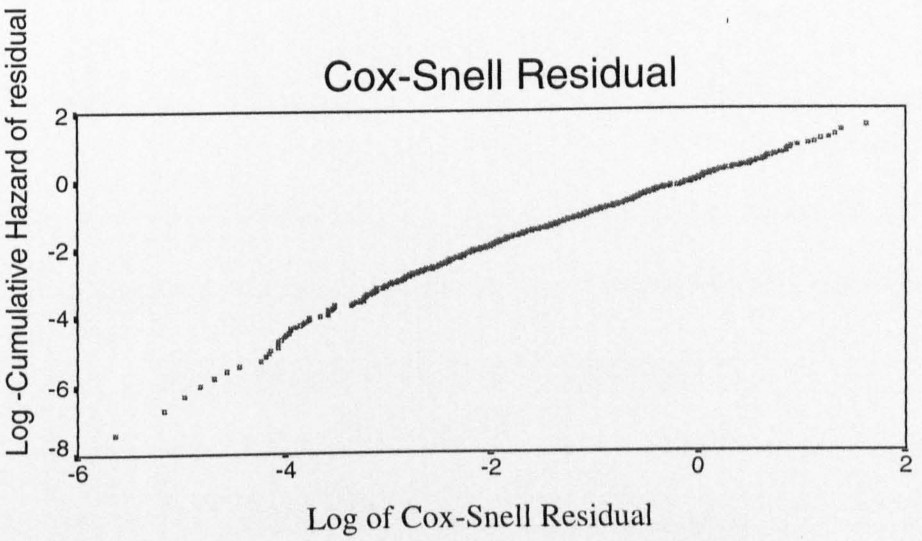


Figure (4.1): Plot of Cox-Snell residuals of breast cancer data. It appears as a straight line with unit slope and zero intercept, indicating no real evidence against the fitted model being adequate.

The use of Martingale residuals is an alternative model validating method, Collett (1994). Figure (4.2) shows plot of the Martingale residuals against the survival time while figure (4.3) shows the plot of the Martingale residuals plot against rank of survival time. Both of the graphs display no discernible pattern in the residuals over time, with only two residuals indicated as potential as outliers, again suggesting no real evidence against the fitted model being adequate. The Martingale residuals can also be plotted against the explanatory variables, since the variables are converted into categories, therefore, there is not much information that can be extracted in this case.

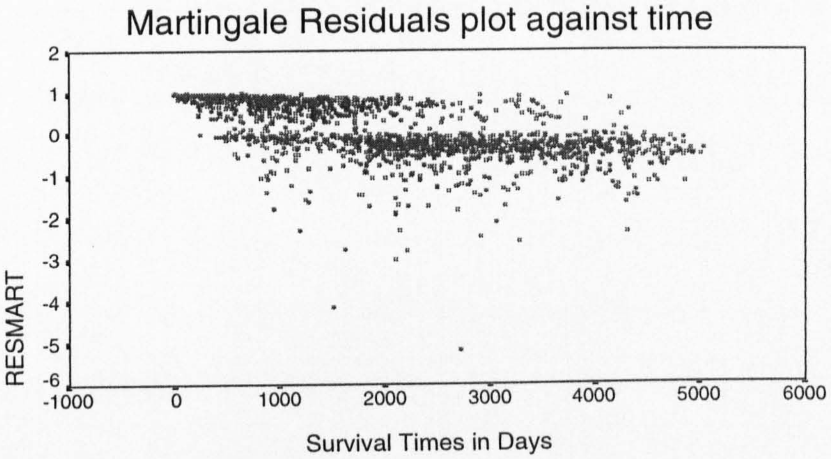


Figure (4.2): Martingale residuals versus survival time. There is no pattern in the residuals over time, indicating no real evidence against the fitted model being adequate.

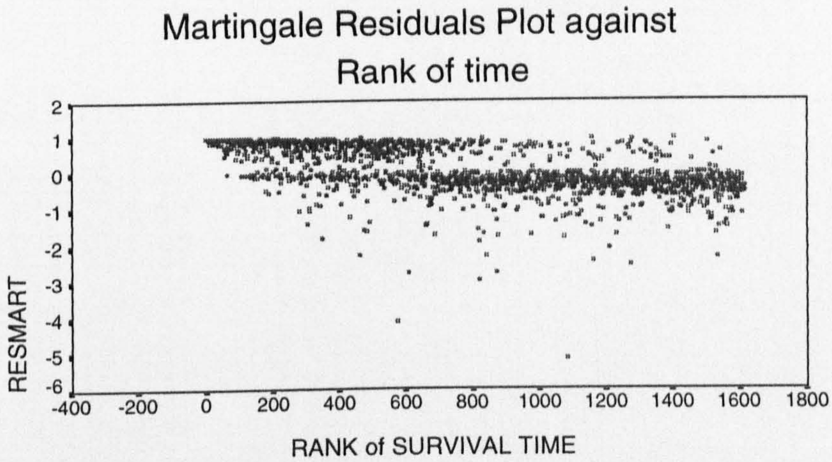


Figure (4.3): Martingale residuals versus rank of survival time. There is no systematic pattern in the residuals over time, indicating no real evidence against the fitted model being adequate.

4.1.3 Assessing the possibility of time dependency of the explanatory variables

This section tests the statistical significance in Cox regression of the time dependency with the previously selected variables. The –2 log-likelihood, $-2\log\hat{L}$, value for the model without time dependent variables was 7114.600 with 27 degrees of freedom. Time dependence was only tested over the first 5 years. The results are summarised in table (4.2) below and indicate that none of the variables display true dependent behaviour when assessed at the 5% level of significance. With the result, there is no evidence to prove that the linear components of the model do not vary with time, indicating that the fitted model is adequate, Collett (1994).

Additional variables added to the model	$-2\log\hat{L}$	**Change of -2 Log Likelihood from the previous model**	**Degree of Freedom**	**p-value**
*Time * Age Group*	7107.198	7.402	35	0.4939
*Time * Histology*	7107.655	6.945	35	0.5426
*Time * Manchester Stage*	7095.575	19.025	43	0.2674
*Time * Number of Nodes Involved*	7102.981	11.619	43	0.7698
*Time * Oestrogen*	7104.520	10.08	39	0.6089
*Time * Pathological Size*	7103.827	10.773	39	0.5484
*Time * Node Stage*	7097.458	17.141	39	0.143

Table (4.2): Significant level for assessing the time dependency of the variables.

4.2 Survivorship Prediction

4.2.1 Prediction of 5 years follow-up survivorship using Cox regression

The survival function of individual patients over a fixed time period can be estimated using Cox regression, as Ohno-Machado (1997) has suggested. The estimated survivor function for the *ith* individual at time t is given by $\hat{S}_i(t)$ where

$$\hat{S}_i(t) = [\hat{S}_0(t)]^{\exp(\hat{\beta}' X_i)},$$

for $t_{(k)} \leq t < t_{(k+1)}$, k = 1,2,...,r-1 of r distinct death times, where $\hat{S}_0(t)$ is the estimated baseline survival function at time t, $\hat{\beta}$ is a vector of unknown parameters and X_i represents the vector of the values of the explanatory variables of the *ith* individual. A 5 years survival curve was produced for each of the subjects. The data set was split into two parts according to whether the record number was odd or even, to produce the training and test set. The $\hat{\beta}$ and the baseline hazard function were estimated using the training set, then applied to the test set for performance evaluation. The training and test split was in line with the neural network approach, allowing a fair comparison of the performance between two methods, where the network parameters are estimated from the training set and the model is applied to the test set. Figure (4.5) shows an example of the estimated survival curves for 10 patients over the 5 year period. By drawing a threshold across the figure at any value on the y-axis, representing the probability of survival, the cross points of the threshold and the survival curves were used to predict the year of death for each patient, which are reflected on the x-axis. The ROC was used to determine the value of threshold giving the most accurate prediction from a range of possible thresholds between 0 and 1. The calculation of the ROC involves the true positive rate (sensitivity) divided by the false positive rate (1-specificity). The sensitivity and specificity sometimes are called the true positive rate and true negative rate, respectively and defined as

$$Sensitivity = \frac{True\ Positive}{The\ number\ of\ positive\ cases},$$

$$Specificity = \frac{True\ Negative}{The\ number\ of\ negative\ cases}.$$

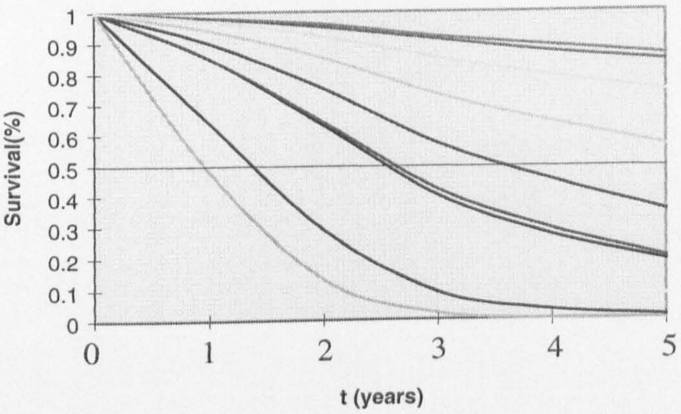


Fig. (4.5): An example of re-estimated survival curves of 10 patients. The horizontal line corresponds to a 0.5 probability of survival as the threshold for the prediction of survival time after surgery for each patient.

The definition of true positive in this study is, that the patient is predicted to die of breast cancer (positive) within a particular time interval and the patient actually does die of breast cancer within the time interval. While the true negative in this study means the patient is being predicted not to die of breast cancer within an time interval and the patient actually does not die of breast cancer within this time interval. An optimal situation would be all of the patients who are to die of breast cancer are predicted to die of breast cancer at the same time interval and all of the patients who are not to die of breast cancer are not being predicted as dying of breast cancer at a particular time interval for some values of the threshold. This is corresponds to the ROC curve passing through the (0,1) point on the graph.

Figure (4.6) displays the ROC curves of prediction death happening up to respectively year 1, year 2, year 3, year 4 and year 5. The calculation is based on taking the difference between the survival functions at each time point and the one before, to obtain the probabilities of death during each year. The time interval that contained the greatest estimated probability of death was interpreted as the predicted year of death. For an example, taking 0.5 as the threshold, the actual year of death is the third year and if the highest estimated probability of death is in the third year band and is greater or equal to 0.5, it is counted as a correct classification.

The results show that the curves reach sensitivity values above 0.6 only for relatively high false negative rates, above 0.2. So this predictive approach is not considered to be very informative, and a new approach for the interpretation of survival models is proposed. This new approach is based on assigning patients into a prognostic risk groups.

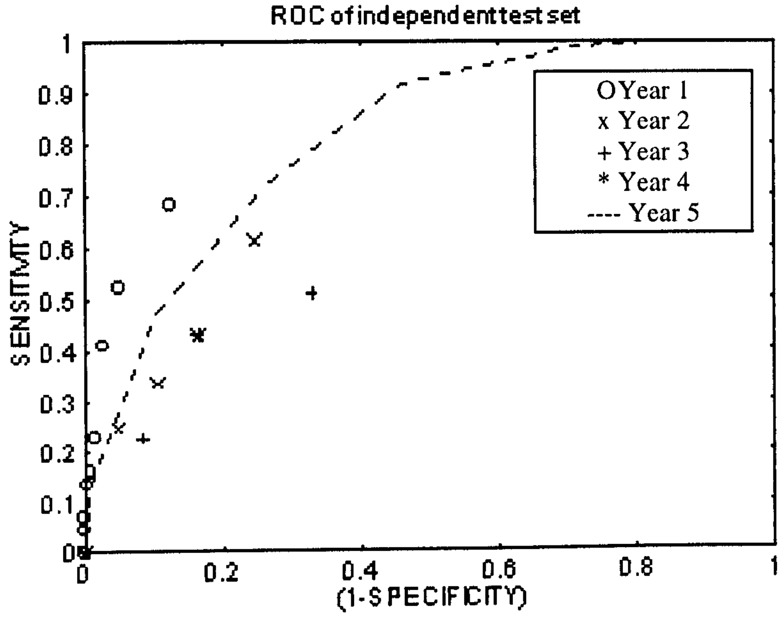


Fig. (4.6): The ROC curves of Cox regression, predicting year of death from breast cancer up to 5 years.

4.3 Prognostic index and Log-rank test

The hazard function for the *i*th individual can be written as

$$h_i(t) = \varphi(x_i)h_o(t),$$

where $\varphi(x_i)$ (>0) is a function of the vector of explanatory variables of the *ith* individual that can be interpreted as the relative hazard compared with an individual for whom x=0. The function $\varphi(x_i)$ is conventionally written as $\exp(\eta_i)$, where

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi},$$

and p is the number of explanatory variables. The quantity η_i is called the prognostic index or risk score for the *ith* individual, Collett (1994).

The prognostic index provides a score for each subject, and can indicate whether the particular patient has a good, intermediate or bad prognosis for survival. Prognostic indexes for a given cohort can be ranked and partitioned into prognostic groups, and their survival curves displayed for each prognostic group. There are several different ways to arrange the prognostic indexes into prognostic groups, by allocating significant amount of samples into each group, Christensen (1987). Or alternatively, by observing the natural distribution of the indexes from the prognostic indexes plot is also considered in this thesis and also use of a well-established statistical method, such as the log-rank test. The log-rank test determines, to a given significance level, whether the population comprises of two subgroups with different survivorship. The disadvantage of the first method is lack of clear guidance about the cut-off point locations. The second method is not convincing, when the scores are crowded and leave no gap between groups, as they are difficult to separate by eye. Finally, the log rank test becomes the preferred method.

The log-rank test was proposed by Peto and Peto (1972). In the two group case, the null hypothesis is that there is no difference in the prognostic scores of the individuals in two groups. The tenability of this hypothesis is tested by considering the difference between the observed number of surviving individuals in the two groups at each time points and the number expected under the null hypothesis.

Let d_{1j} and d_{2j} be the number of deaths at $t_{(j)}$, j=1,2,...r, in group 1 and group 2, respectively, r is the number of distinct death times, and the n_{1j} and n_{2j} be the number of individuals at risk at time $t_{(j)}$ in group 1 and group 2, respectively. Therefore the expected number of individuals e_{1i} who die at time $t_{(j)}$ in group 1 is given by

$e_{1j} = n_{1j} d_j / n_j$,

where $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$.

The overall measure of the deviation of the observed values of d_{1j} from their expected values is calculated by summation of the differences $d_{1j} - e_{1j}$ over the total number of time intervals, r, in the two groups. The test statistic is given by

$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j})$ with the variance of d_{1j} being given by

$v_{1j} = \dfrac{n_{1j} n_{2j} (n_j - d_j)}{n_j^2 (n_j - 1)}$, so that the variance of U_L is $\text{var}(U_L) = \sum_{j=1}^{r} v_{1j} = V_L$.

Note that U_L has an approximate normal distribution when the number of death times is not too small, Collett (1994), so that H$_0$: $U_L / \sqrt{V_L}$ has approximately the standard normal distribution and can be written as

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1).$$

In addition, note that $\frac{U_L^2}{V_L} \sim X_1^2$, where X_1^2 denotes the chi-squared distribution with one degree of freedom.

The larger the value of the statistic $W_L = U_L^2 / V_L$, the greater the evidence against the null hypothesis. A 5% significance level is used here.

4.4 Low-risk cohort analysis of Cox regression

4.4.1 Model selection

A new variable *nodes ratio* was considered at this point. This is the ratio *of the number of positive nodes* to the *number of nodes removed.* The *number of positive nodes* has already been selected as an important variable, see section (4.1.1). Considering that two patients with same number of positive nodes may have different prognoses since clinicians tend to have different prognoses depending upon the total number of nodes that have been removed. For example, a patient who has 5 positive nodes out of 5 nodes removed is more severely affected than a patient who has 5 positive nodes out of 30 nodes removed. The *nodes ratio* variable was designed to take into account this consideration.

By applying the low-risk cohort selection criteria described in section (3.3), a total of 917 cases were selected. A forward stepwise elimination model selection was performed once again as described in section (4.1.1). However, four variables were excluded from the pool of variables in order to identify the prognostic factors which provide information on the survivorship of the patient regardless of treatment. The four variables are namely ***treatment***, ***surgery***, o***estrogen*** and ***adjuvant radiotherapy***. O***estrogen*** is a measurement of female hormone, and the remaining three variables are decided by the doctors according to the symptoms of the patients. The selected variables are ***node stage, nodes ratio, histology and pathological size.*** Akaike's information criterion (AIC), Akaike (1973), was also employed to measure and establish the significance of the value of $-2\log$-likelihood, $-2\log\hat{L}$, on adding new terms into a model or deleting existing terms from the model. The AIC statistic is

$$AIC = -2\log\hat{L} + \alpha q\,,$$

in which q is the number of unknown parameters in the model and α is a predetermined constant that specifies the weighting between fit accuracy, which is measured by $-2\log\hat{L}$, and model complexity and α =3 is recommended for general use. The final model chosen was that with the smallest value of the AIC, and in fact, the final model contained 4 variables from the original 6, namely *histology, pathological size, node stage* and *nodes ratio.* Vonta et al (1998) also used the AIC test statistic to select the best subset of variables to be included in the final Cox model and found that the *lymph nodes, tumour size (pathological size)* and *grade (tumour stage)* have significant impact on the survival times of breast cancer as our model showed. The details are given in table (4.3).

Model selected from SPSS	$-2\log\hat{L}$	**Parameters in model**	**AIC**
Node stage	3734.798	2	3740.798
Node stage + Nodes ratio	3700.972	7	3724.872
Node stage + Nodes ratio + Histology	3680.338	10	3710.338
Node stage + Nodes ratio + Histology + Pathological size	3669.689	12	3705.689
Node stage + Nodes ratio + Histology + Path Pathological size + Age group	3661.818	15	3706.818
Node stage + Nodes ratio + Histology + Path Pathological size + Age group + Diameter	3651.835	19	3708.835

Table (4.3): The AIC measurement for each step of the variable selection process for the low-risk cohort.

4.4.2 Natural distribution of the prognostic groups of the low-risk cohort

The low-risk cohort was split into training (458 records) and test sets (459 records), based upon the odd or even record number. A 5 years analysis was again conducted, therefore all patients who survived more than 5 years were viewed as being censored at year 6. The training and test split process is in line with neural networks analysis for a fair modelling methods comparison. By considering the natural distribution of the prognostic indexes, the test set data were partitioned into 7 prognostic groups as illustrated in figure (4.7). The Cox regression and the Kaplan-Meier estimated survival curves for each of the prognostic groups are displayed in figure (4.8).

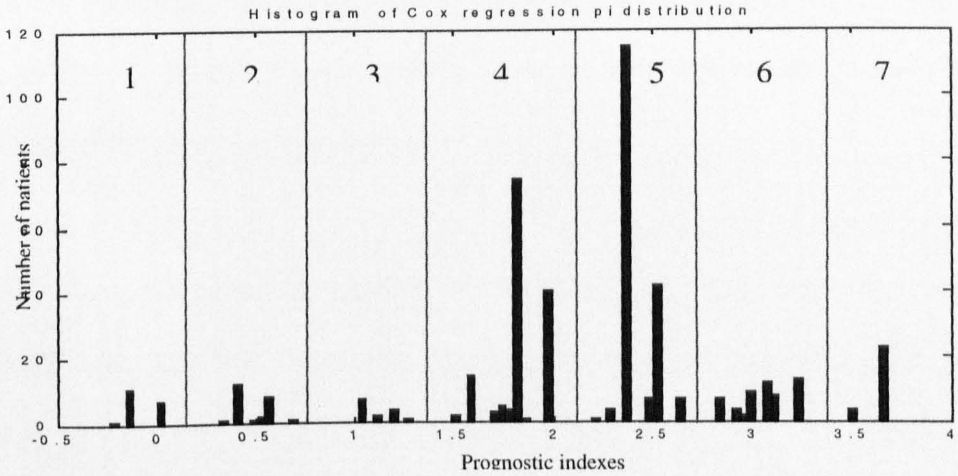
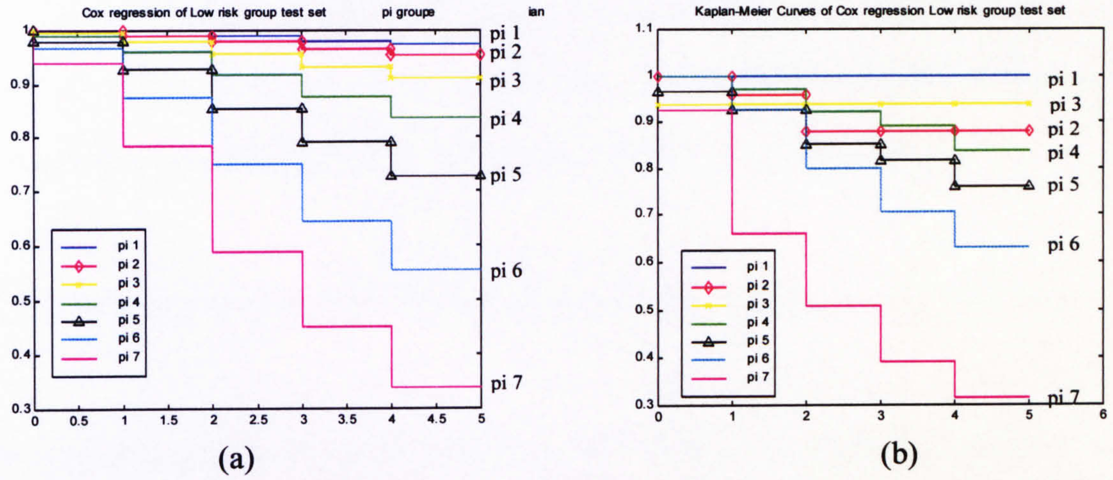


Figure (4.7): Each band represents one prognostic group, labeled from 1 to 7, they are aggregated by observing the natural grouping behaviour of prognostic indexes.



(a) (b)

Figure (4.8): (a) The predicted survivorship for each of prognostic group using Cox regression in the low-risk cohort and (b) the corresponding Kaplan-Meier estimate of the survivor function, which represents the observed survivorship for each group. The results show the estimated performance is acceptable in general by comparing the two graphs, except for groups 2 and 6 which show an over and under estimation of the survivorship of the groups, respectively.

When comparing the estimated survival curves using Cox regression with the Kaplan-Meier estimated survivor functions, which are used to described the observed probability of survival, the accuracy of the survival estimation was varied over the prognostic groups. The smaller the value of the prognostic index, the greater the survival probability of the subject/group will be. In addition the prognostic index has been arranged in mortality risk order, so that the prognostic group1 to 7 are in the order from the highest to the lowest degree of survival. However, the results show that some of the curves overlapped or crossed. This suggests that some of the prognostic groups potentially could be combined.

The natural grouping approach becomes inaccurate when the separation of the prognostic groups is not clear. We thus consider an alternative grouping method and adopted the log-rank test for the partition of the data into prognostic groups.

4.4.3 Partitioning the low-risk Cohort into prognostic groups using Log-rank test

The log-rank test, described in section (4.3), was adopted to replace the visualisation grouping method, in which the survival curves of two groups is compared by measuring the significance level arising out of a test of the equality of the two survivor functions.

The process begins with choosing the cut-off point from the lowest prognostic index value to the highest. The log-rank test was performed separately at each cut-off point; therefore a set of p-values was obtained. The optimal cut-off was chosen at the cut-off point with the highest p-value and the group was split only if this is significant at least at the 5% level. A subset of patients, whose prognostic index was greater than the optimal cut-off point, were removed from the data and regarded as one prognostic group. The whole process was repeated until no more prognostic groups could be defined.

Figure (4.9) displays the four groups obtained via the above log-rank test based approach. The study was conducted with monthly time resolution over 60 months. The performance measure was no longer only applied to the test set, but a 5-fold cross validation was introduced. Furthermore, 95% confidence interval bands were also included with the Kaplan-Meier curves, as this helps to identify the separation between prognostic groups, assessing the uncertainty of the data as displayed in figure (4.10). By displaying the variable profiles of each prognostic group as shown in figure (4.11), the contribution of each variable is monitored, hence illustrating which variables play an important role in each prognostic group.

Note that the number of prognostic groups obtained using this approach is fewer that the number obtained using the previous approach and none of the survival curves crossed over or overlapped. The log-rank test is a well-developed method for survival curves comparison. The results indicate that it is better than visualisation grouping method, the survival curves are well separated and also the estimated survival rate was improved for each group.

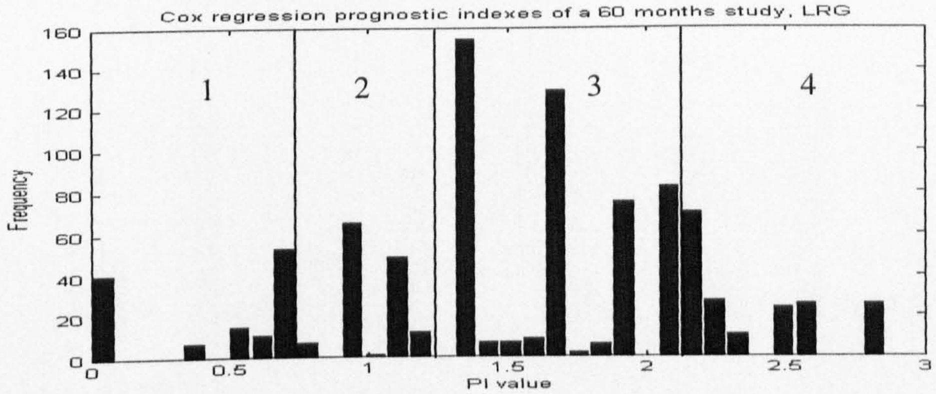
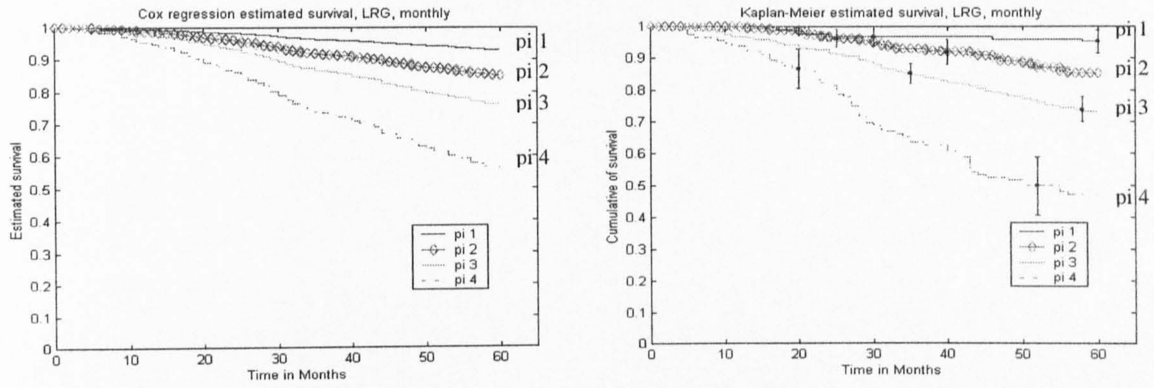
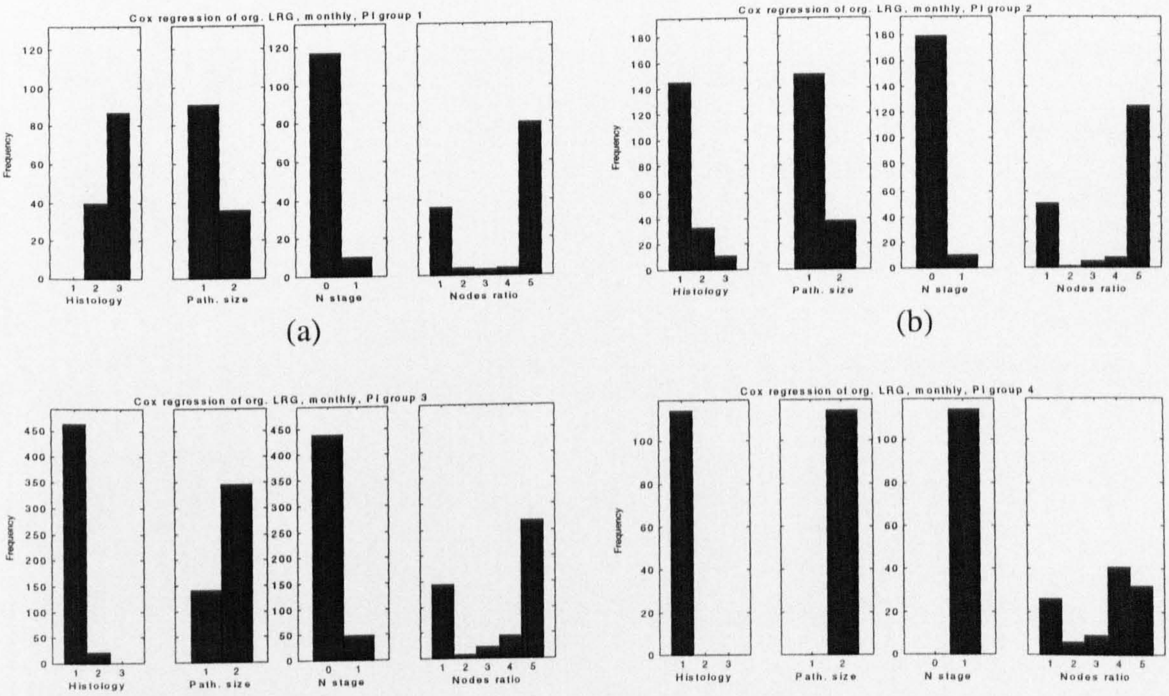


Figure (4.9): A total of 4 prognostic groups were aggregated by the log-rank test and labelled from 1 to 4, contained 127, 189, 487 and 114 patients, respectively.



(a) (b)

Figure (4.10): (a) Each curve (labelled in turn pi 1 to pi 4) represents a Cox regression estimated probability of survival for a prognostic group over 60 months. (b) The corresponding Kaplan-Meier estimate of the survival functions. Each of the Cox regression estimated survival curves fall within their confidence interval bands and thus indicating that the per formance of the Cox regression is acceptable.



(a) (b)

(c) (d)

Figure (4.11): Attribute profiles for the prognostic groups illustrating which variable dominates the prognostic group.

4.5 Cox regression for the high-risk cohort

4.5.1 Model selection

A total of 633 records were left after the low-risk cohort was removed from the design data set, which are regarded as a high-risk cohort. The model selection procedure was repeated with the AIC criterion, as described in section (4.1.1). The selected model comprised the variables *menopausal status, node stage, pathological size, clinical staging* and *nodes ratio. Node stage, pathological size* and *nodes ratio* are again being selected as for the low-risk cohort. The details of each stage in the model selection process and the value of the respective AIC values are displayed in table (4.4).

Vaiables in the model	$-2\log\hat{L}$	**Degrees of freedom**	**AIC value**
Clinical stage	3805.973	3	3814.973
Clinical stage, Pathological size	3789.091	6	3807.091
Clinical stage, Pathological size, nodes ratio	3773.244	10	3803.244
Clinical stage, Pathological size, nodes ratio, Menopausal status	3764.722	12	3800.722
Clinical stage, Pathological size, nodes ratio, Menopausal status, Node stage	3754.329	15	3799.329

Table (4.4): The AIC value of each variable during the model selection process for the high-risk cohort.

4.5.2 Applying the selected model to the high-risk cohort

The analysis was carried out on a monthly bases for 60 months using 3-fold cross validation. Again the log-rank test was adopted to identify the prognostic groups. The results are summarised in the order of figures (4.12) to (4.14). Figure (4.12) displays the log-rank test aggregated prognostic groups from the prognostic indexes, and figure (4.13) the estimated survival curves over 60 months for each prognostic group and the Kaplan-Meier estimate of the survivor functions. The variables profile for each prognostic group is illustrated in figure (4.14).

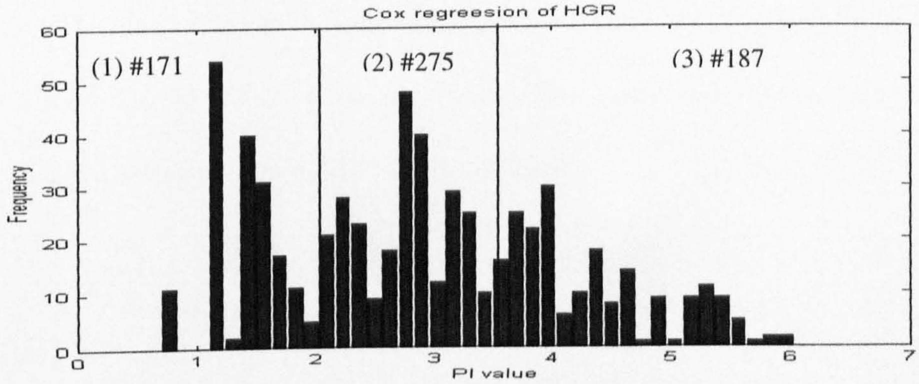
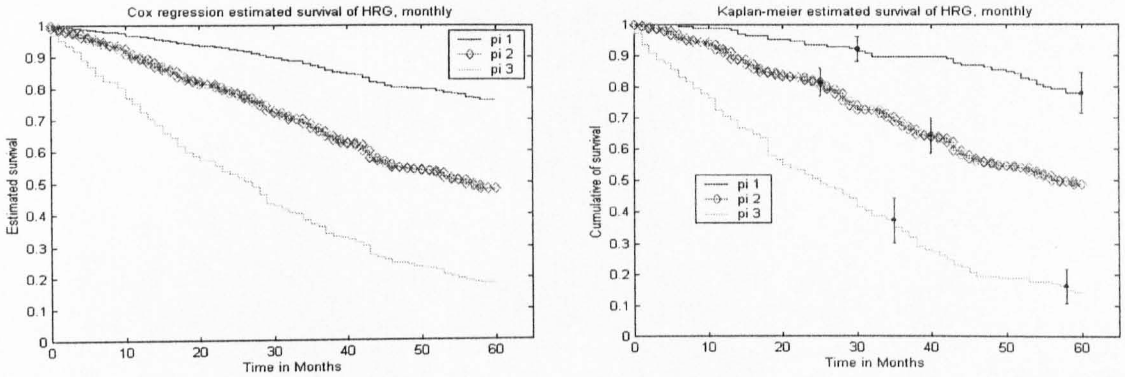


Figure (4.12): The 3 groups obtained via the log-rank test by means of prognostic indexes and labelled from 1 to 3 as illustrated. Their sample size is displayed next to their group labelling. Group 2 contained almost 45% of the patients in the high risk group.



(a) (b)

Figure (4.13) (a): The Cox regression estimated survival curve over 60 months for 3 prognostic groups and labelled as pi1, pi2, and pi3 from top to bottom. Right: Their corresponding Kaplan-Meier estimated survival curves. The curves are well separated on the graph and their Cox survival estimations are within their confidence bands when compared with the observed survival curves. The small confidence interval bands suggests that the variance within each group is small.

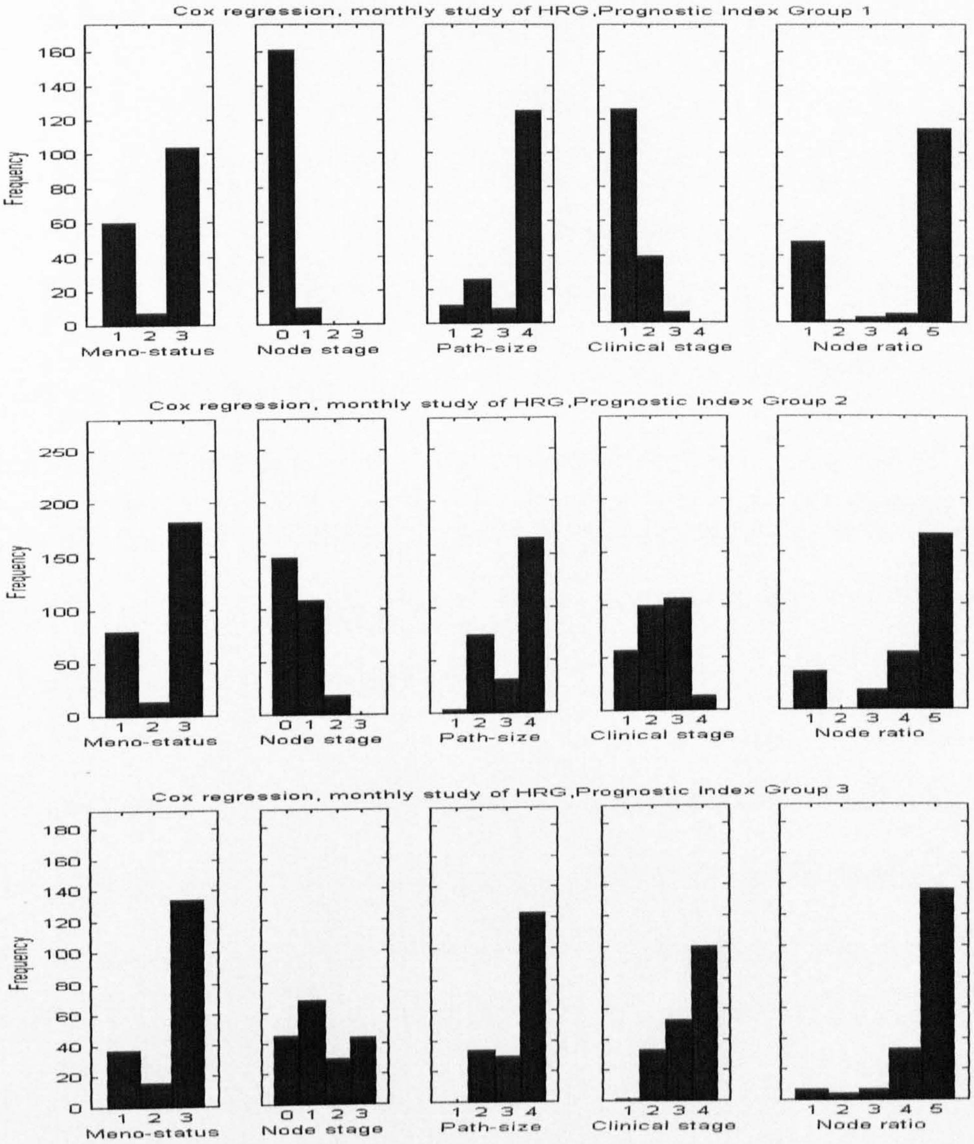


Figure (4.14): A display of the variable profiles within each of the prognostic groups to observe the category shifting behaviour of each variable. The display shows that (except the *menopausal status*) the variables display a different degree of the category shifting movement over the prognostic groups. The category 4 of *pathological size* represents the missing data of the variable.

The display of the prognostic indexes in figure (4.12) has demonstrated again the importance of the log-rank test in this study by providing an objective criterion to assign patients into prognostic risk groups. Figure (4.13) also suggests that the performance of the Cox regression approach in predicting the mean actual survival for each group is acceptable. The major variables dominating the prognostic group 1 and 2 are the *node stage* and *clinical stage*, with the *node stage* changing from category 0 to 1 while the *clinical stage* shifts from category 1 to either categories 2 or 3. The other variables also show some degrees of category shifting behaviour as *pathological size* shifts from category 1 to 2 which corresponds to <2cms and 2-5cms respectively. This shifting sequence is continued into prognostic group 3.

The results show that the survival probability of prognostic group (pi) 1 after 5 years is 0.8. A question raises 'Does this group truly belong to the high-risk cohort?' The *pathological size* was one of the criteria for defining the low- and high-risk cohorts, but itself also contained missing data on 414 out of 1,530 records. As these 414 records did not have confirmed small tumour diameter, they were left in the high-risk cohort. However, 203 records partially fit into the criteria of low-risk cohort on the basis of *tumour stage*, in which subsets of 120, 75 and 8 records were allocated to the prognostic group 1,2 and 3 respectively. These records probably do belong to the low-risk cohort and it is interesting that they were identified as low-risk even using *pathological size*, coding missing value as a separate attribute. The prognostic group 1 in the high-risk cohort may correspond to the prognostic group 2 and 3 in low-risk cohort. Removing the 120 records from prognostic group 1 of 171 records, leaves only 51 records, they might be the true members of prognostic group 1. The subset of 75 records may also correspond to the prognostic group 4 in the low-risk cohort where the survival probability was 0.5.

The results have demonstrated that the approach of predicting the year of death for individuals is not informative. On the other hand, the second approach produced some interesting results, which involved defining prognostic groups and estimating group survivorship.

After completing the analysis for the low-risk and high-risk cohorts using Cox regression, one further interesting point was found. The motivation behind the separation of the data into two cohorts was to try to understand the survivorship of the disease and to enable a precise analysis of each cohort to be made. The above analysis has shown that there was no clean cut-off point for separating the data between low- and high- risk cohorts. There is a group of patients which overlaps the two cohorts.

4.6 Discussion and Conclusion of Cox regression analysis of the breast cancer data

Two analytic approaches are illustrated in this chapter, prediction of 'year of death' and survival prediction for prognostic groups. The unsuccessful attempt of estimating the likely year of death is possibly caused by the large amount of censored data.

The second approach, in which the low- and high-risk cohorts are both further divided into distinct prognostic groups, gives more promising results. The Cox regression predicted survival for the prognostic groups agrees well with the corresponding Kaplan-Meier estimated survivor function, and falls within the confidence bands. This is especially true when the log-rank test is used to partition patients into prognostic groups.

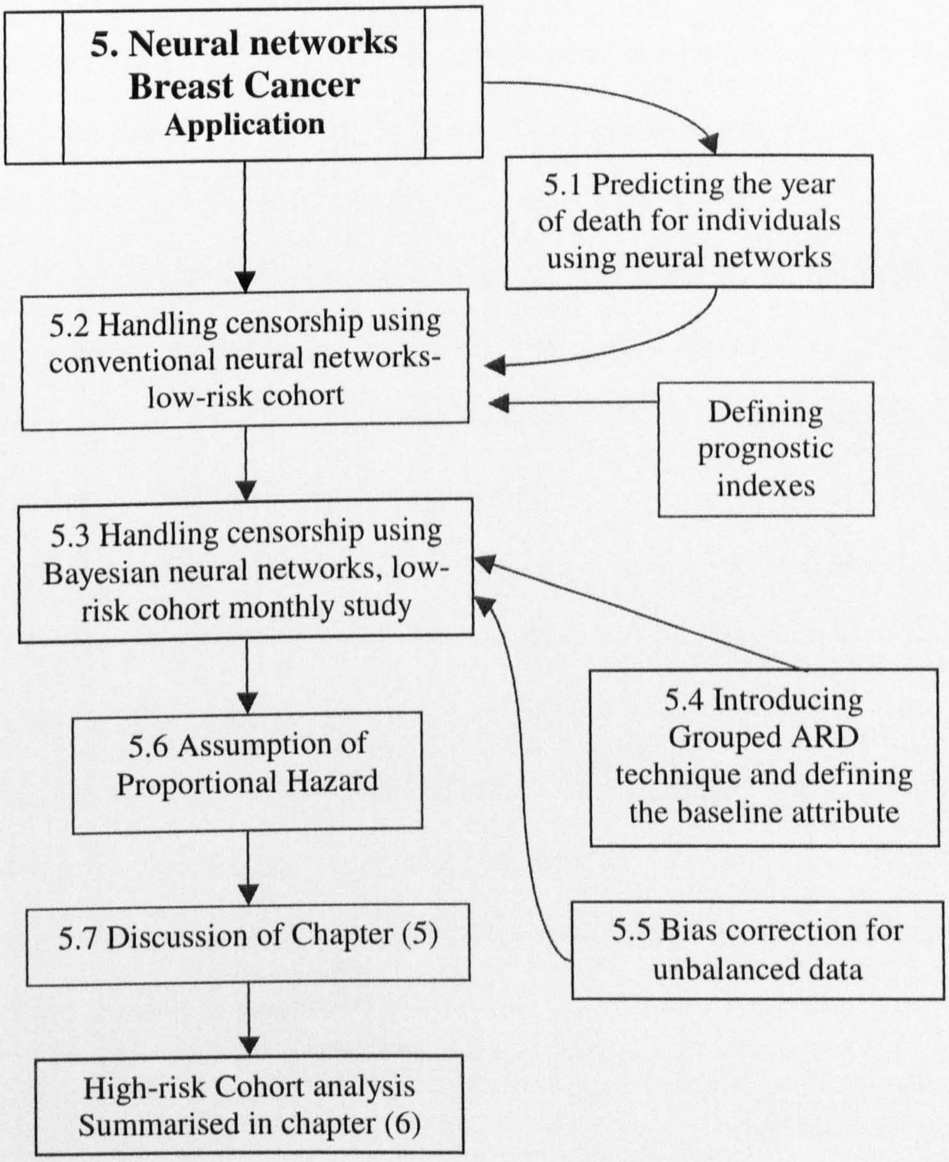
The fact that the *pathological size* variable contained many missing values in the high-risk cohort caused confusion in the cohort assignment, where the records containing missing values of *pathological size* were allocated into the high-risk cohort, since *pathological size* is one of the main separation criteria. The identified prognostic group 1 in the high-risk cohort appears to be a high survival group and contains a substantial number of patients. Most of the patients in this group have the *pathological size* labelled missing. It is possible that they are the patients really showing a high survivorship from the high-risk cohort or it is the confusion caused by the number of clinical data separation criteria containing missing data. Missing data cannot be avoided when collecting a large amount of data and there is no definite solution or method for handling categorical missing data without the potential for introducing bias into the analysis. The data separation criteria may need to be redesigned to include verifying incomplete variables against likely indictor, for example, using *tumour stage* when *pathological size* is unavailable.

Chapter 5

"If you can one day renovate yourself, do so from day to day. Yea, let there be daily renovation."

(Confucius: The Great Learning, Chapter II)

苟日新
日日新
又日新
（大學 第二章）



5. Neural networks Breast Cancer Application
5.1 Predicting the year of death for individuals using neural networks
5.2 Handling censorship using conventional neural networks- low-risk cohort
Defining prognostic indexes
5.3 Handling censorship using Bayesian neural networks, low-risk cohort monthly study
5.4 Introducing Grouped ARD technique and defining the baseline attribute
5.6 Assumption of Proportional Hazard
5.5 Bias correction for unbalanced data
5.7 Discussion of Chapter (5)
High-risk Cohort analysis Summarised in chapter (6)

5. Neural networks breast cancer Application

This chapter summarises the results of two survival analysis approaches using neural networks. Due to the restrictions of the network structure, censorship is not considered at the first approach. In this approach, the probability of death 5 years after surgery is calculated for individuals and presents with the ROC curve, then benchmarks with the Cox regression of the same approach.

In the second approach, the role of analysis has changed. A 5 years survival function is predicted for individuals, in which the neural networks model is modified to be capable of handling censorship, by implementing a partial logistic model. Also, the data are divided into low- and high-risk cohorts. In each cohort, patients are allocated into mortality risk groups, and the corresponding survival function is calculated by the average of the 5 years survivorship prediction of the group. Hence, the accuracy of the prediction is assessed by the Kaplan-Meier estimation of survival from the observations for that group. Moreover, two different neural network approaches are adopted, including the most commonly used MLP trained by back-error propagation and the neural network trained with a Bayesian framework. In the Bayesian neural network approach, the results report substantial bias introduced to the network estimation because of the skewness of the distribution of target values, which is solved by marginalising the outputs to the averaged hazard of the data. This chapter also introduces the more advanced automatic relevant determination (ARD) technique which carries out soft pruning of the model.

5.1 Prediction of 5 years survivorship using the MLP network

Neural networks are non-linear modelling methods, with successful applications in many fields. Medical analysis is one of the fields that adopts this method, in which the probability of a disease occurrence is frequently the variable of interest.

In order to calculate the probability of death 5 years after surgery for breast cancer patients, a 5 year survival curve for individuals is required. Gore *et al* (1984) proposed using the cross point of a threshold that crosses the survival curve to predict the year of death for individual patients. Two different network frameworks are used, the ordinary MLP and the neural network trained with Bayesian framework, the details are given in section (5.1.1) and (5.1.2), respectively.

For each neural network, the 1376 records were split into two groups of 688 records each, selection was dependent on the odd or even record number. One set was used to train the network for parameter generalisation and the other set was used for testing. The number of patients who survived beyond 5 years in the training set and test set is 437 and 436 respectively. The rest of the records are spread over the other 5 years of classes. The outputs can be interpreted as the probability of death at a particular time interval and the cumulative probability of death for *ith* individual is given as

$$\hat{h}_i(t) = \frac{\sum_{l=1}^{t} n_l}{\sum_{l=1}^{k} n_l}$$

for t = 1,2,...,k where k is number of time intervals, n_l is the network output of each time intervals and $\hat{h}_i(t)$ is the cumulative probability of death at particular time interval. Figure (5.1) shows an example of estimated cumulative probability of death of a patient over the first 5 years and beyond 5 years. To predict the year of death for this patient, a threshold is identified for the y-axis and the predicted year of death for an individual patient is identified by the cross-over between the cumulative probability of death and the pre-specified threshold.

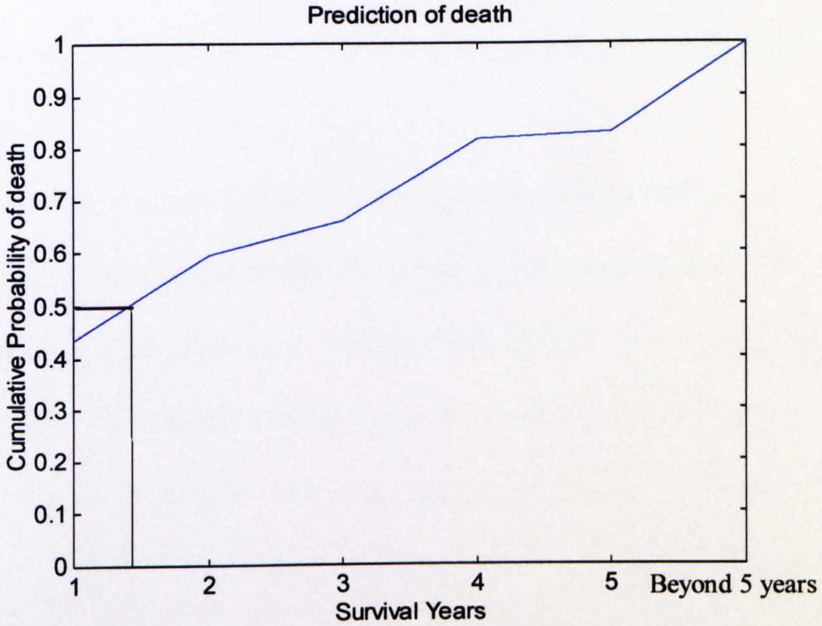


Figure (5.1): An example of neural network estimated cumulative probability of death of an individual patient over the first 5 years and beyond 5 years. Using 0.5 as the threshold to predict the year of death, it is predicted the death is most likely happened in year 2.

5.1.1 MLP network with early -stopping

Six different MLP networks are employed, each network represents one-year interval of 5 years and beyond 5 years. For the patients who survived in that time interval, the target is labeled with 0, or 1 for death of breast cancer. The network consisted of 35 input nodes, one hidden layer of 8 hidden nodes, which has been tested for convergence and performance, and 1 output node. The 35 binary input nodes were transformed from the 8 Cox regression selected variables, reported in section (4.1.1). Since censorship was not considered at this stage, the patients who survived beyond 5 years were considered to be dead after 5 years and those censored before 5 years are discarded.

Early-stopping was employed to overcome the over-fitting problem, where the network training was stopped when the smallest error was archived with respect to new data. During a typical training session, the training data error generally decreases as a function of the number of iterations in the algorithm, whereas the test error first reduces than slowly increases, achieving a minimum value where generalisation is optimal.

Gradient descent was the adopted parameter optimisation algorithm for its simplicity and efficacy and the sigmoid function was the chosen activation function for the network of which restricted the output value to be between 0 and 1, and can be interpreted as probability of death. Each network was trained for 120 iterations.

The calculation required for ROC curves was discussed in section (4.2.1), which involved sensitivity and specificity. The network output can be read as predicting an independent probability of death for each year, same approach as the Cox regression in section (4.2). Figure (5.2) displays the ROC curves for each year and the results are similar to those obtained with Cox regression, and inconclusive.

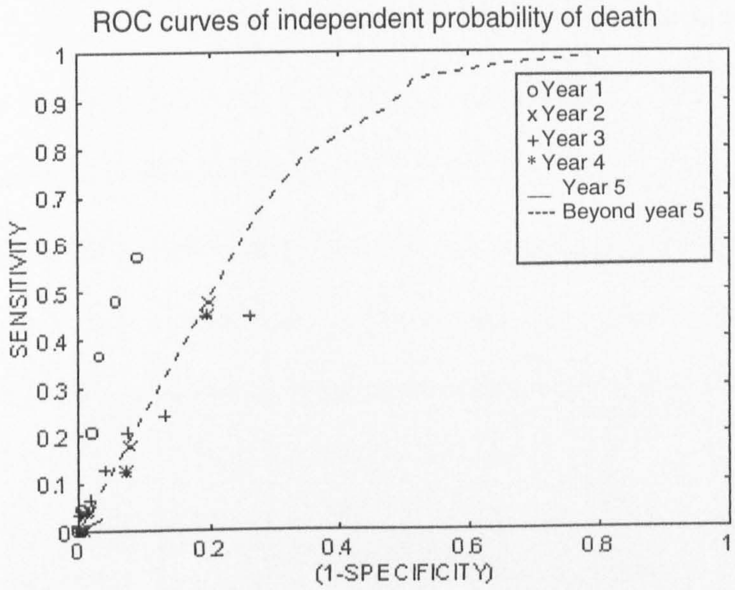


Fig. (5.2): The ROC curves of independent probability of death.

5.1.2 Bayesian Approach

We began with the consideration of the architecture of neural networks, number of layers, number of hidden nodes and choice of activation function. In the conventional maximum likelihood approach, a single 'best' set of weight values is determined by minimising a suitable error function. By contrast, the Bayesian approach considers a probability distribution function over weight space and this can be obtained by calculating the posterior probability distribution given some prior distribution. Once the data has been observed, the prior distribution can be converted to a posterior distribution through the use of Bayes' theorem. The posterior distribution can then be used to evaluate the predictions of the trained network for new inputs, Bishop (1995), chapter 10. Aston's Netlab software was used, this software is specially designed for neural network classification problems.

In the Bayesian model with automatic relevance determination (ARD), there were 38 distinct weights decay parameters, one for the fan-out weights fan from each input node; one for the bias of hidden nodes; one for the output node weights and the last one for the output node bias. The network was trained until all parameters had converged. Figure (5.3) is the ROC curves using the neural network trained with Bayesian framework to predict the year of death of breast cancer for 5 years and beyond 5 years.

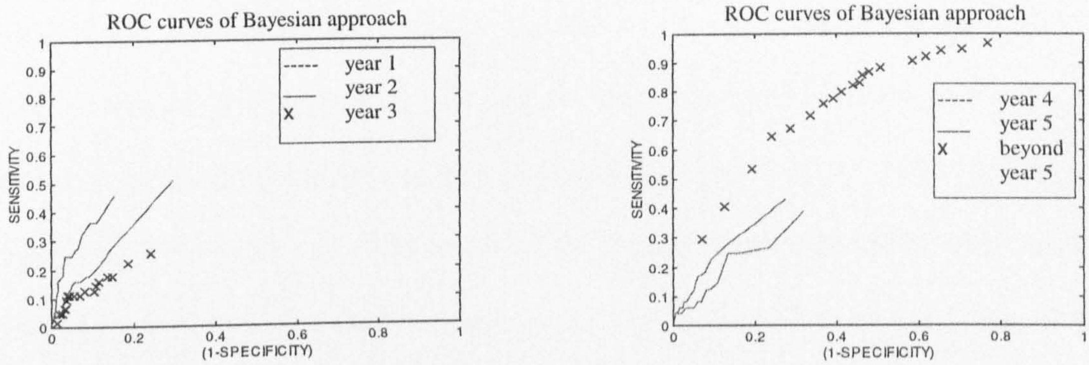


Figure (5.3): The ROC curves of Bayesian regularised neural network for year 1, year 2, year 3, year 4, year 5 and beyond year 5.

5.1.3 Summary of death year prediction with a neural network

By comparing the results of the Bayesian model and the conventional MLP over the independent probability of death method, the Bayesian model shows a better performance on year 2, 3 and 4, while the second shows better performance of prediction in the first year after surgery. There is no apparent difference between the performance of methods for predicting death beyond year 5.

The results also suggests that the neural networks Bayesian approach can perform as well as the Cox regression, figure (4.6). The neural networks perform better in the prediction of death for all time intervals except year 1, which the Cox regression shows better prediction performance, concluding that neural networks are marginally better in long-term outcome prediction. However, none of the results can be considered as being significant for clinical use and some other studies have highlighted that omitting censorship may bias the result, Brown *et al* (1997) and Radvin and Clark (1992), therefore this approach was ended.

Although the results show the neural network performance is marginally better than the Cox regression, both methods failed to produce interpretable results. At this stage, censorship is not considered. Dealing with censorship in the development of neural network model for survival analysis is essential. The *Partial Logistic Artificial Neural Networks* (PLANN) model, Biganzoli (1996), was identified from the literature review to be preferred solutions to handle censorship. The application of PLANN model to the breast cancer data is summarised in the following sections.

5.2 Neural network modelling of censored data

The new approach aims to accurately estimate the cumulative probability of survival for each individual up to a maximum time period, putting all the subjects are partitioned into prognostic groups via the use of prognostic indexes. A predicted mean survivorship for each prognostic group can then be evaluated. Throughout this chapter, all survival analyses were based on 5 years or equivalently 60 months.

5.2.1 Defining a prognostic index in neural network model

Prior to identifying distinct risk groups, it is necessary to rank all of the patients in order of mortality risk. This ranking uses prognostic indexes that are defined separately for both the Cox regression and the PLANN model.

In neural networks, the equivalent of the βx exponent used in Cox rgression is obtained by treating the Multi-Layer Perceptron (MLP) structure as a non-linear extension of logistic regression, and taking the logit of the hazard prediction. However, as this is time-dependent, a cumulative index is obtained by averaging it over the time-span of the study, to give

$$PI_{NN} = \frac{\sum_{i=1}^{T} \log it(y_i)}{T},$$

where T is the number of time intervals (Lisboa *et al* 2000)

5.2.2 The preliminary test of applying the PLANN model to the neural networks using the low-risk cohort

A conventional MLP network was first adopted to implement the PLANN model. The data set used is the low-risk cohort which was split from the data following the criteria given in section (3.3). The training set contained 458 samples with 459 samples in the test set according the odd and even record number. The input variables of the networks were the variables that Cox regression selected namely; *node stage, histology, nodes ratio* and *pathological size* in section (4.4.1). These variables were then transformed into 12 binary attributes, together with the time covariate, formed the input layer of the network. Hence, the value of the time covariate was the mid-point of each time interval. Due to the characteristic of the PLANN model, records in the training set were replicated extensively for each time interval until the patients dropped out from the study. The target label for an observed time interval was 0 where the patient was observed alive and 1 when the event of interest occurs in that time interval. Therefore no sample replication and target labeling was allocated to the subjects after they were dropped out from the study. Unlike the training set, all subjects in the test set were replicated for all time intervals. Only one output node was needed for this model, which represented the conditional probabilities of death from breast cancer in a time interval, therefore the model predicts the hazard mortality.

The networks contained a single hidden layer of 12 hidden nodes and adopted the scaled conjugate gradient (SCG) algorithm as the parameter optimisation method replacing the gradient descent algorithm. This algorithm is claimed to be faster to reach convergence and has fewer pre-set parameters, Bishop (1995). Different values of weight decay parameter have been tested, finally 0.075 was chosen for the wideness range of prognostic indexes and also better separation and grouping of mortality risk groups. In order to overcome the over-fitting problem, early stopping was adopted. The network was trained with only 30 loops.

5.2.2.1 Result Implementation

Since each subject in the test set was replicated 6 times with a different value for the time covariate for a 5 years study, therefore each subject is associated with 6 output values and 6 prognostic indexes, and each output value was independent of the others. They are recorded independently after training, the 6 prognostic indexes are averaged to represent the risk score (prognostic index) of an individual and the 6 output values were transformed to cumulative probability of survival function over 5 years and the calculation for *i*th individual is

$\hat{S}_i(t_j) = \prod_{k=1}^{j} p(t \le t_k \mid t > t_{k-1})$, where j is the number of time intervals.

Hence,

$\hat{S}_i(t_j) = \prod_{k=1}^{j} (1 - y(t_k))$, y_k is the network output at time *k*.

By plotting a histogram of the prognostic indexes of all subjects in the test set, the indexes are naturally gathered into a number of small groups, which can be identified by eye, as shown in figure (5.4). Each band represents one prognostic group, 5 groups were identified in this case. The network predicted mean survivorship for each prognostic group together with the corresponding Kaplan-Meier estimate of the survivor functions including 95% confidence interval are displayed in figure (5.5). The Kaplan-

Meier estimated survivorship function is used to describe the observed survivorship for each prognostic group, which allows assessing the accuracy of network prediction for prognostic groups.

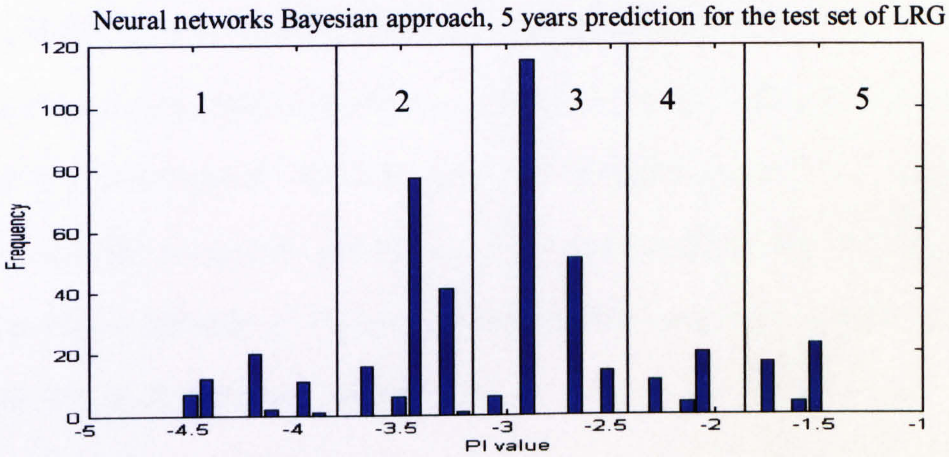
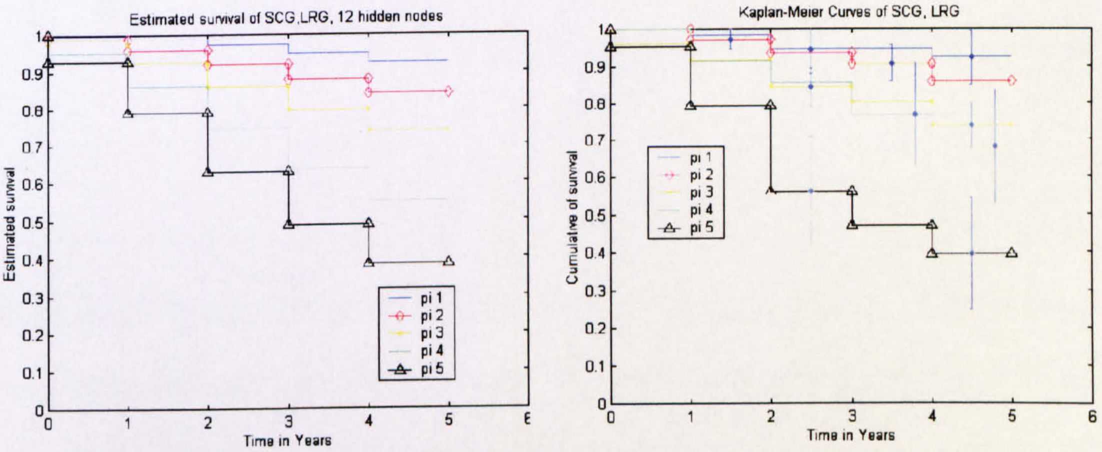


Figure (5.4): Subjects were divided into five mortality risk groups by the eyeballing method from the ranked prognostic indexes



(a)

(b)

Figure (5.5): The neural network SCG approach predicted survivorship over 5 years for prognostic groups, together with the corresponding Kaplan-Meier estimate of the survivor functions. In general, the performance of the neural network survival estimation was acceptable, except group 4, the probability of survival was under estimated by 0.13 at year 5.

The results show that some of the survival curves are close together and their confidence intervals actually overlapped, suggesting that some of the prognostic groups can be potentially combined, i.e. group 1 and 2, group 3 and 4. The combined results are displayed in figure (5.5), finally 3 prognostic groups are left, each contains 194, 221 and 44 subjects, respectively. The confidence interval bands for group 1 and 2 are clear, with no serious over-lapping, and also the accuracy of the estimated survivorship for prognostic groups compared with that observed has been improved. The result after combining specific prognostic groups has given strong statements that the neural network model is capable of handling censored data, and give accurate survival predictions with small confidence intervals.

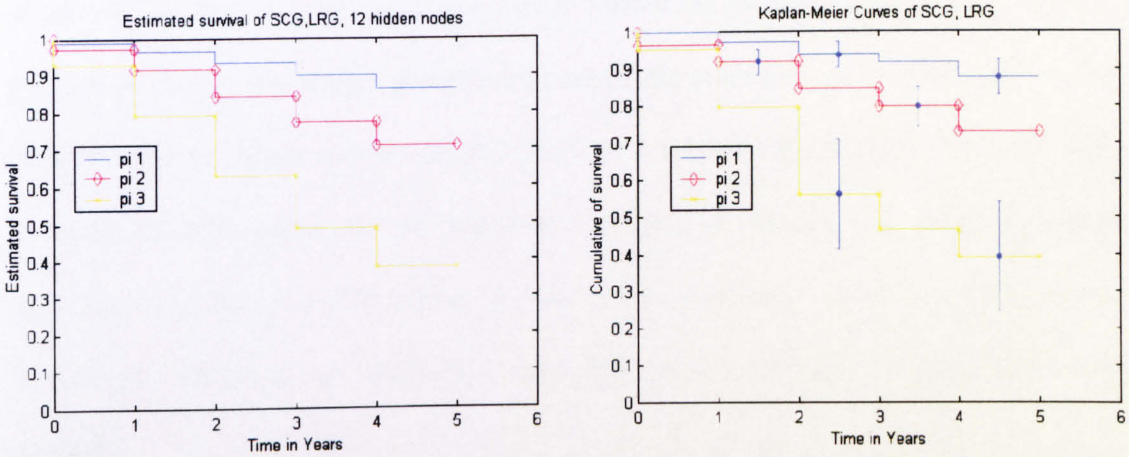


Figure (5.5): The predicted survivorship for the 3 prognostic groups after combining some of the groups from figure (5.4) and the corresponding Kaplan-Meier estimate of the survivor functions. The results display better prognostic group separation and survival estimate accuracy.

5.2.2.2 Discussion of the first neural network model handling censored survival data

The Cox regression, is the most commonly used conventional statistical tool for survival analysis, and its the high popularity is due to commercial availability and its robustness and ease of interpretation. Whereas, the neural network model produced some interesting results which are different from those of the Cox regression. Using the same approach, with fewer prognostic groups, and more accurate survival estimation for prognostic groups. One disadvantage of the neural network approach is the time spent on obtaining the optimal network structure, the correct number of hidden nodes and the weight decay value. Even though a good network design is not always guaranteed the result will be better than the conventional statistical method, it should be at least as good as it. So far, it is only a preliminary test of the potential use of neural networks for censored survival data and the result has given a positive agreement. The next stage is to repeat the SCG approach but implement it with the 5-fold cross validation method, then applying the PLANN model to the neural networks trained with a Bayesian framework, which is an alternative approach to the network weights optimisation method.

5.2.3 SCG approach of low-risk cohort using cross validation procedure

The SCG training and test split approach has suggested that the neural network PLANN model is capable of handling censored survival data. The analysis was repeated once again using the SCG approach but trained with a 5-fold cross validation procedure, which allowed better understanding og the nature of this data in general. The results are summarised in figure (5.6) - (5.7) as in sequence of, dividing mortality risk groups by eye judgement, the network predicted survivorship for each prognostic group together with the corresponding Kaplan-Meier estimate of the survivor functions.

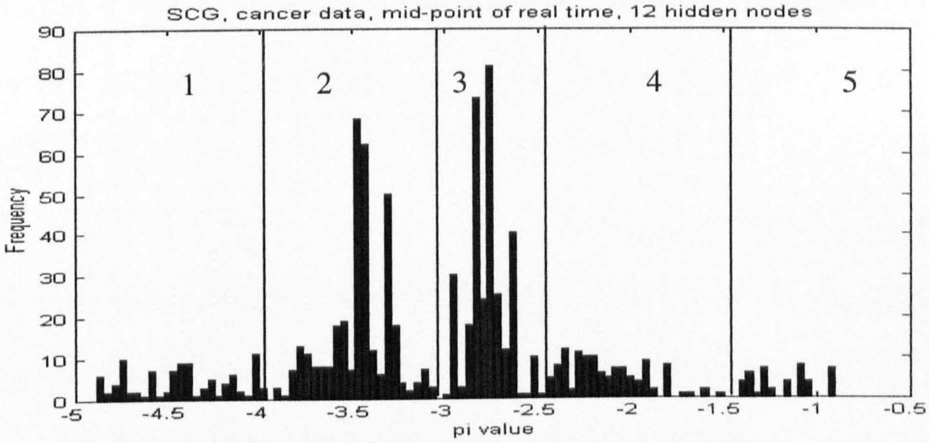


Figure (5.6): The five partitioned mortality risk groups using visualisation grouping method from ranked prognostic indexes which is calculated by the network over 5 fold cross validation sets.

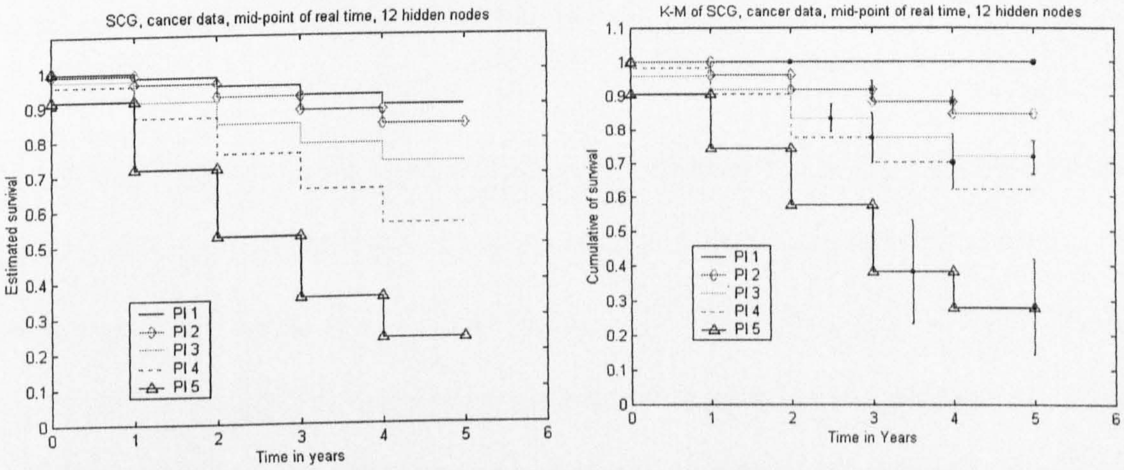


Figure (5.7): The neural network predicted mean survivorship for each of the prognostic groups and the corresponding Kaplan-Meier estimate of the survivor functions. The curves are nicely separated and the survival estimation gives a good agreement.

The result shows that the five prognostic groups are nicely separated. Moreover, the prognostic group 3 and 4 have the potential to be combined and more interestingly, prognostic group 1 shows a 100% of survival chance over 5 years. The accuracy of the estimated survival has not been improved by the cross-validation method but the confidence intervals for each prognostic group are narrower than the training and test split approach.

The SCG approach has already produced some useful results. The next step will test the PLANN model with the neural network Bayesian approach, since it overcomes the over-fitting naturally and the ARD technique can be added on to tune down the irrelevant input variables from affecting the network calculation.

5.3 Bayesian framework for the PLANN model

The evidence approximation to the Bayesian neural network is an alternative parameter regularisation framework. This approach uses a hyperparameter that controls the strength of weight decay. Only a single value of hyper-parameter α is considered at this stage, in which all input variables share same value of alpha. Multiple alpha values will be considered later in the ARD approach, reported in section (5.4). The input variables were those selected by the Cox regression which allows comparison over different weight optimisation approach. A single layer of eighteen hidden nodes was adopted for better network estimates, wider range of prognostic indexes and better mortality group separation. One output node was used. The analysis was completed with a 5-fold cross validation and yearly bases of 5 years. The result of prognostic group partitioning was that the network predicted survivorship for prognostic groups and the corresponding

Kaplan-Meier estimate of the survivor functions are summarised in the figure (5.8) - (5.9), respectively.

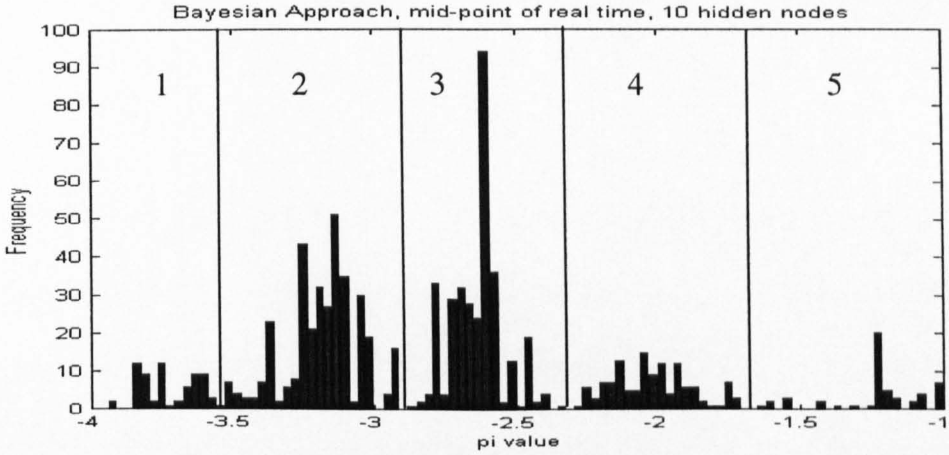


Figure (5.8): The neural network Bayesian approach evaluated prognostic indexes of 917 low-risk cohort data, 5 prognostic groups were partitioned by visualisation grouping method.

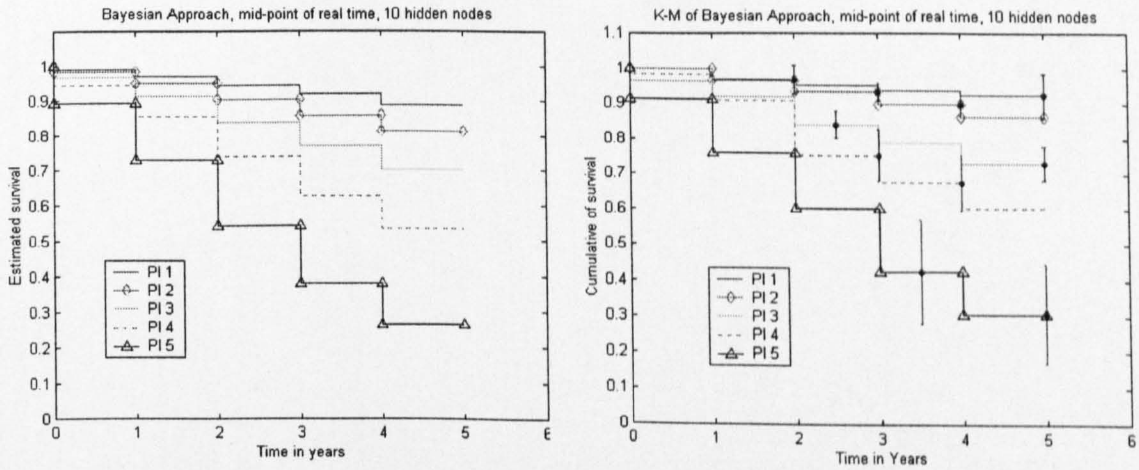


Figure (5.9): The network predicted survivorship for each of the prognostic groups and the corresponding Kaplan-Meier estimate of the survivor functions.

Figure (5.9) shows that the survival curves of 5 prognostic groups are nicely separated between 0.97 to 0.3 at year 5. The majority of patients are partitioned into groups 2 and 3. The SCG approach aggregated highest survival group which showed a 100% survivorship has disappeared, the predicted survival of the newly formed prognostic survivorship has disappeared, the predicted survival of the newly formed prognostic group 1 is 0.97, still a very high survival group. In term of the accuracy of survival prediction, the Bayesian approach is marginally better than the SCG approach in general.

The result is comparable with the conventional statistical tools, the Cox regression in this study. Another special feature of the Bayesian framework is the use of the ARD technique where the input variables that are least relevant to class differentiation can be determined. Although the input variables were selected by Cox regression and have been proved to be effective in prediction, neural networks may act differently on these variables since they are non-linear methods unlike linear Cox regression.

5.4 The PLANN model of Bayesian framework using ARD

The ARD technique was based on the use of a separate hyper-parameter for each input variable. Each alpha controls the optimisation of the network weights that fan out from the each input variable. The assumption is that irrelevant, or noisy covariates, develop large hyperparameter values that penalise the objective function, E, driving down the values of the regression coefficients (or weights) associated with them. Therefore, the bigger value of the alpha, the smaller value of the corresponding weights to be. In other word, the alpha value is a measurement that determines the irrelevant input variables and minimises their influence towards the network output. This is called *soft pruning*.

The network input variables were those by selected Cox regression. The role of hyperparameters in here is to examine how these variables have been handled in the network.

5.4.1 Group ARD Concept

Owning to the structure of categorical data, each of the input variables was transformed into several binary input attributes in the network. Originally, the ARD technique assigns a single value of the hyperparameter alpha to each input variable. In this new approach, instead of assigning a value of alpha to the group of weights which fan out from each input node, a single value of alpha is associated with the weights that fan out from all of the input attributes which correspond to a single variable. This is called the *grouped ARD technique*.

This approach was applied to the low-risk cohort and implemented with a 5-fold cross validation again and same input variables were used. Different numbers of hidden nodes were tested, the best network output estimation was given by 18 hidden nodes. The results are summarised in figure (5.10) - (5.11).

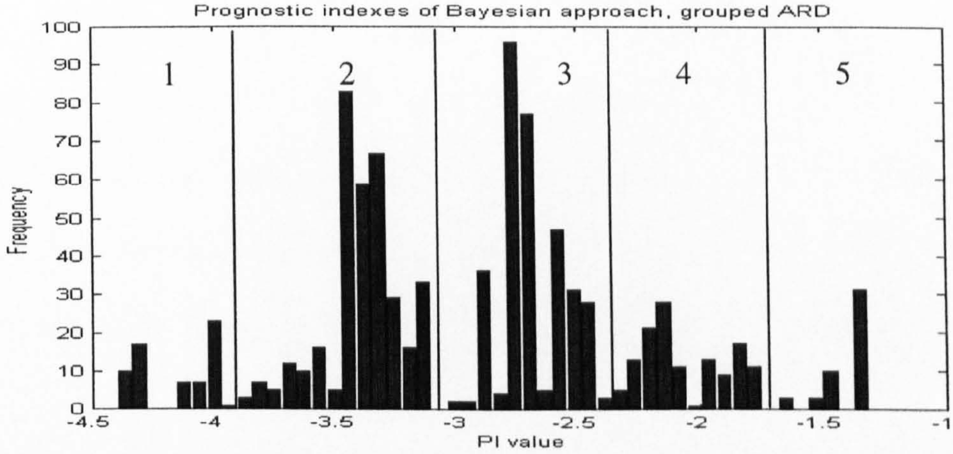


Figure (5.10): Prognostic indexes that evaluated by the grouped ARD Bayesian approach for the low-risk cohort and five prognostic groups were partitioned by visualisation grouping method.

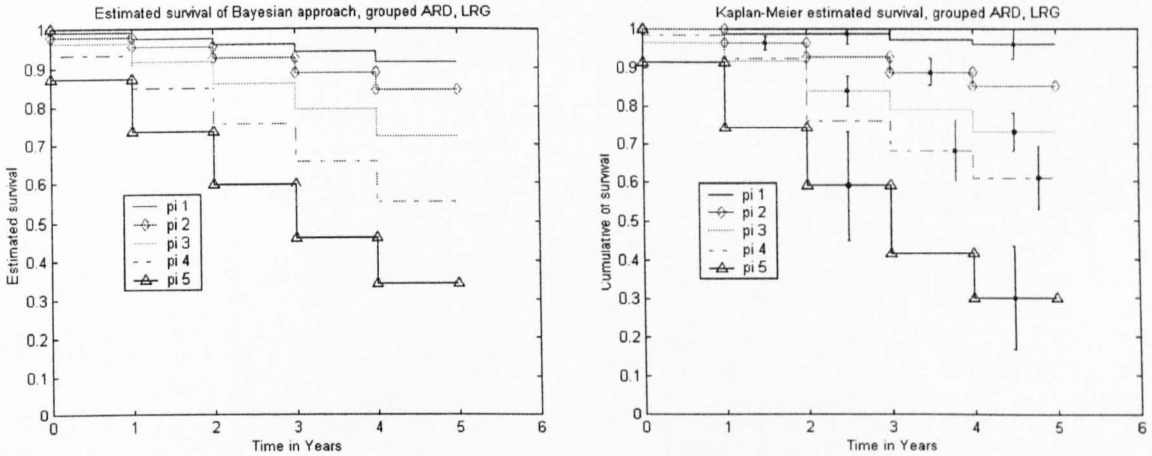


Figure (5.11): The grouped ARD network predicted mean survivorship for prognostic groups and the corresponding Kaplan-Meier estimate of the survivor functions. The separation of prognostic groups over 5 prognostic groups is better than the single alpha ARD approach and also better survival estimation for each of the prognostic groups.

The results show that the grouped ARD technique is successful and the accuracy of the estimated survival for prognostic groups have also been improved. However, the value of alpha hyper-parameters for each input variable are strangely large, further investigation will be reported in section (5.4.2). Hence, the neural network analysis will be based on the use of grouped ARD technique and also the prognostic groups partitioning method. We will be using the log-rank test to choose the optimal position of the thresholds to aggregate prognostic groups from the ranked prognostic index values.

5.4.2 Baseline attributes determination

The value of alpha hyper-parameter in the single alpha approach was around 6. However, the alpha values of the grouped ARD approach varied between 22.57 to 84.71. The two sets of alpha were very different. The value of alpha corresponds to an inverse variance, as the bigger the posterior variance of weights, the smaller value of alpha would be, which leads to higher significant influence of the corresponding input variable to the output estimation.

The large value of grouped ARD suggests some redundancy between the group of attributes corresponding to a single input variable. After all, the attributes for each variable must sum to one, imposing a constraint on their values.

5.4.2.1 Applying the conventional ARD technique

The network setting has not been changed, still trained with 18 hidden nodes, each of input variable was received a separate hyper-parameter alpha. Table (5.1) displays the values of alpha of each attribute.

Histology	**1**	**2**	**3**		
Alpha value	11.94	1513.8	70.216		
Pathological size	**1**	**2**			
Alpha value	14.789	461.5			
Node stage	**0**	**1**			
Alpha value	14.726	369.03			
Nodes ratio	**<=20%**	**20-30%**	**30-60%**	**60%+**	**Unknown**
Alpha value	11.149	2082.2	2030	12.759	8447.2

Table (5.1): The reading of alphas that are corresponding to each of the input nodes. One of the alpha values within a variable is distinguishably large.

It is clear that for each variable one attribute may be regarded as irrelevant. In order to maintain the consistency, the attribute becomes the baseline is same as the Cox regression, the lowest hazard attribute of the variable, the value of the baseline for that variable is coded by all remaining attributes equal to zero.

5.4.2.2 Adding the baseline attribute assumption to the training criteria

By considering the baseline attribute approach, the network was retrained with the grouped ARD technique and included the baseline attributes assumption. The chosen baseline attribute for each variable was same as the baseline category for Cox regression. The results show that the alpha value of each variable has been significantly reduced. The new reading of alphas are 1.4353, 1.1534, 1.6545, 3.0963 and 2.6333 for variables *histology, pathological size, node stage, nodes ratio* and *time* respectively. The prognostic indexes are still within the range of –4.5 to –1 as displayed in figure (5.12) and the distribution of samples are similar to our previous results, the 917 samples have been successfully partitioned into 4 prognostic groups by the log-rank test, as illustrated in figure (5.13). Their predicted mean survivorship at year 5 is varied between 0.97 to 0.22 and each respectively contains 75, 341, 460 and 41 number of subjects. The network survival predictions in general can be concluded as acceptable. Although the network estimation for group 1 and group 4 show a small deviation from the corresponding Kaplan-Meier estimate of the survivor functions, the curves are still being included within the confidence interval bands, considering that their confidence interval bands are bigger than the other two groups.

The results show there is a need to define baseline population when handling categorical data. The baseline attributes are given values of zeros and the same attribute in each variable is allocated to both of the neural networks and the Cox regression.

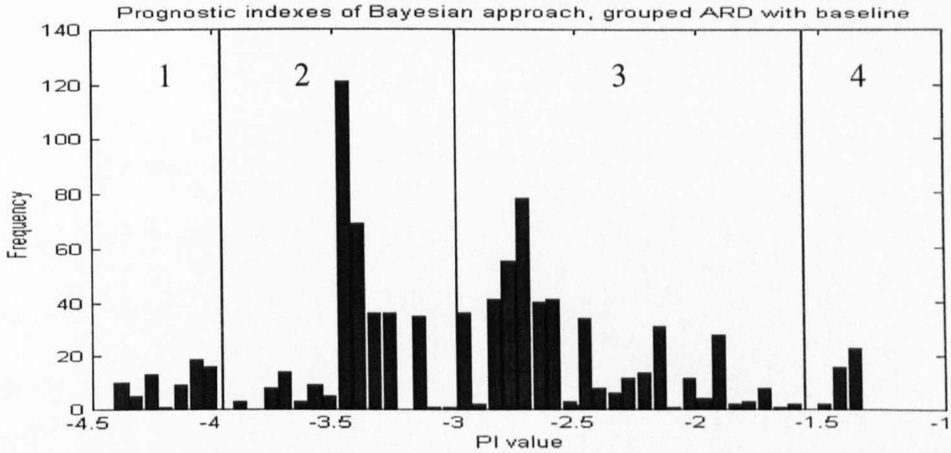


Figure (5.12): The neural networks computed prognostic indexes using the grouped ARD technique and baseline attributes assumption. Four prognostic groups were divided by the log-rank test.

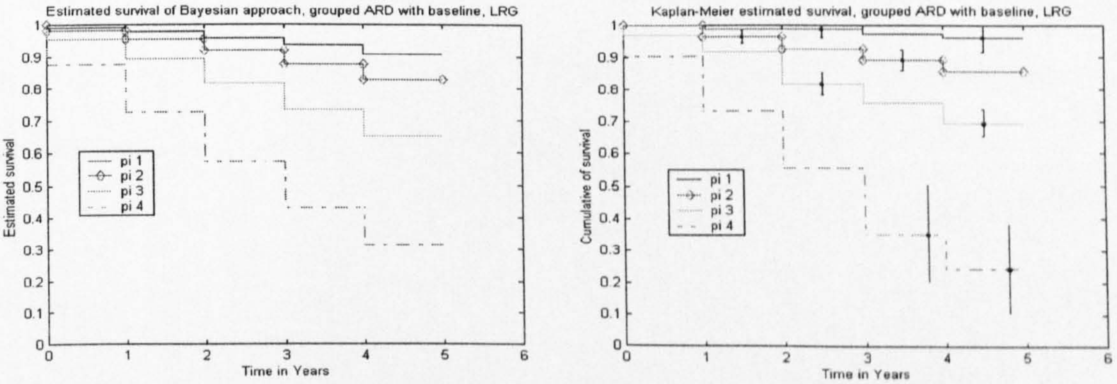


Figure (5.13): The network predicted mean survivorship for prognostic groups after applying the grouped ARD technique and the baseline attributes assumption and the corresponding Kaplan-Meier estimate of the survivor functions. The curves are well separated and the network prediction is acceptable.

5.4.2.3 Displaying the attribute profiles of each prognostic group

Since the techniques of data analysis using neural networks has been refined, it is necessarily to determine the characteristic of each prognostic group, in which the variables that play a leading role of in each prognostic group are examined. Therefore, the variable attribute histogram for each prognostic group is displayed in figure (5.14).

Clearly, particular attributes of variables are dominated in particular prognostic group. Group 1 is the highest survival group of the data, the attribute of *histology* moves from attributes 2 and 3 to mainly attribute 1 from group 1 to 2. Over the 4 prognostic groups, the *pathological size* and *node stage* move gradually from attribute 1 to 2 and attribute 0 to 1, respectively. Finally, all the variables are concentrated on their particular attribute that creates prognostic group 4, which is the lowest survival group.

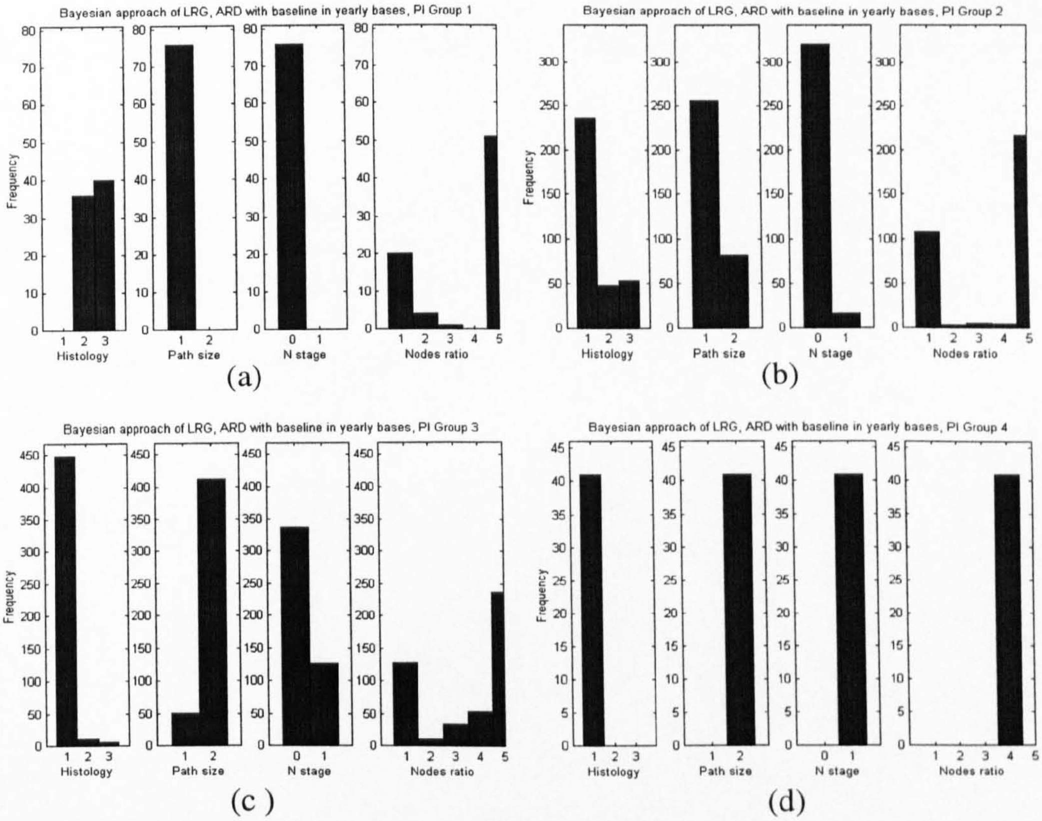


Figure (5.14): The attribute profiles of prognostic groups which generated by the neural networks with the application of grouped ARD technique and the baseline attribute assumption in which the behaviour of each variable over prognostic groups can be monitored.

5.5 Bias correction of network output due to heavily skewed binary data

The Bayesian framework does not take account of the skewed distribution of target labels in binary classification problems, with the consequence that all network outputs are marginalised to the mid-range of the value. In the evidence approximation to the integral of

$P(c_1 \mid x, D) = \int g(a)P(a \mid x, D)da$, then

$$P(c_1 \mid x, D) \cong g\left((1+\frac{\pi s^2}{8})^{-1/2} a_{MP} \right),$$

where s^2 is the variance of the sample distribution, $a \sim N(a_{MP}, s^2)$. The $P(c_1 \mid x, D)$ is the probability of class membership c_1 given the data x and training data set D.

For a two classes problem, the network output is adjusted to minimise the probability of misclassification of the given input data with the decision boundary, corresponding to a network output of $p(C_1 \mid x, D) = 0.5$. The form of the logistic sigmoid activation function determines that $a_{MP}(x, w_{MP}) = 0$. The $p(C_1 \mid x, D) = 0.5$ statement is no longer held when the data are heavily skewed. Some modifications to the network error and estimates have been proposed, Lisboa *et al* (2000). Firstly, by weighting the cost function using Bayes' theorem so the network is as if trained with an equal prior. In applying Bayesian neural networks to the modelling of censored data, it is necessary to re-weight the error function to equalise the heavily skewed distribution of mortality indicator indexes follows

$$LL = -\sum_n \left(\log(y_n).\left(\frac{t_n}{2d}\right) + \log(1-y_n).\left(\frac{1-t_n}{2(1-d)}\right)\right),$$

where d represents the frequency of death, t_n is the target labels and y_n is the network outputs. This modification also applies to the gradient and the Hessian calculations in the similar manner.

The conditional network estimates are then compensated using Bayes' theorem, to take account of the true prior distribution for the target labels, resulting in conditional network estimates that marginalise to the priors, d, which is the averaged hazard in this study, i.e.

$$y_g(x) = \frac{\tilde{y}_g(x)d}{\tilde{y}_g(x)d + (1-\tilde{y}_g(x))(1-d)},$$ where $\tilde{y}_g(x)$ is the network output.

When $\tilde{y}_g(x) = 0.5$, it follows that

$$y_g(x) = \frac{0.5d}{0.5d + (1-0.5)(1-d)} = d.$$

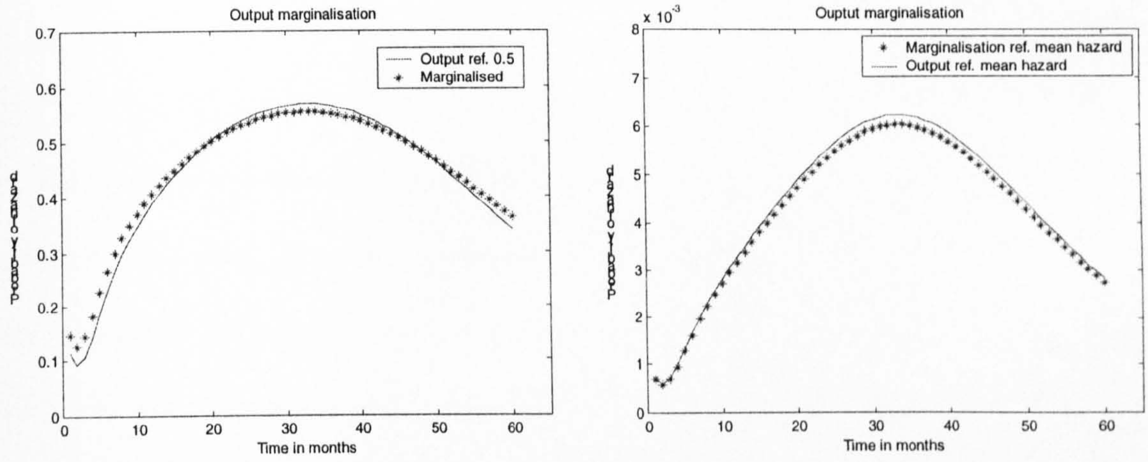
So, $\tilde{y}_g(x)$ marginalises to 0.5 while $y_g(x)$ marginalises to d.

The calculation of the averaged hazard involved two parts, (i): the probability of death at each time interval is the total number of death within the time interval divided by the total number of patients at risk at the beginning of the time interval, (ii): then averaging the probabilities of death by the number of time intervals.

5.5.1 Demonstration of the effect of network output marginalising towards class prior

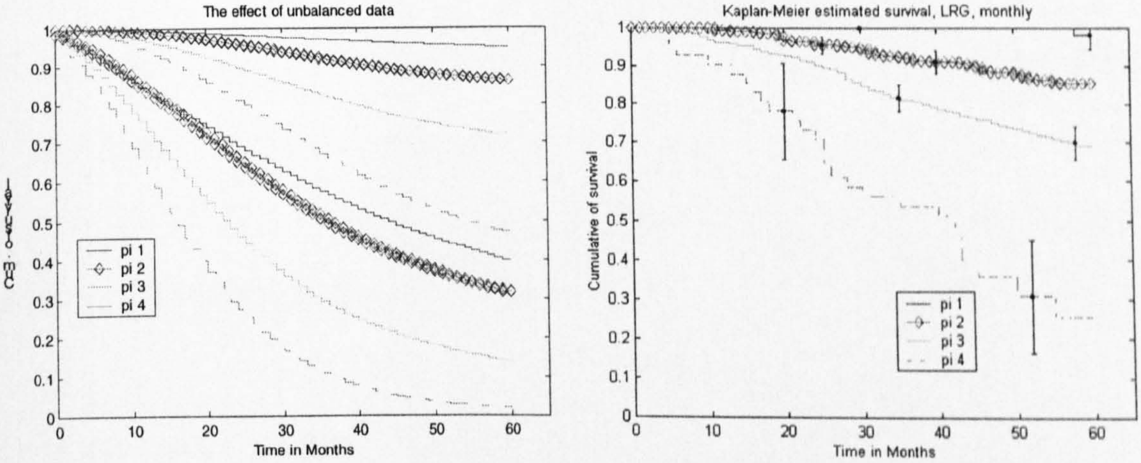
The breast cancer analysis was repeated using the PLANN model, but refined into monthly bases. The network estimates were marginalised towards the averaged hazards of the data, which was calculated by averaging the hazard of each time interval. The network was still using 18 hidden nodes, the grouped ARD technique and baseline attributes assumption were also applied, mortality risk groups were partitioned by the log-rank test, finally, the analysis was implemented using the 5-fold cross validation.

The calculation of cumulative survivorship involves a series of network output multiplication, described in section 5.2.2.1. Therefore, any bias in the calculation of each hazard rate causes a huge bias after multiplication over several time intervals. Figure (5.15) – (5.16) demonstrate the effect of marginalisation towards midpoints and the class priors. Figure (5.15) displays four different network outputs, the original output, the network output marginalised towards midpoint; the network output marginalised towards averaged hazard from the original output; and marginalised towards averaged hazard from the network output had marginalised towards midpoint. Nevertheless, the top 4 curves in figure (5.16a) are the network survival prediction for 4 prognostic groups which marginalised towards averaged hazard and the lower 4 curves, the corresponding network prediction marginalised towards midpoint. Figure (5.16b) is the corresponding Kaplan-Meier estimate of the survivor functions. The network outputs are seriously damaged when the bias correction is not applied.



(a) (b)

Figure (5.15) (a): The original output predicted by the neural network (-) and the marginalised output (*), averaged from all patients in low-risk cohort. (b): the original network output marginalised towards averaged hazard directly (-) and the midpoint marginalised result marginalised towards averaged hazard (*).



(a) (b)

Figure (5.16), (a): The top 4 curves are the network survival prediction which marginalised towards averaged hazard and the lower 4 curves are the survival prediction which marginalised towards midrange and (b) is the corresponding Kaplan-Meier estimate of the survivor functions.

5.5.2 Complete the low-risk Cohort analysis into monthly study

After defining the necessary techniques to correct the marginalisation to take account of the skewness of the target distribution, still using 18 hidden nodes, the network was retrained, same input variables as before and with grouped ARD. The time intervals have been refined into a monthly study over 60 months. Thus, data were replicated more frequently than the yearly study and the mean hazard per time interval was correspondingly smaller. For the individuals who survived more than 60 months were censored at month 61.

Figure (5.17) demonstrates when the log-odds ratio, $\tilde{a}(x, w_{MP})$, is zero, the marginalisation of $\tilde{y}_g(x)$ is toward 0.5 as if the case of network trained with equal priors. After compensation for the time value of the prior, $y_g(x)$ marginalises to the averaged hazard in this case 0.0032. The new range of prognostic indexes lies between –3 and 5. Four prognostic groups are partitioned by the log-rank test and the patients allocation is 56, 359, 460 and 42 respectively, with a majority of patients still allocated to group 3, as illustrated in figure (5.18).

Only the network predicted mean survivorship for the lowest survival group is not accurate showing 0.16 error when comparing figure (5.19a) and (5.19b). However, all survival curves are included within the Kaplan-Meier estimated confidence interval bands. Since the network outputs were marginalised toward the mean hazards, thence, the mean survival rate at year 5 is around 0.7, therefore the error generated by prognostic group 4 can then be explained. One solution to solve this situation is to model each prognostic group separately, which would allow accurate prediction of survival for the patients in each group, given the prognostic group allocation already decided on the basis of the most likely values of a, namely a_{MP}.

There is a significant difference observed from the attribute profiles over prognostic groups in figure (5.20), when comparing with the Cox regression approach in figure (4.14). The profile of variables in each prognostic group is more highly concentrated on a particular attribute, thus reducing the overlap between prognostic groups.

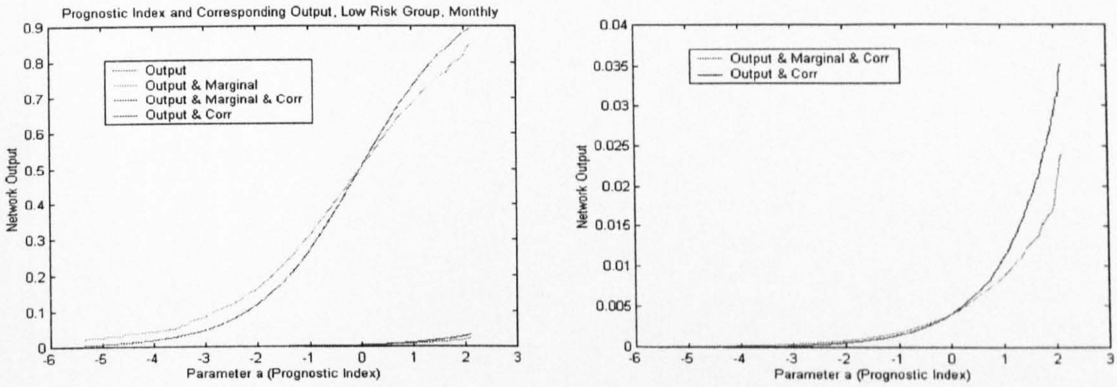


Figure (5.17): The network outputs and their associated prognostic index in four format, namely original network output (Output), network output marginalised towards midpoint (Output & Marginal), network output marginalised towards average hazard using the outputs marginalised towards midpoint (Output & Marginal & Corr) and network output marginalised towards average hazard directly (Output & Corr).

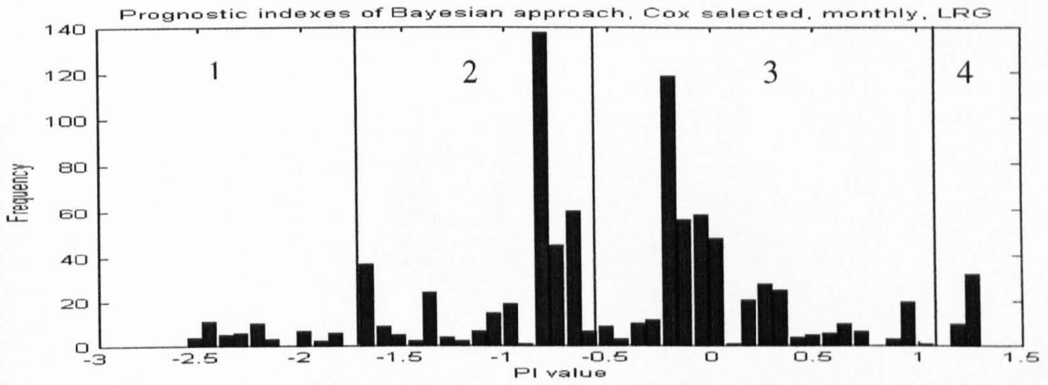


Figure (5.18): The log-rank test partitioned 4 prognostic groups or the low-risk cohort, the network outputs are marginalised towards the averaged hazard of the data.

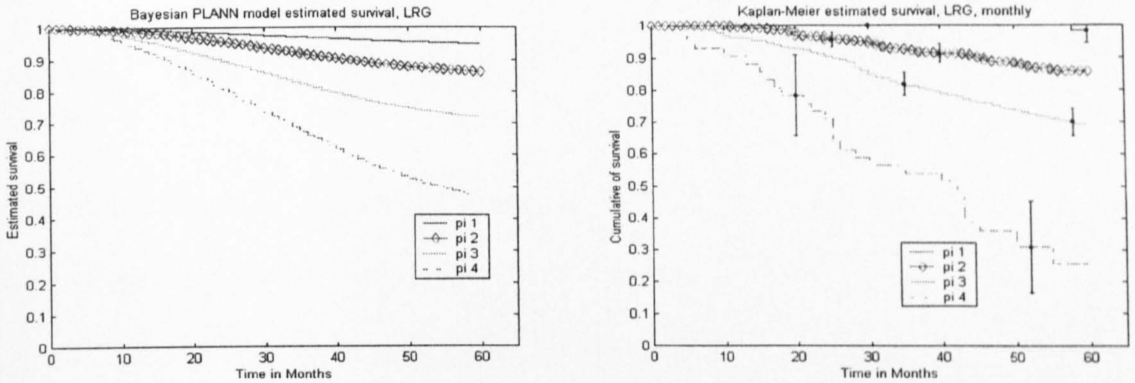


Figure (5.19): The network estimates mean survivorship for prognostic together with the corresponding Kaplan-Meier estimate of the survivor functions. The network is trained with bias correction technique and the outputs are marginalised towards averaged hazard. The survival prediction for prognostic groups is concluded as accurate, except group 4.

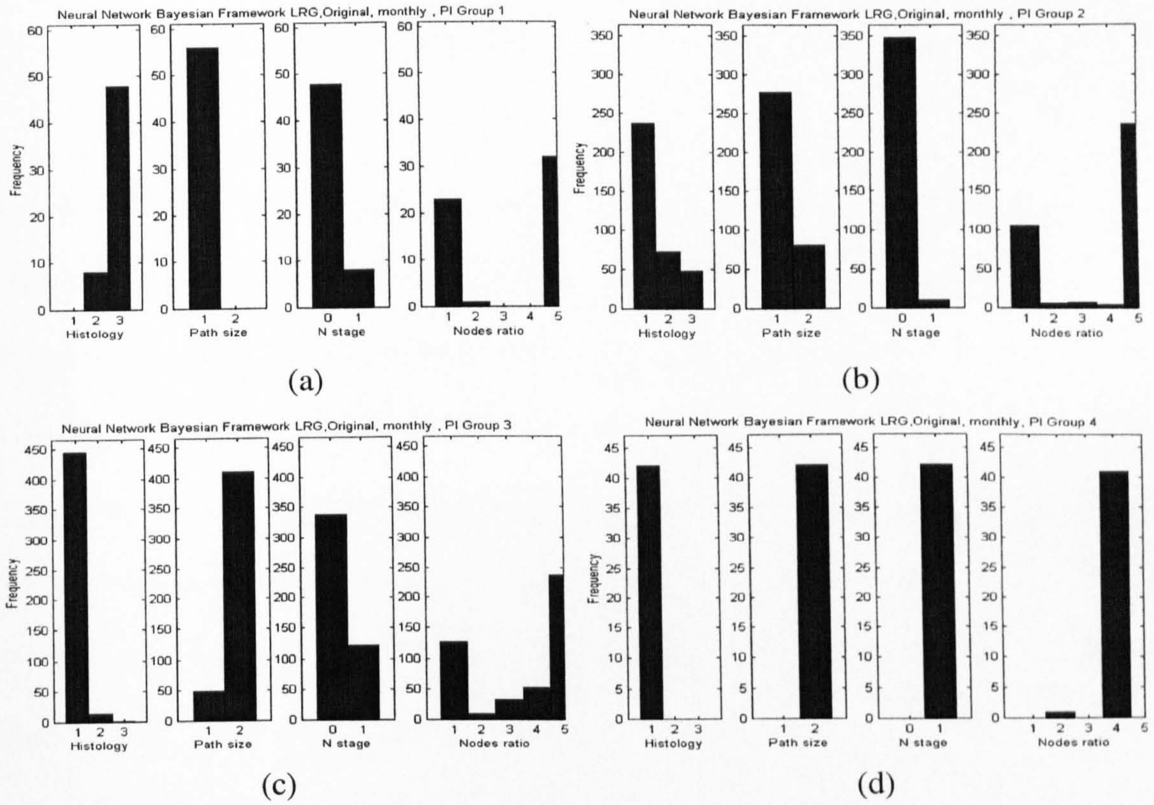


Figure (5.20): The attribute profiles for prognostic groups, which the network outputs are marginalised towards averaged hazard. The distribution of patients has been changed significantly from the previous neural network result, they are more concentrated on particular attribute in each variable.

When the analysis has been refined to a monthly study, more detailed information can be extracted from the data, including a smooth prediction of the hazard, shown in figure (5.21). The results so far indicate that a small group of 42 patients in this cohort has a relatively low survival. This group of patients will be examined further and the results are summarised in section (6.4).

5.6 Assumption of proportional hazard

The proportional hazards model allows a non-constant hazard rate to be modelled without making any assumption about the underlying distribution of the hazards in the different groups, but is requires the hazards in the groups remain proportional over time. Therefore, the time dependence of the hazard is that observed for the baseline population. This assumption was assessed by the commonly used Cox-Snell residual plot or some other residual plots, all methods have confirmed no significant evidence that the data were not fitted into the proportional hazard assumption, as reported in section (4.1.2). However, such residual plots are not precise in verifying hazard proportionality between prognostic groups.

Handling censorship is a main feature of the PLANN model; nevertheless it is also capable of generating a smooth hazard rate over time. Figure (5.21) displays the mean hazard for each of the 4 prognostic groups over time and shows that the hazard of each prognostic group was not uniformly proportional to each other. The peak hazard for each group is retarded slightly as the hazard increases, indicating only a minor deviation from the proportionality assumption over the time frame of the study. Gore *et al* (1984) confirmed that if time to peak hazard is earlier in some prognostic groups than in others, then the proportional hazard assumption is no longer sustained and has been the case in breast cancer.

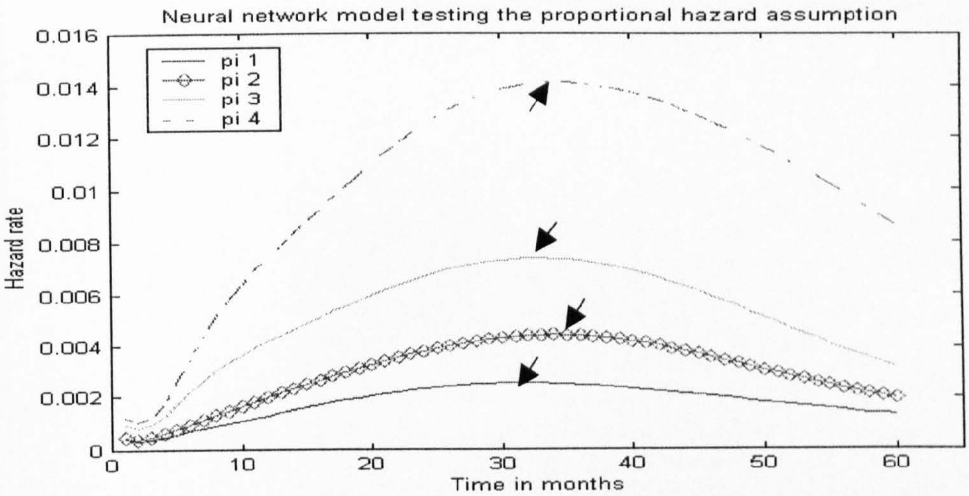


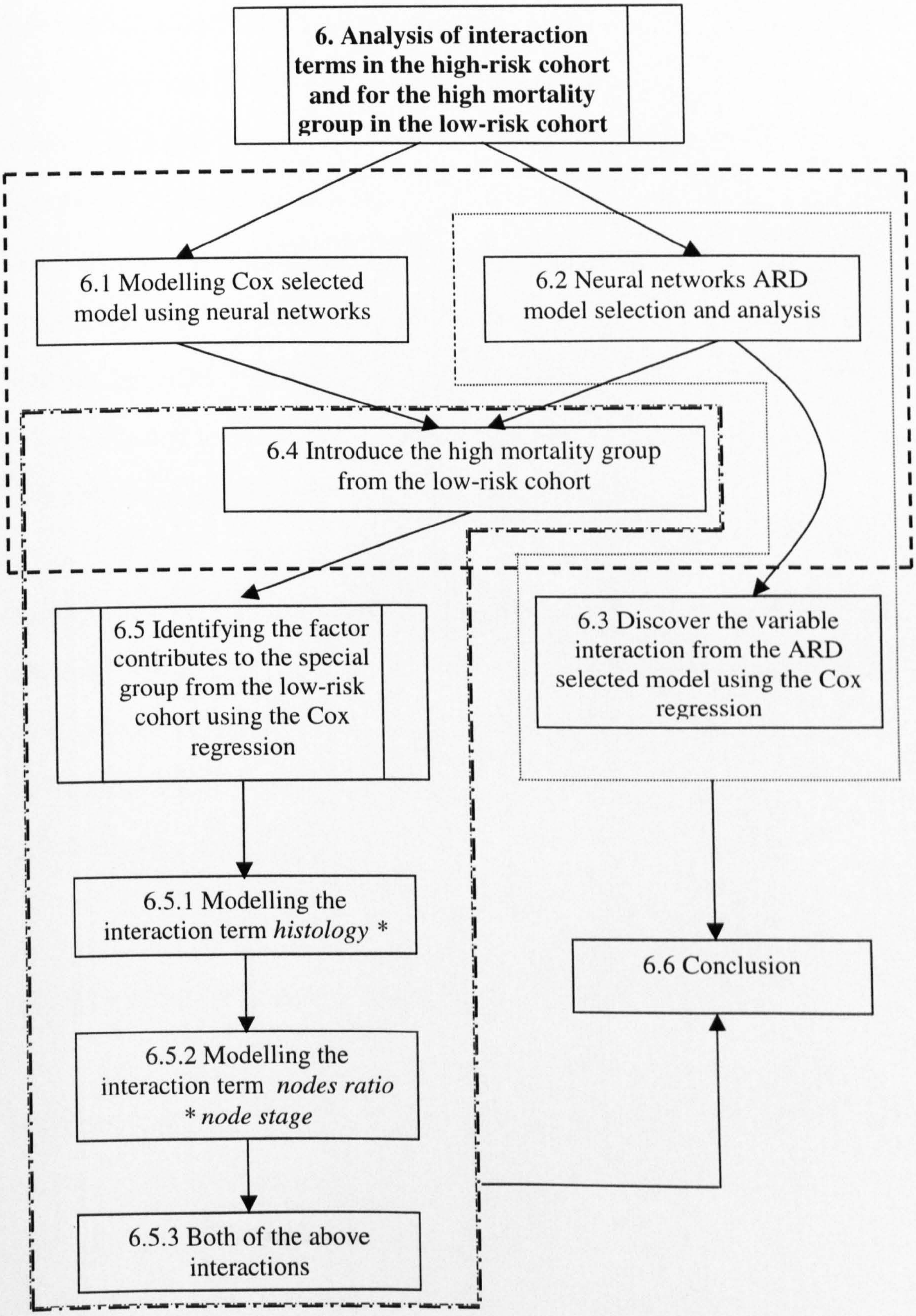
Figure (5.21): The network predicted hazard probability of the 4 prognostic groups over 60 months for the low-risk cohort. The arrow point at each curve is where the corresponding peak hazard occurred.

5.7 Discussion of chapter (5)

In this chapter, it was demonstrated the PLANN model is capable of handling censored survival data and can be adopted easily by a standard neural network model for classification problem. The Bayesian neural network has performed as well as the Cox regression, although they responded differently and produced slightly different allocations into the prognostic groups, the neural network being more specific in attribute profiles in each risk group. Moreover, the robustness of the Cox regression has also been demonstrated when the non-proportionality of hazard has been confirmed within the data.

The same data analysis will be repeated for the high-risk cohort which is summarised in chapter (6). Variable interactions are also investigated using the neural network model regularised with ARD for high-risk cohort.

Chapter 6


6. Analysis of interaction terms in the high-risk cohort and for the high mortality group in the low-risk cohort
6.1 Modelling Cox selected model using neural networks
6.2 Neural networks ARD model selection and analysis
6.4 Introduce the high mortality group from the low-risk cohort
6.5 Identifying the factor contributes to the special group from the low-risk cohort using the Cox regression
6.3 Discover the variable interaction from the ARD selected model using the Cox regression
6.5.1 Modelling the interaction term *histology* *
6.6 Conclusion
6.5.2 Modelling the interaction term *nodes ratio* * *node stage*
6.5.3 Both of the above interactions

6 High-risk cohort

The high-risk cohort contains all the remaining subjects who did not fit into clinical separation criteria for the low-risk cohort. This includes any occurrence of large tumours, fixed affected nodes in the axilla, and distant metastases. However, it also includes the subjects with *pathological size* coded unknown, making a total of 633 subjects. The following sections contain the neural network analysis using different models, Cox selected variables and ARD selected variables. All of the analyses for the high-risk cohort are implemented with a 3-fold cross validation, in order to reduce the computational time, rather than a 5-fold cross validation as for the low-risk cohort, which still leaves significant amount of samples for network training.

An analysis of the high-risk patient group identified from the low-risk cohort by the neural network in previous chapters, was also included, with the aim of determining the characteristics of survivorship in this group. Additionally, variable interactions are investigated in this chapter, by identifying variables with ARD then including explicit interactions into Cox regression.

6.1 Neural networks analysis using the Cox selected variables

Forward-stepwise elimination was again employed to select the optimal Cox model for the high-risk cohort from the original 18 variables, the details are summarised in table (4.4). The selected variables and the time variable formed the network input layer, altogether 16 input nodes when the baseline attribute is removed. The baseline population was chosen as in the Cox regression, to be the lowest hazard categories. A single hidden layer of 18 hidden nodes was used for consistency with previous results. The single output node represented the hazard rate of an individual at a particular time interval. Grouped ARD was also used.

The analysis consisted of a monthly study over 5 years, marginalising the hazards towards the average hazard of the training data. The log rank test was, again, employed to define prognostic groups. After completing the network training process, the final values of the hyper-parameter alpha were ranked and used to identify the main contributing variables, which are *menopausal status, node stage, pathological size, clinical stage* and *nodes ratio.*

A total of 3 prognostic groups were identified, containing 248, 174 and 211 patients, respectively. The thresholds determined by the log-rank test are indicated in the plot of the prognostic indexes, figure (6.1). The network predicted mean survivorship for prognostic groups and the corresponding Kaplan-Meier estimate of the survivor functions are shown in figure (6.2). Good survival prediction for all groups is obtained, as with the low-risk cohort. There is a minor inaccuracy in the estimates of survival for prognostic group 2, which is the consequence of the network output marginalisation towards the mean hazard for all of the groups. The overall mean hazard has been suppressed towards 0 due to the patients in the lower risk group, prognostic group 1. The highest survival group is highly populated with *node stage* 0, representing negative node, and *pathological size* coded as unknown. The remaining variables used for the clinical data separation criteria indicate they may belong to the low-risk cohort.

Finally figure (6.3) shows the attribute histograms for each of the prognostic groups. In each prognostic group, one or two variables are most prominent, thence each of the variables contributes differently to each group. However, the *menopausal status* and *pathological size* do not show clearly differentiated attribute profiles over the three prognostic groups. Moreover, apparently, the *menopausal status* was given the largest value of alpha, and its profile is similar across all of the prognostic groups. This indicates that *menopausal status* contributes to the survivorship estimation through interactions with the other variables.

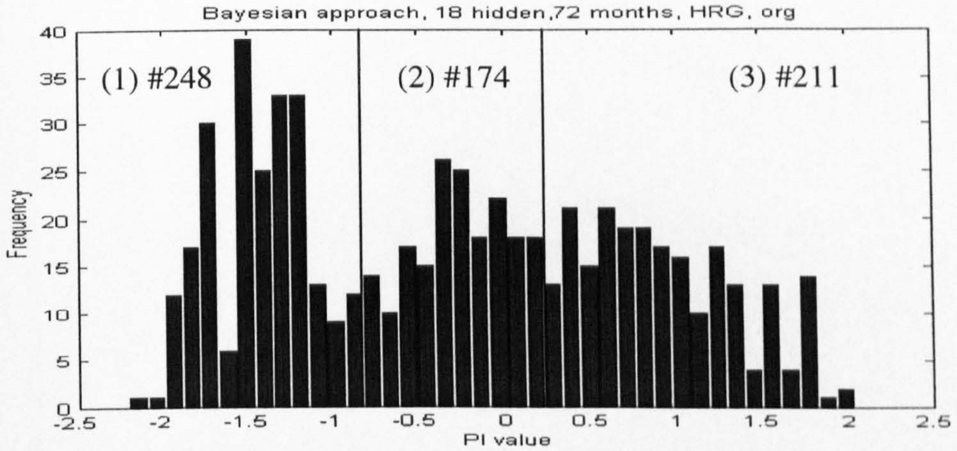
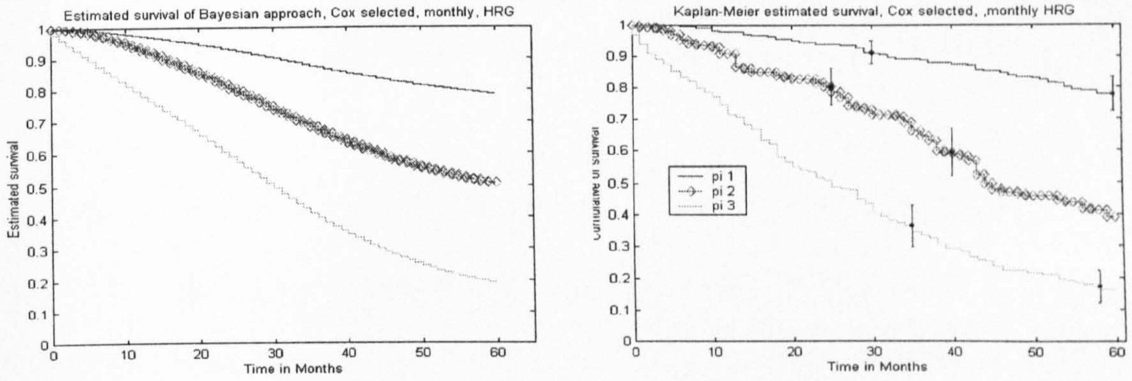
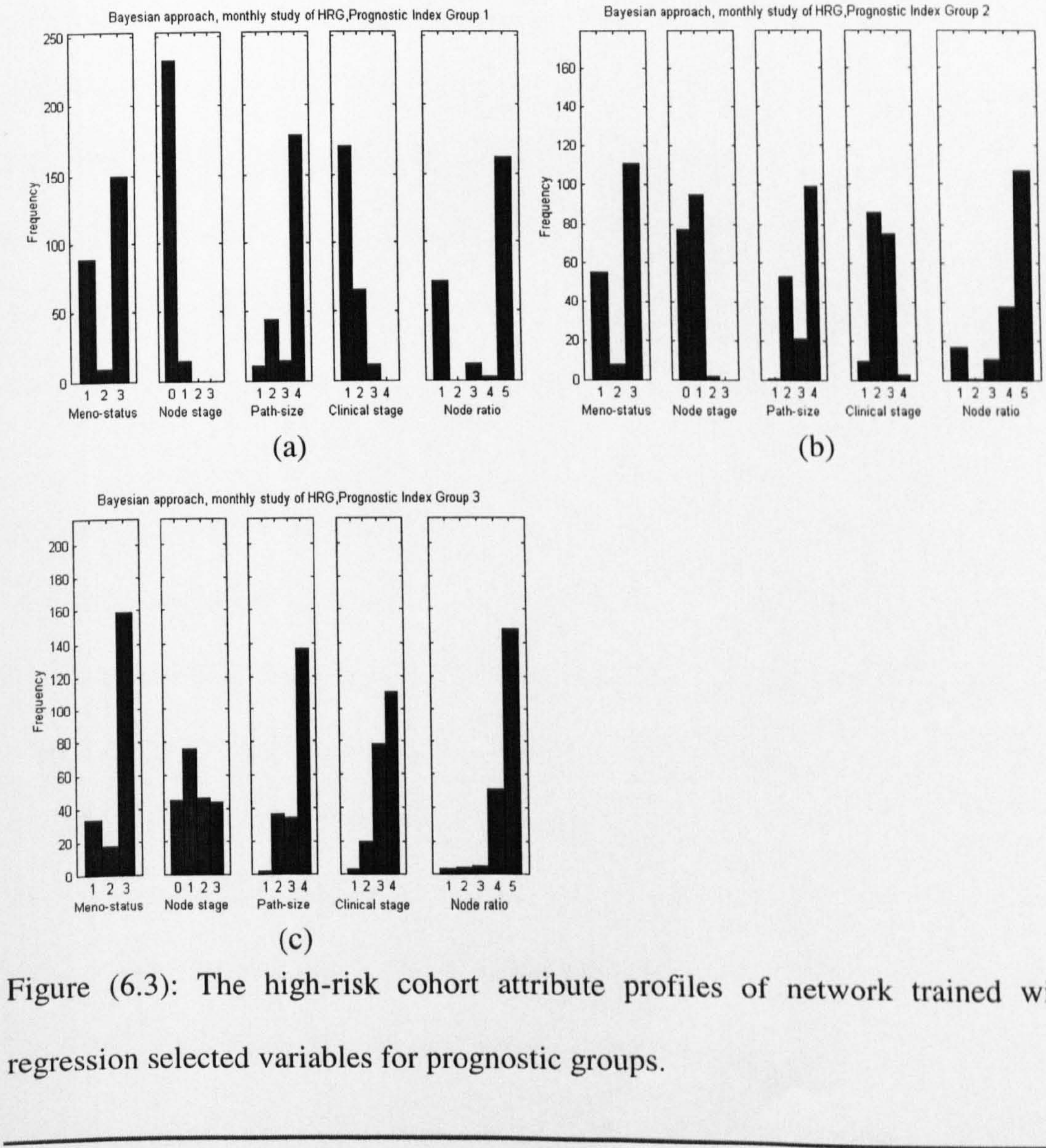


Figure (6.1): Prognostic index plot of high-risk cohort and the log rank test partitioned prognostic groups.



(a) (b)

Figure (6.2): (a) Neural networks predicted survivor function for prognostic groups using the 5 independent Cox regression selected variables and (b): The corresponding Kaplan-Meier estimate of the survivor functions.



(a) (b)

(c)

Figure (6.3): The high-risk cohort attribute profiles of network trained with Cox regression selected variables for prognostic groups.

6.2 Neural networks model selection using ARD technique

When the regularisation coefficient assigned to a particular input grows large, the attached network weights are damped down towards zero. This is how the technique of Automatic Relevant Determination (ARD) controls the irrelevant input variables from damaging the network output performance.

The process of model selection begins with including all of the independent variables in the model, resulting in a set of values for the hyper-parameter α. In this case, a group of attributes corresponding to same variable shares same value of alpha. The network output marginalisation to the mean hazard ratio and baseline attributes were also used. The variable removal criterion consists ranking the alpha values by size, then gradually removing the input variables with significantly large alpha values from the model, until no more variables could be removed without substantial detriment to model performance. This amounts to backward stepwise elimination.

The model selection for the high-risk cohort using the ARD started with all the 14 variables, excluding the four surgical variables, as described in section (4.4.1). Although the time covarates are given a large value of alpha, they are not considered as candidate for variable selection and kept in the model. Removing time from the input variables would result in a survival model with time independent hazard. Therefore survival would be exponential to time. *diameter, pathological size, clinical stage* and *number of nodes involved* were the first set of variables to be removed from the model. Finally, 6 variables were left, namely, *menopausal status, predominant site, tumour* *stage, node stage, histology* and *nodes ratio*. The details were summarised in table (6.1). The ARD selected variables are slightly different from those selected by Cox regression for this cohort, the common variables of two models being *menopausal status, node stage* and *nodes ratio*. Even though the rest of variables from two models are different, some of the variables represent similar kinds of information such as *tumour stage* and *pathologic size*. Only the *predominant site* and *histology* are selected differently in the ARD model. These newly selected variables will be tested for their predictive power in section (6.2.1).

Variables	Value of Alpha			
	(1st stage)	(2nd stage)	(3rd stage)	Final Model
Menopausal Status	7.0357	2.9952	3.1	6.443
Age Group	4.3999	3.9710	8.3102	
Predominant Site	7.6109	4.6426	3.4912	6.6387
Side	5.3807	11.7232		
Diameter	20.9377			
Tumour Stage	5.2955	2.4935	1.9996	5.1673
Node Stage	8.7874	2.6625	4.155	4.4299
Metastasis Stage	5.29336	10.54498		
Pathological Size	12.8370			
Manchester Stage	12.3407			
Histology	5.4018	2.5584	3.5155	6.3668
Nodes Involved	10.7831			
nodes ratio	5.1505	2.6657	3.3531	4.4423

Table (6.1): The value of alpha of variables involved in the different stages of ARD model selection process.

6.2.1 Network trained with the ARD selected model

The network was re-trained with the ARD selected variables, still with 18 hidden nodes, and with bias correction terms to marginalise the network output towards the average data hazard. The results are displayed in figure (6.4) to (6.6). Three distinct mortality risk groups are identified containing 244 171 and 218 patients, and the observed survivorship at month 60 are 0.72, 0.4 and 0.21 respectively. Although each of the network predicted mean survivorship falls into the confidence interval bands estimated by Kaplan-Meier, the survival estimation of prognostic group 2 is not as accurate as for the other groups, owing to the effect of marginalisation towards the overall averaged hazard. This can be solved by modelling each prognostic group separately. Also, the survival curves are not as well separated as with the neural network using Cox selected variables. However, the model is still useful to identify candidate variables that may act through interactions with other variables.

These are the variables that have similar attribute profiles for different prognostic groups, namely, ***menopausal status, predominant site*** and ***histology***. A further analysis base on these variables for variable interactions is summarised in section (6.3).

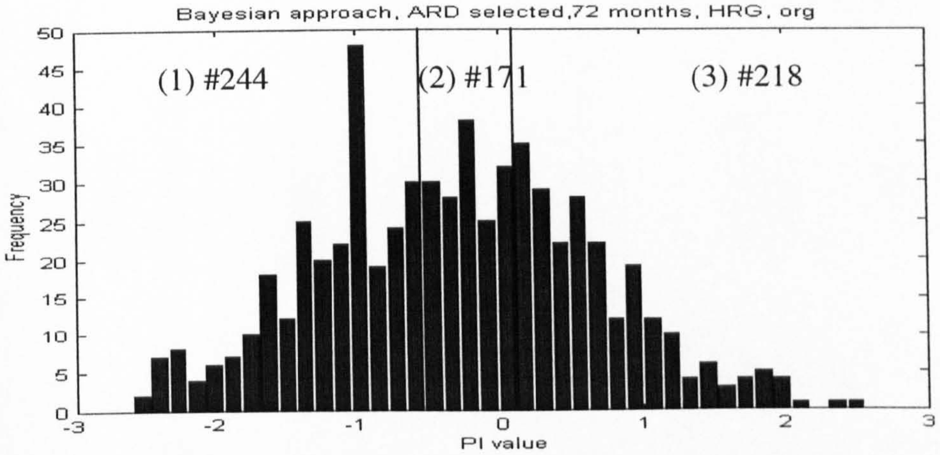
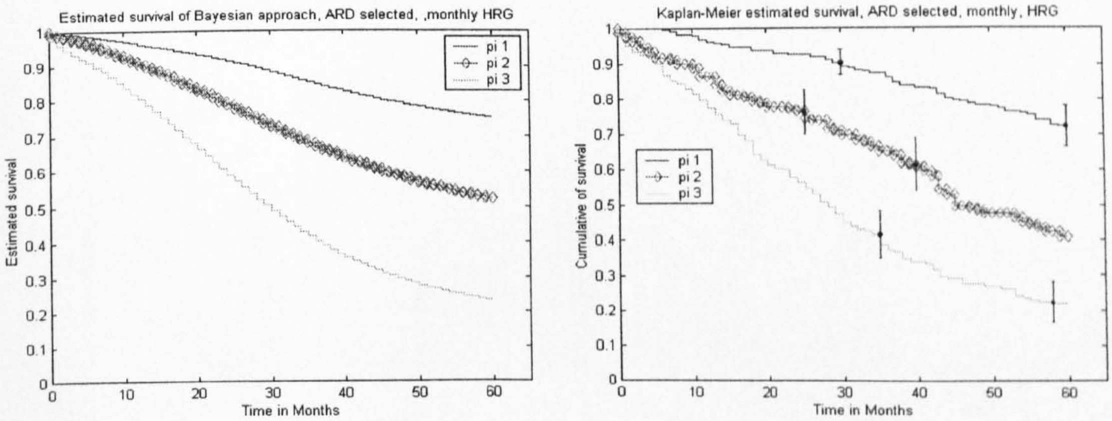
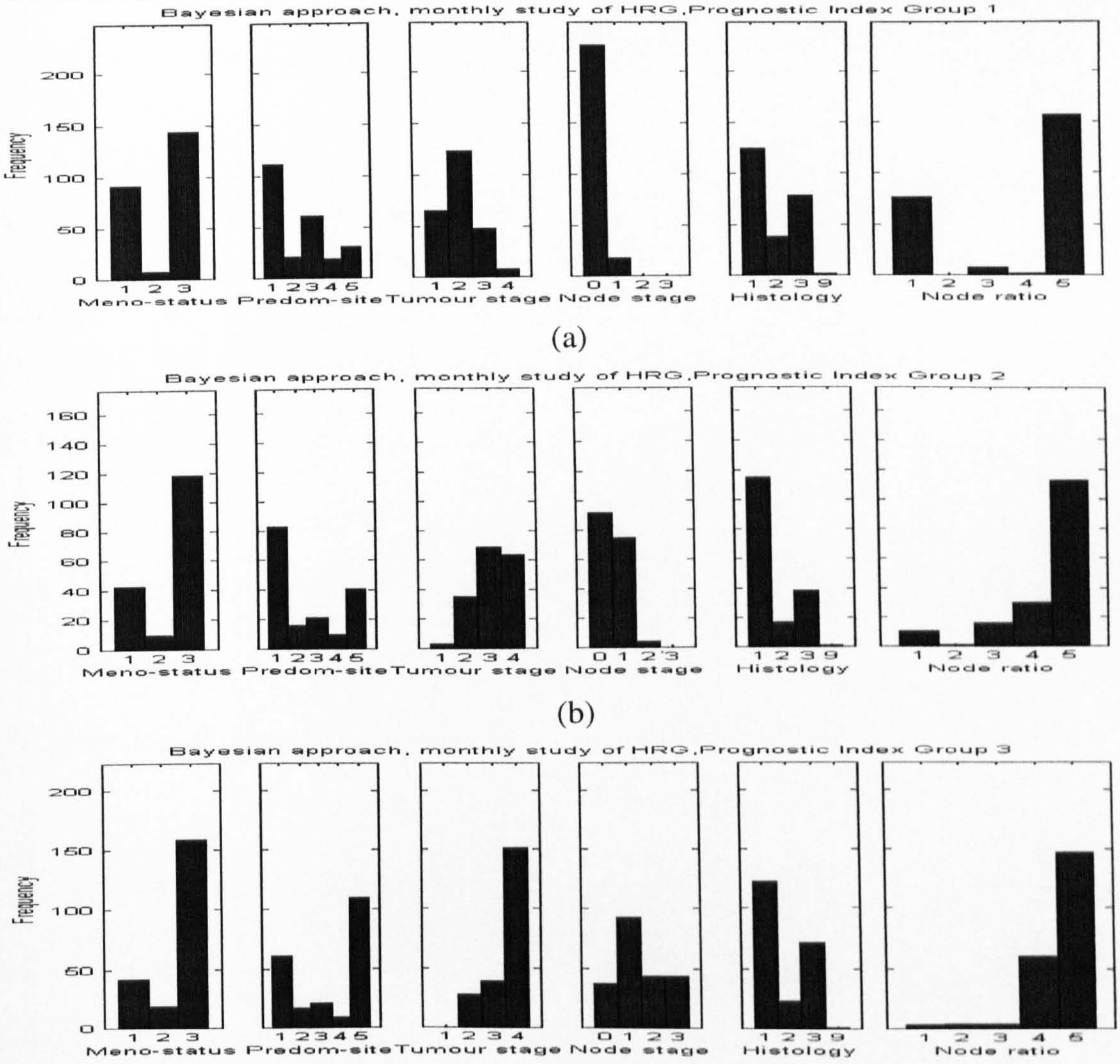


Figure (6.4): The 3 partitioned prognostic groups using the ARD selected variables for the high-risk cohort.



(a) (b)

Figure (6.5): The neural network predicted survivor function using the 6 ARD selected variables for the prognostic groups and the corresponding Kaplan-Meier estimate of the survivor functions.



(a)

(b)

(c)

Figure (6.6): The attribute profiles of network trained with ARD selected variables for prognostic groups.

6.3 Assessing pairwise interactions for the high-risk cohort

Although the network trained with the ARD selected variables did not produce better results than the network trained with the Cox selected variables, the variables selected differ and this may indicate the presence of variable interactions, which influence the constitution of prognostic groups, and the prediction of survival for individuals.

It is very difficult to determine explicitly the functional form of the implicit interaction between variables in the neural network model. Cox regression allows interaction terms to be included explicitly in the model, and hence tested for their statistical significance.

The 6 ARD selected variables were divided into 2 categories, specific and non-specific variables. Non-specific variables have similar attribute profiles for different prognostic groups in figure (6.6), which specific variables show a gradual transition in the univariate profiles across the prognostic groups. The variables classified as non-specific variables are ***predominant site***, ***menopausal status*** and ***histology***. The process of identifying interactions term for the high-risk cohort using Cox regression was divided into three stages. Firstly, the 3 pairwise interactions between the non-specific variables were used alone to model the data; then the 3 pairs of specific variables were used as interaction terms; finally, the 9 cross-terms from the two sets of variables were used to model the data. The results from these studies are listed in table (6.2).

Significant interaction terms between **non-specific** variables	Significant interaction terms between **specific** variables	Significant interaction terms between **non-specific** and **specific** variables
*Histology * Predominant site*	*Nodes ratio * Node stage*	*Menopausal status * tumour stage*
Hsitology Menopausal status*	Nodes ratio * Tumour stage	*Menopausal status * nodes ratio*
		*Histology * Tumour stage*
		*Histology * node stage*

Table (6.2): The significant interaction terms of ARD selected variables for high-risk cohort.

Following this preliminary pre-filtering of candidate interaction pairs, the 8 pairs of variable interaction terms were put together with the 14 independent variables and the optimal Cox model was identified by forward stepwise model selection. This resulted in a final model comprising an independent variable, *clinical stage*, together with the pairwise interaction between *nodes ratio * tumour stage*.

The results of Cox regression fitted with this optimal model with a 3-fold cross validation are summarised in figures (6.7) to (6.9). The Cox prognostic indexes, including a contribution from the interaction term, and the Cox predicted and Kaplan-Meier estimates of the mean survival function for each the 3 partitioned mortality risk groups, are shown together with the variable attribute histograms for each of prognostic group. Furthermore, figure (6.10) displays the values of *tumour stage* for each value of *nodes ratio* in each of the 3 prognostic groups, to show the interaction between these two variables. There is an improved separation between the expected survival of prognostic groups, and the patients allocated to each group are, 214, 278 and 139, respectively. Moreover, the survivor probability at 60 months for the highest risk group has gone down below 0.1. Also all of the variables show clearly different attribute profiles over prognostic groups, as shown in figure (6.9). The results have confirmed that a variable interaction is present in the data and the interaction term *tumour stage* * *nodes ratio* contributes to the identification of high-risk prognostic group.

The ARD technique is a useful tool for seeking variable interactions in the data and combining with the Cox selected independent variables, yields a powerful predictive model. The predictive power of the ARD selected model for high risk patients can also be tested once more by introducing the network identified special high-risk patients group from the low-risk cohort to the high-risk cohort. The analysis is reported in the next section.

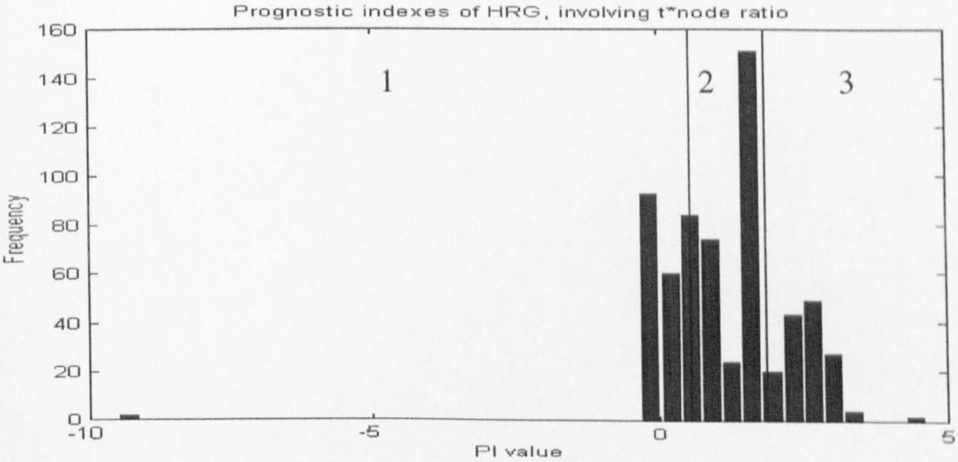
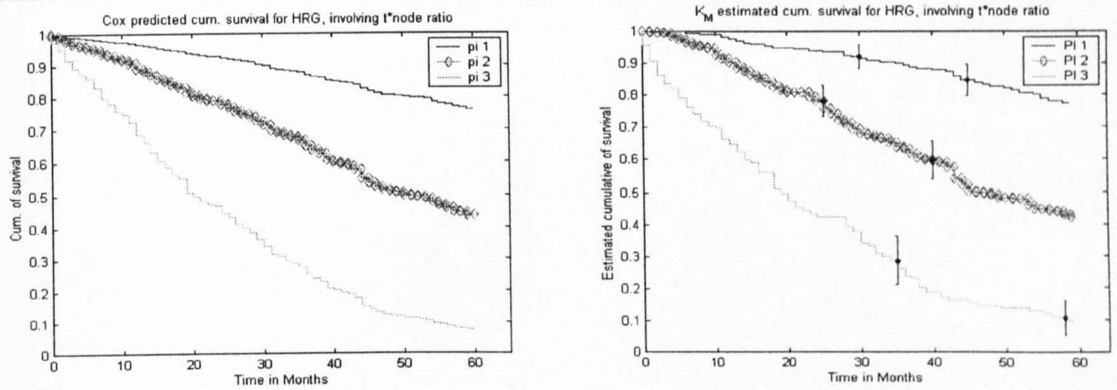


Figure (6.7): Cox regression defined prognostic groups involving a pairwise interaction between variables *tumour stage* and *nodes ratio* for the high-risk cohort.



(a) (b)

Figure (6.8) (a): The Cox regression predicted survival function involving the interaction term *tumour stage * nodes ratio* for prognostic groups from high-risk cohort and (b) the corresponding Kaplan-Meier estimates survival function. The prognostic groups are well separated and the survivorship of prognostic group 3 has driven toward below 0.1.

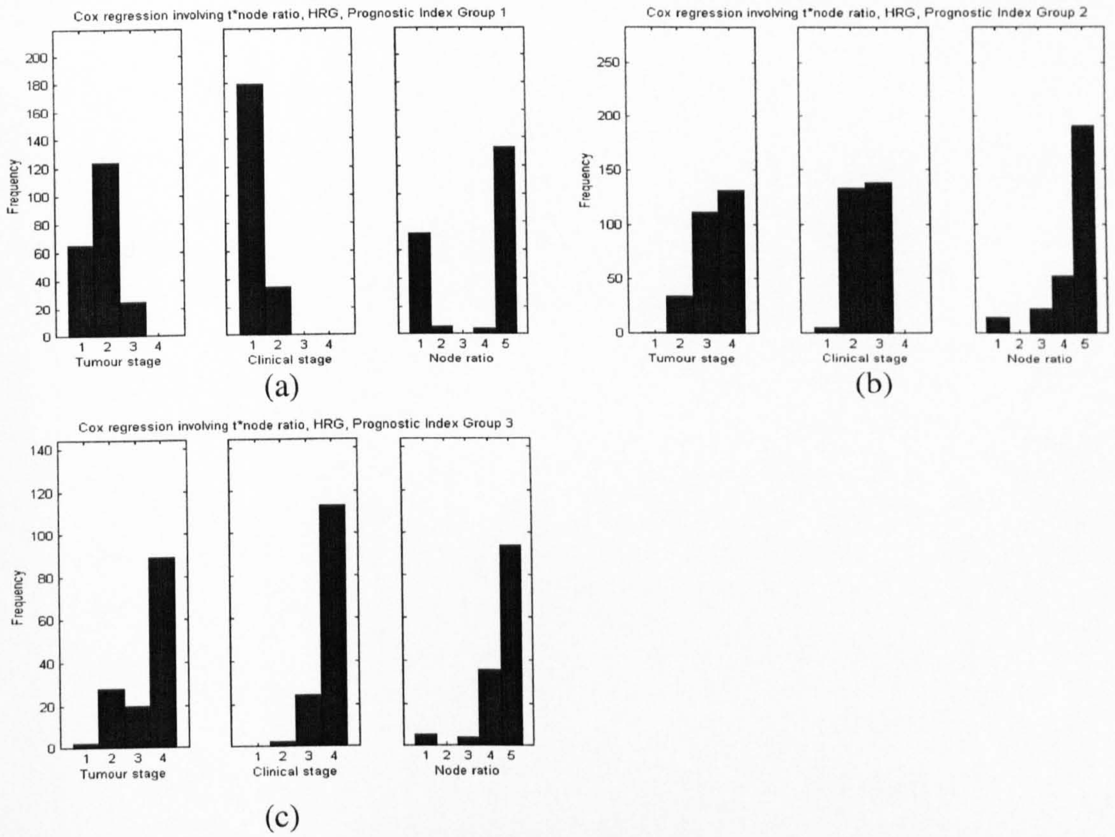


Figure (6.9): The variable profiles of Cox regression involving variable interaction term *turmour stage* and *nodes ratio* for high-risk cohort. Each of the variables has shown a clear attribute profile over the prognostic groups.

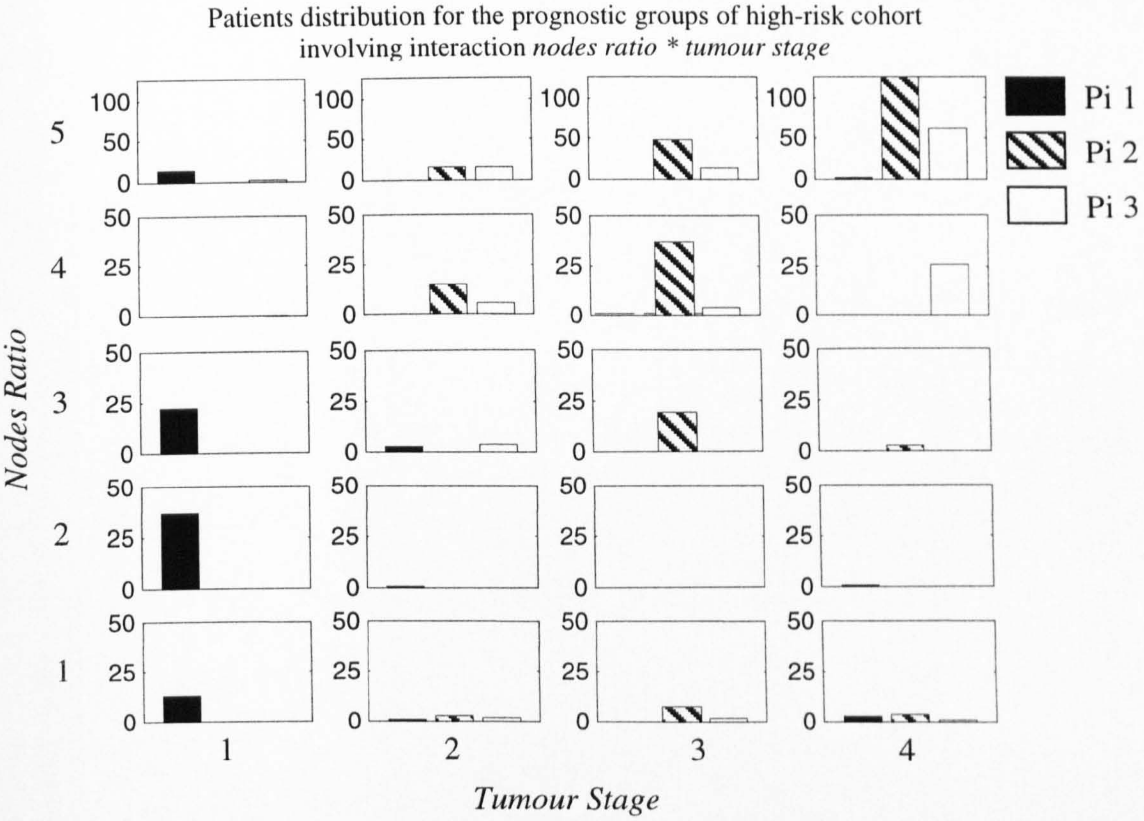


Figure (6.10): Distribution of patients for prognostic groups over the interaction term *tumour stage* and *nodes ratio*, where *nodes ratio* 5 represents the missing data attribute.

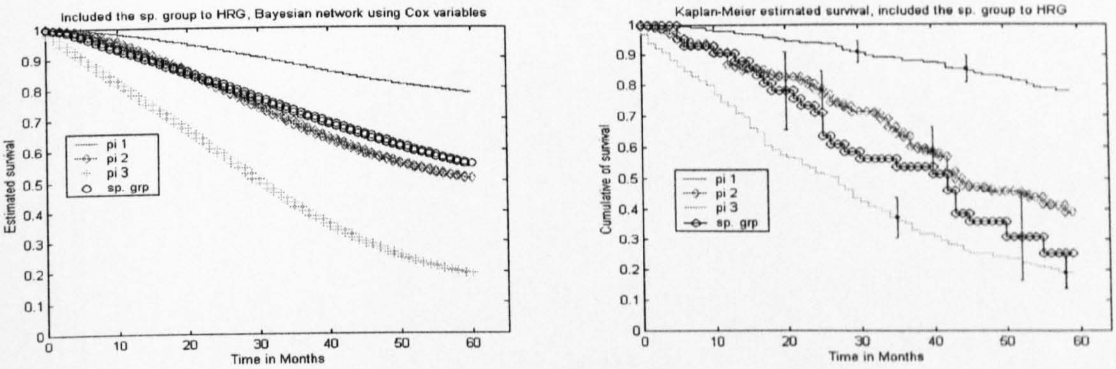
6.4 Inclusion of the highest risk group from the low-risk cohort into the high-risk cohort

The predictive models fitted to the high-risk cohort were also applied to the high mortality group from the low-risk cohort. Both the ARD selected variables, and the Cox selected variables were used with neural network, in order to compare the predictive power of each set of variables for these patients. The analysis does not involved retraining of the network, simply re-uses the weights calculated by the network trained previously with the ARD and Cox selected variables with 3-fold cross validation. The estimated hazards for this high mortality group of each model were gathered and averaged over 3 set of results. Then the mean estimated survivorship over 60 months was projected each onto the graph along with the other prognostic groups generated by the ARD and Cox selected model for the high-risk cohort.

Figures (6.11) - (6.12) display network predicted survivor function for the high mortality group together with the original prognostic groups and the corresponding Kaplan-Meier estimates of survival using Cox and ARD selected variables respectively. The figures show that the Cox selected variables do not accurately predict the survivorship of these 42 patients. In contrast, the ARD selected variables show an accurate prediction and similar prediction to the PLANN model developed for the low-risk cohort. The survival curve of these patients crosses over the survivorship of the 3 prognostic groups in the high-risk cohort that ARD generated. During the first 7 months following surgery, this group of patients displays a similar survivorship as group 1, then the survival gradually decreases from month 8 to month 42 where it crosses over group

2, and finally joins with group 3 from month 42 onward, reducing survival probability to 0.26 at month 60.

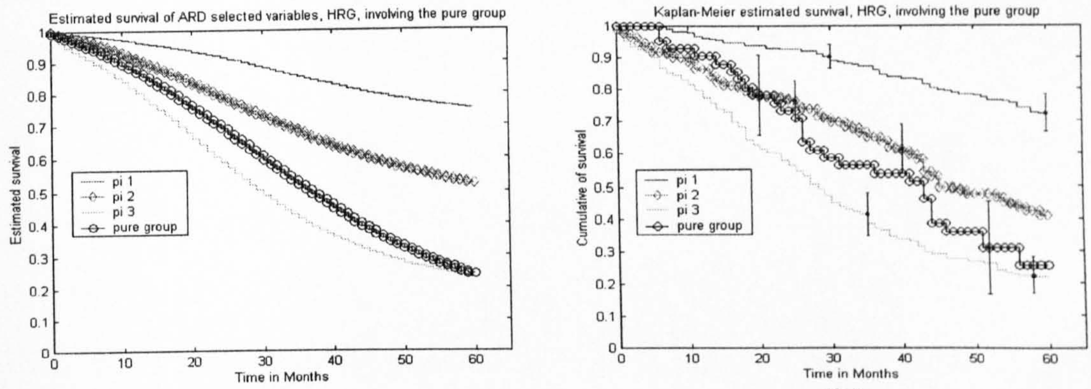
When examining the three sets of models carefully, the Cox selected for the low-risk cohort, the Cox and ARD selected for the high-risk cohort, only the *histology* existed in both of the ARD selected model and the Cox model for the low-risk cohort, but was absent in the Cox selected model for the high-risk cohort. The *histology* could be one of potential variables that describe the survivorship of the high mortality group, or indeed, a variable interaction could be the alternative possibility. Further investigation is summarised in section (6.5.1).



(a)

(b)

Figure (6.11): (a) The neural networks predicted survivor function for the high-risk cohort prognostic groups and the specific group using Cox selected variables and (b): the corresponding Kaplan-Meier estimate of the survivor functions.



(a) (b)

Figure (6.12): (a) The neural networks predicted survivor function for the high-risk cohort prognostic groups and the special group using ARD selected variables and (b): the corresponding Kaplan-Meier estimate of the survivor functions.

6.5 The key variable to model the high mortality group in the low-risk cohort

Previous results suggested that *histology* is the factor that best separates this special high mortality group from the rest of the low-risk cohort. The following contents in this section are the test of this statement and to consider possible variable interaction terms within the low-risk cohort.

6.5.1 Detecting the variables that **histology** *interacted with*

The 3 possible pairwise interaction terms from the Cox selected variables for the low-risk cohort, *histology* node stage*, *histology * pathological size* and *histology * nodes ratio* were included to the Cox regression model selection procedure alone to model the data, and resulting that they were all significantly responded to the survivorship of the data. These three pairs of variables were then entered into the model selection process again, together with the 14 independent variables. The final model contains *pathological size, histology, nodes ratio* and *histology * node stage*. Figure (6.13) displays the Cox partitioned prognostic groups using this model. The Cox predicted survivorship for the mortality risk groups are displayed in figure (6.14) together with the corresponding Kaplan-Meier estimate of the survivor functions. The attribute histograms of prognostic groups are displayed in figure (6.15) and figure (6.16) shows the attribute histograms within the interaction term.

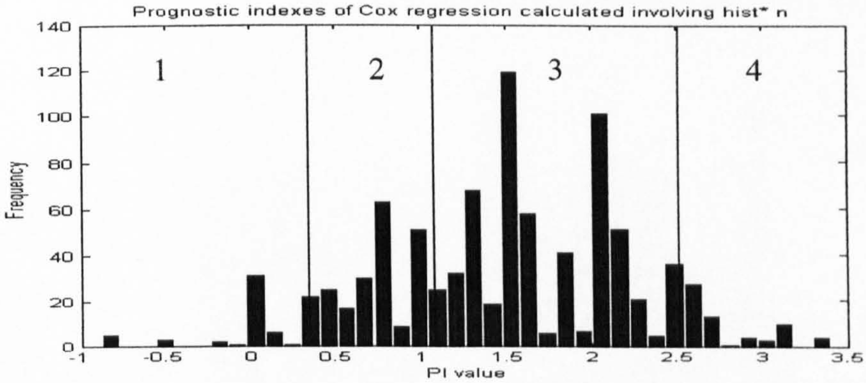
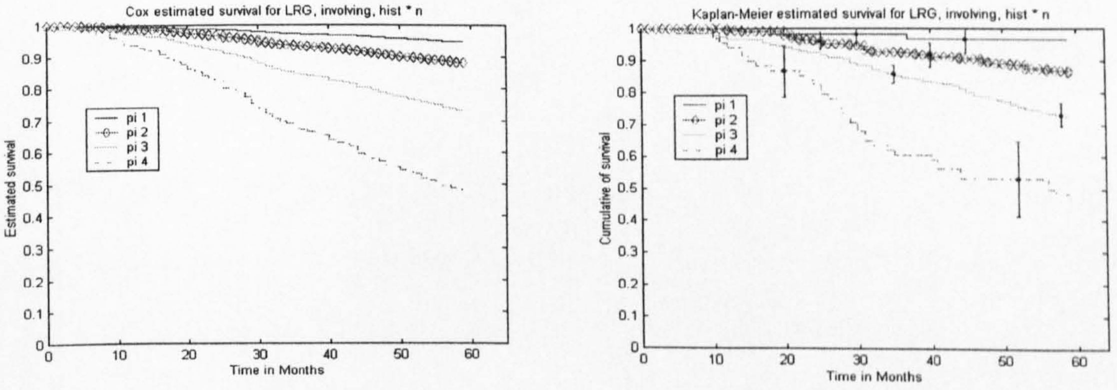


Figure (6.13): Cox regression partitioned mortality risk groups for the low-risk cohort involving the interacted variables *histology* * *node stage* and contained 61, 207, 579 and 68 patients respectively.



(a) (b)

Figure (6.14): (a) Cox regression predicted survivor function for the 4 prognostic groups of low-risk cohort involving the interacted variables *histology* * *node stage* and (b) the corresponding Kaplan-Meier estimate of the survivor functions.

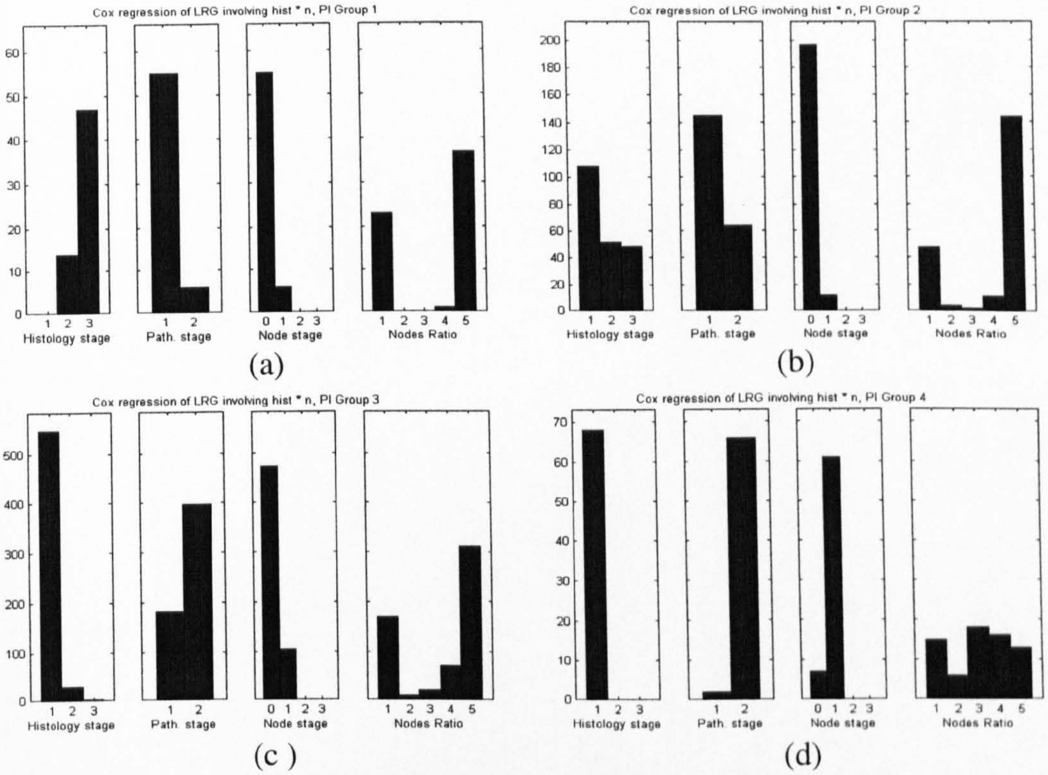


Figure (6.15): Attribute histograms of the model involving *histology* * *node stage* interaction term over prognostic groups for the low-risk cohort. The prognostic group 1 contained fewer patients and the profile is similar to the result of neural networks using Cox selected variables for the low-risk cohort.

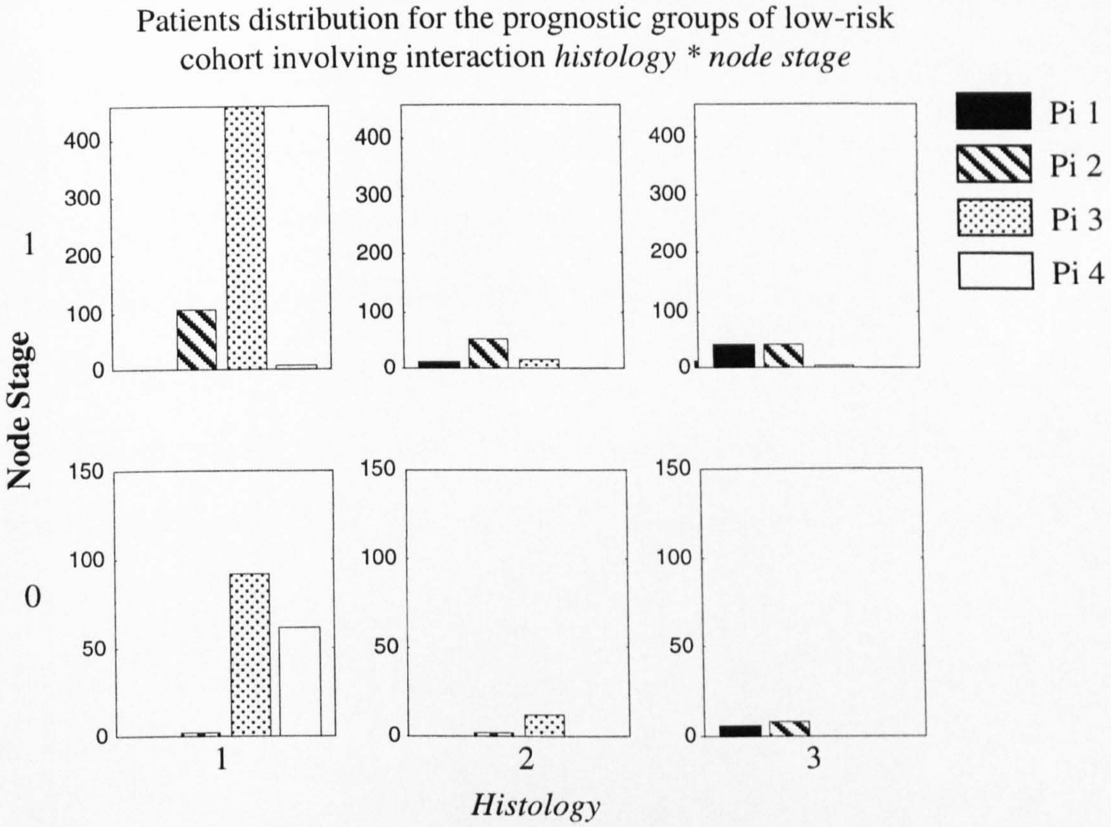
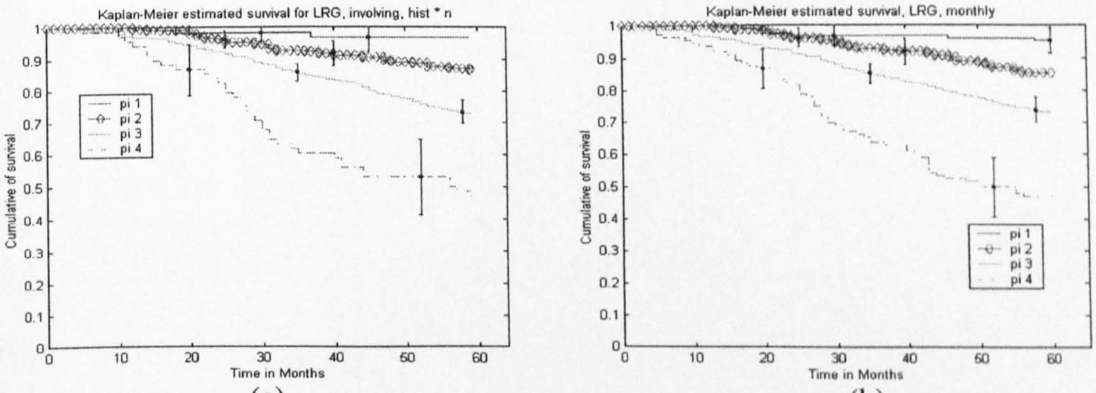


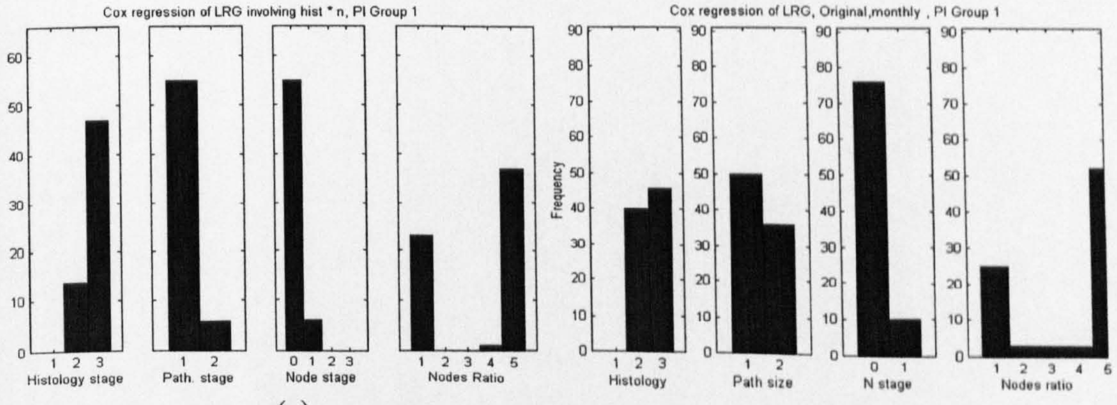
Figure (6.16): Distribution of patients over the interaction term *histology* and *node stage*.

The results show that *histology * node stage* actually is not the main factor that describes the survivorship of the high mortality group as expected, even though *histology* is one of the predictive variables being selected. However, *histology * node stage* contributes to the highest survival group when comparing the expected survivorship of prognostic groups with the result without involving the interaction term. The expected survivorship involving, *histology * node stage* of prognostic group 1 has been improved better, closer to 1, and also the corresponding attributes profiling even more specific and contained less patients, as illustrated in figure (6.17)-(6.18).



(a) (b)

Figure (6.17): (a) The Cox predicted survivor function for the prognostic groups involving *histology * node stage* and (b) without involving interaction term.



(a) (b)

Figure (6.18): (a) Attribute histograms of prognostic group 1 involving interaction term *histology * node stage* and (b) without involving interaction term.

6.5.2 Determine the interaction term that describes the survivorship of the high mortality group in the low-risk cohort

Since the *histology* is not the factor that describes the survivorship of the high mortality group, the remaining possible interaction terms from the ARD selected model for the high-risk cohort are *nodes ratio * node stage*, *nodes ratio * pathological size* and *pathological size *node stage*. The results show that *nodes ratio * node stage* and *pathological size * node stage* were both significant to the survivorship of the data. The final model selected with forward stepwise elimination contained three independent variables, namely, *node stage*, *pathological size* and *histology*, and an interaction term *nodes ratio * node stage*.

This model is fitted to the Cox regression with 5-fold cross validation again. Figure (6.19) displays the Cox partitioned prognostic groups involving interaction term *nodes ratio * node stage*. The Cox regression predicted survivor function for the mortality risk groups with the corresponding Kaplan-Meier estimate of the survivor functions are displayed in figure (6.20). The attribute profiles for prognostic groups are shown in figure (6.21).

The results indicate that the interaction term *nodes ratio * node stage* is the factor that best differentiates the survivorship of the high mortality group. The attribute profiles are more specific over the prognostic groups. However, the attribute histograms for the prognostic group 1 is not as specific as the model involving *histology * node stage*. Including the two interaction terms together in a model may retain good differentiation for the highest survival group, which *histology * node stage* contributed to, and the lowest survival group, where the interaction between *nodes ratio * node stage* is significant.

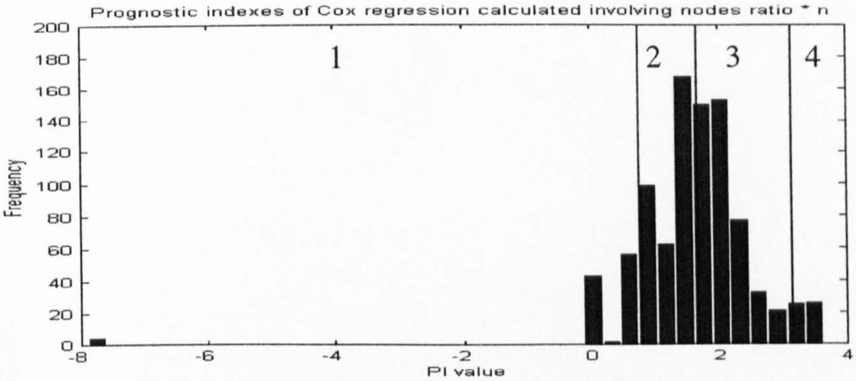
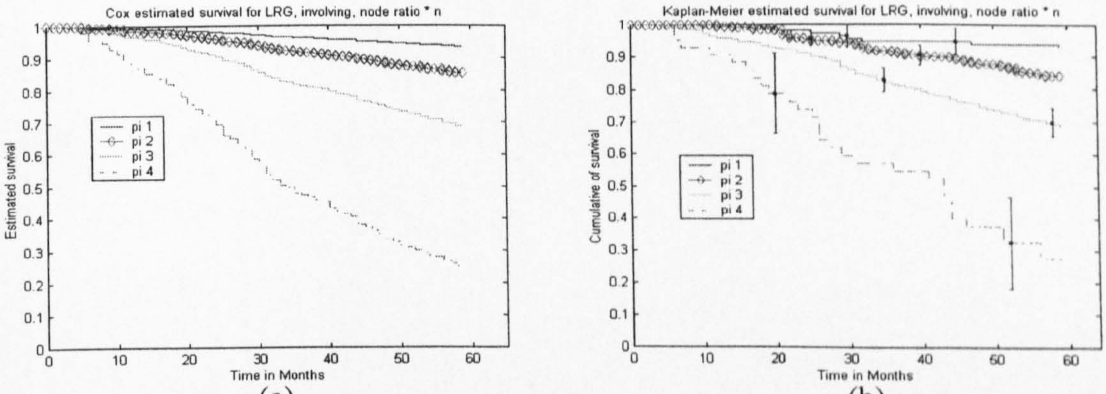
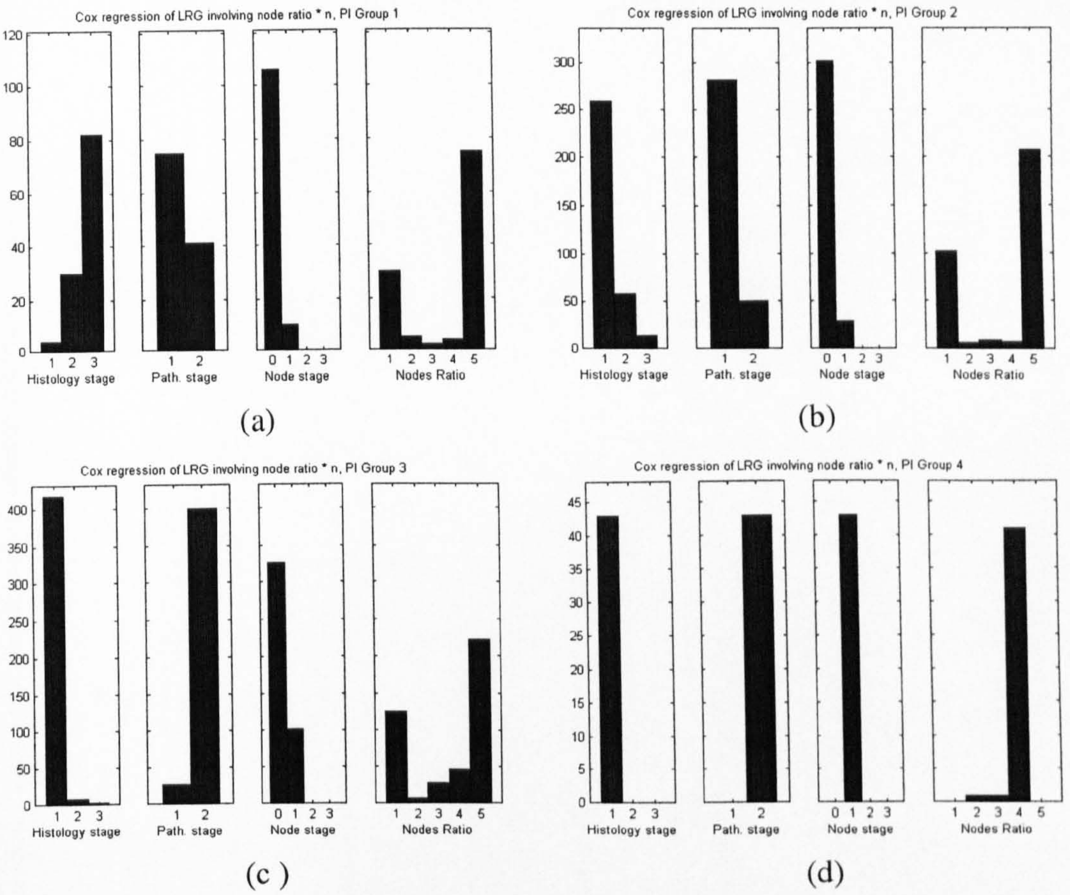


Figure (6.19): The Cox regression partitioned mortality risk groups for the low-risk cohort using the model involved interaction term *nodes ratio * node stage* and contained 116, 331, 427 and 43 patients respectively.



(a) (b)

Figure (6.20): (a) Cox regression predicted survivor function for the 4 prognostic groups in low-risk cohort involving the interaction term *nodes ratio * node stage* and (b) the corresponding Kaplan-Meier estimate of the survivor function.



(a)

(b)

(c)

(d)

Figure (6.21): Attribute histograms of the prognostic groups involving *nodes ratio* * *node stage* interaction for the low-risk cohort.

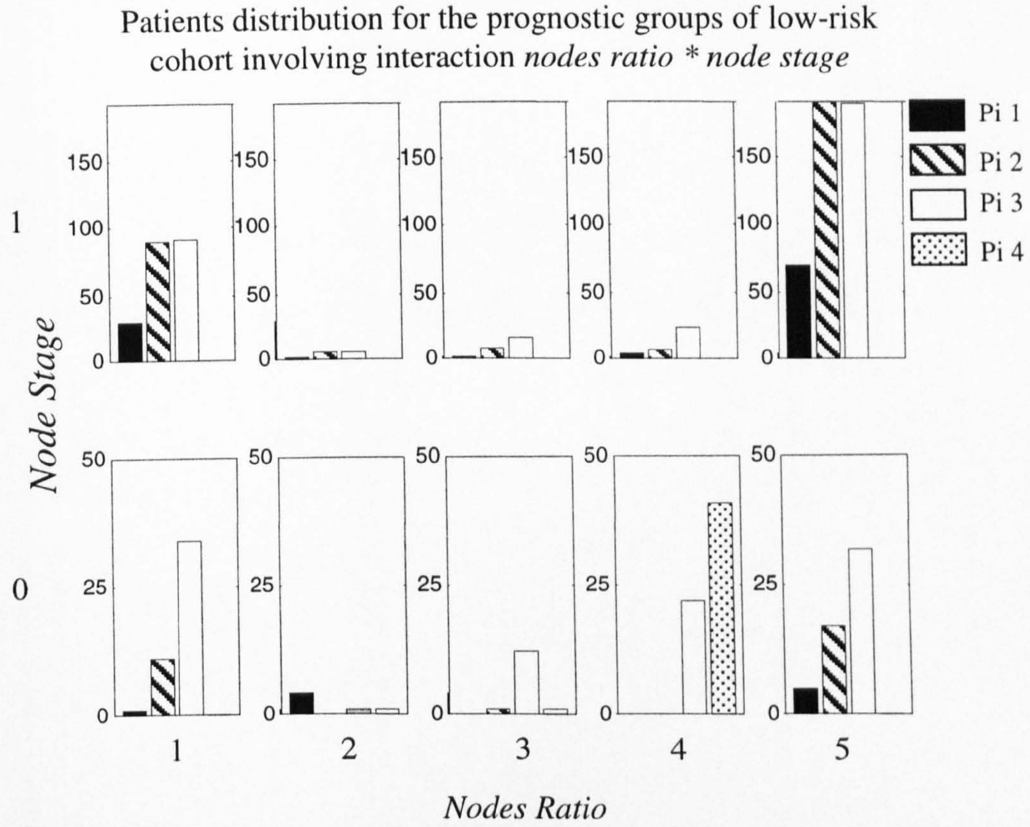


Figure (6.22): Distribution of patients over interaction term *nodes ratio* and *node stage,* where *nodes ratio* 5 is represents the missing data attribute.

6.5.3 Modelling both interaction terms together for the low-risk cohort

A new model was selected for the low-risk cohort, which contained the interaction *histology * node stage* and *nodes ratio * node stage*, in addition to *histology* and *pathological size*. Forward stepwise elimination was employed once again. The Cox regression results are presented in figure (6.23) to (6.25) in the same sequence as the previous study but not including the cross distribution of the interacting variables.

The results with Cox regression including 2 interactions terms are poorer than with the neural network using the same base variables. In other words, the ARD model with *histology, pathological size, node stage* and *nodes ratio* has a better separation between the prognostic groups. In the Cox model with two interaction terms, the prognostic group 1 and 4 contained more patients and were less specific. It appears that in the linear Cox regression model, the two interaction terms are working against each other. Although the factors that describe the survivorship of the highest survival and lowest survival groups have been identified individually, their power cannot be merged and expanded using Cox regression. It demonstrated the strength of neural networks in applying the interactions selectively to different prognostic groups.

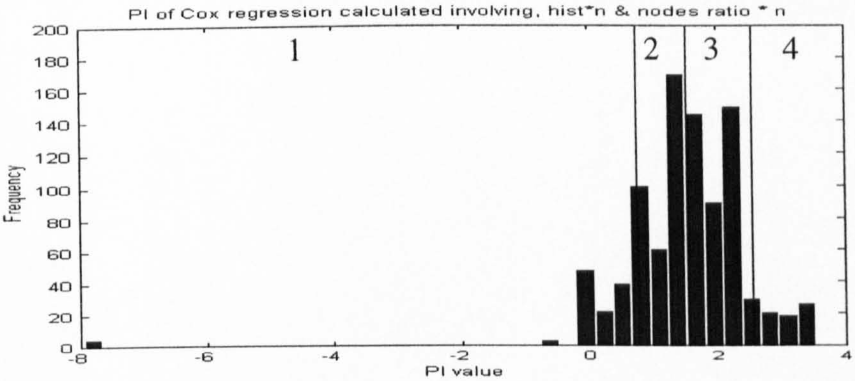
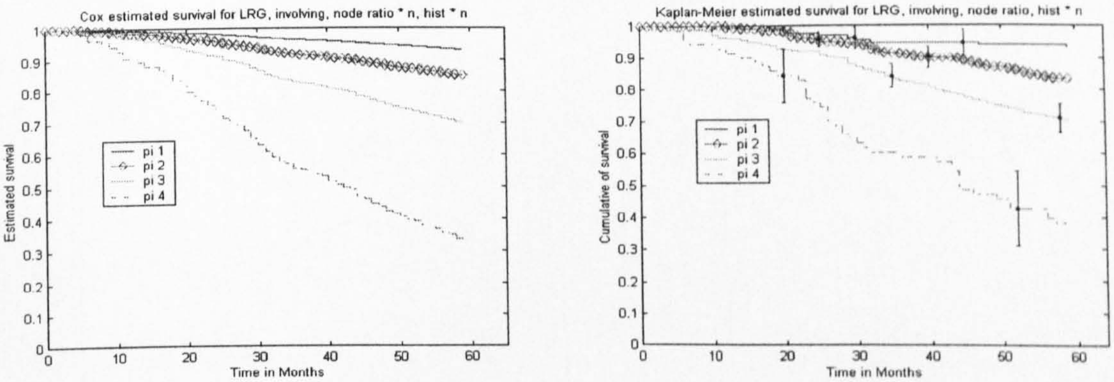


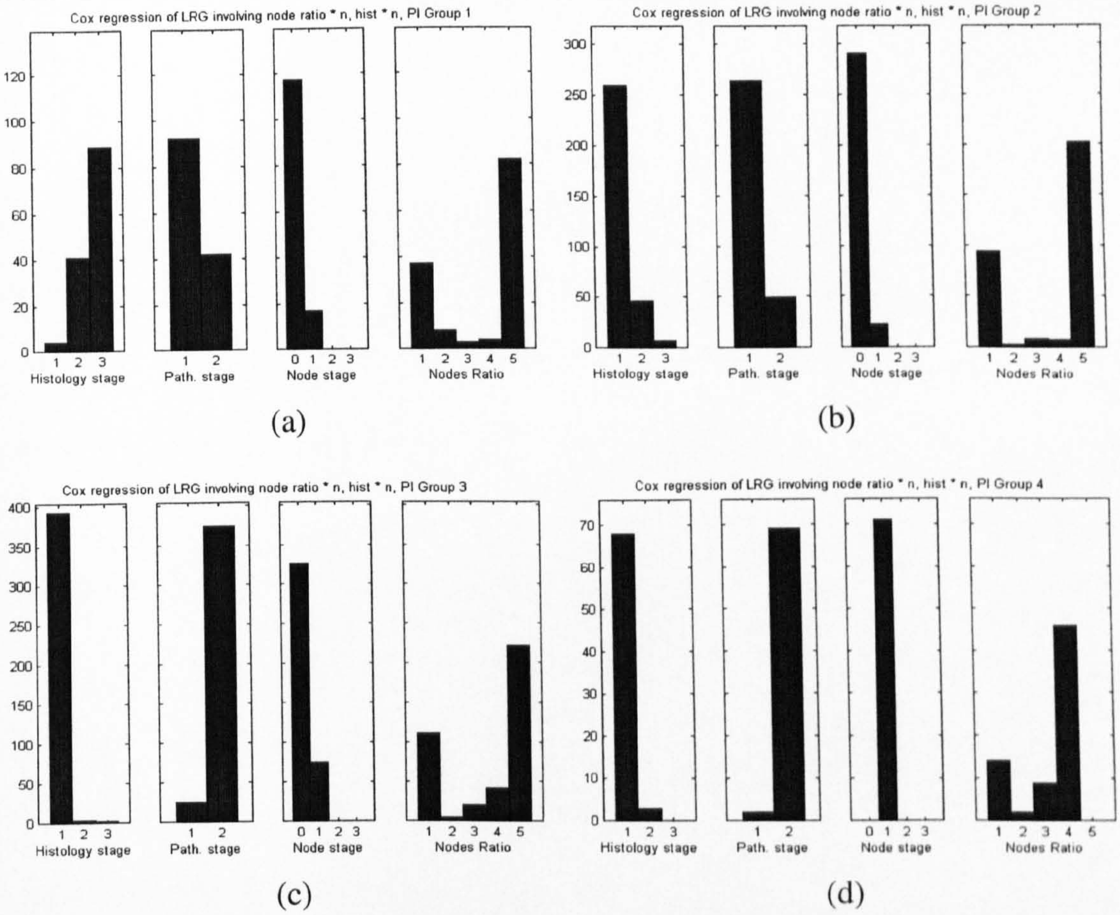
Figure (6.23): Cox regression partitioned mortality risk groups involving the two interaction terms *nodes ratio * node stage* and *histology * node stage,* contained 134, 313, 399 and 71patients respectively.



(a)

(b)

Figure (6.24): (a) Cox regression predicted survivor function for the prognostic groups involving the interaction terms *nodes ratio * node stage* and *histology * node stage* and (b) the corresponding Kaplan-Meier estimate of the survivor functions. The outcome does not meet the expectation. Neither the specified low survival nor the high survival group was displayed.



(a)

(b)

(c)

(d)

Figure (6.25): The attribute histograms for the prognostic groups which involving interaction terms *nodes ratio * node stage* and *histology * node stage* for the low-risk cohort.

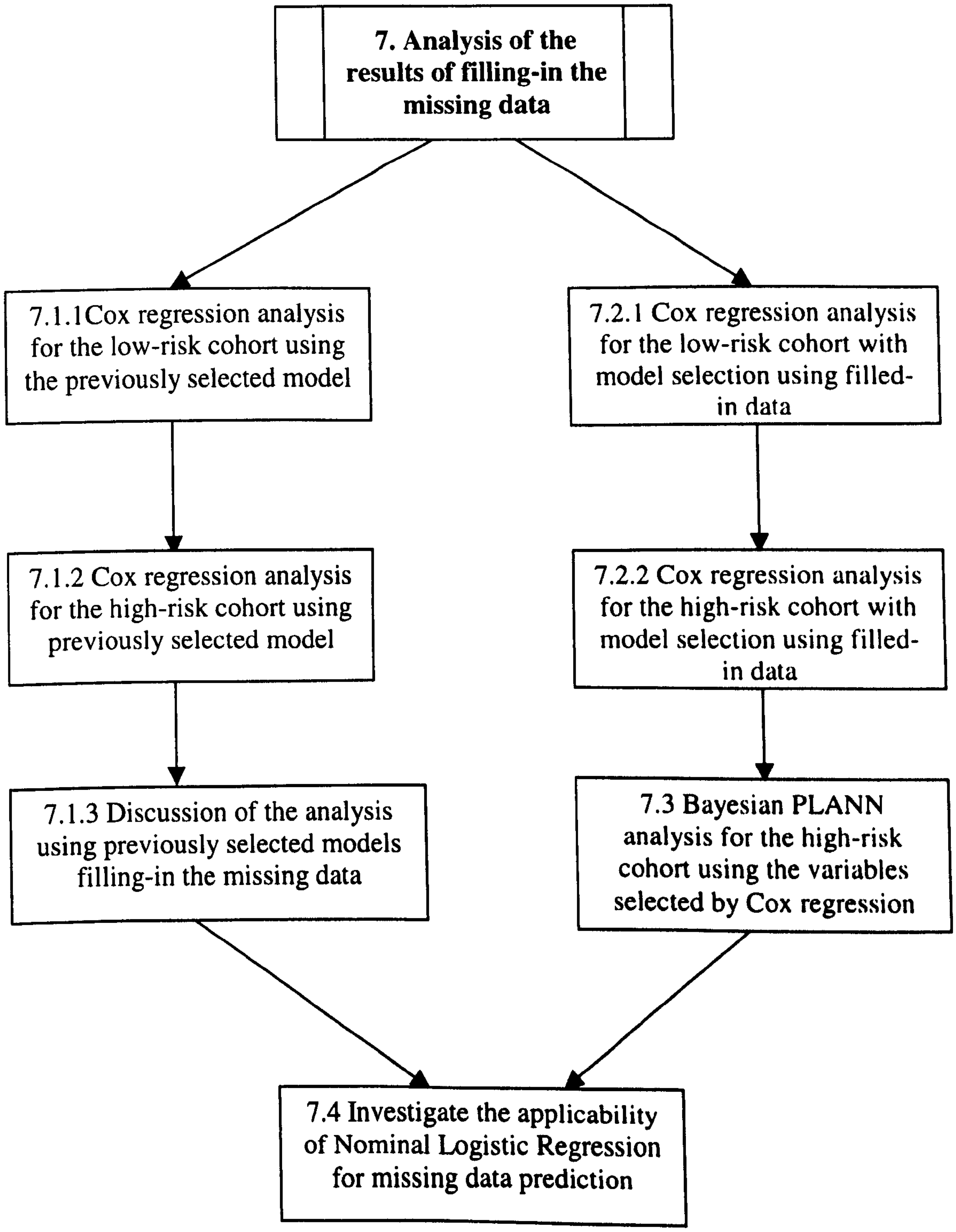
6.6 Conclusion

In this chapter, the power of neural network ARD model selection towards seeking and handling variable interactions has been demonstrated. It was also combined with Cox regression, to find an optimal model for each of the cohort that separates well the prognostic groups and has specific attribute profiles.

In the high-risk cohort, the final model only contained one independent variable, *clinical stage*, together with the ARD identified interaction term *tumour stage* nodes ratio.* Note that *clinical stage* is a composite variable, combining *tumour*, *node* and *metastasis* stages. The survival predictions generated by this model are the best among all models for the high-risk cohort.

The situation in the low-risk cohort is not straightforward. Although the interaction terms contributing to the highest and lowest mortality groups have been identified separately, the results show that they cannot be put together into a single Cox model. The best result is given by the neural network model trained with Cox selected variables for the low-risk cohort.

Chapter 7


7. Analysis of the results of filling-in the missing data
7.1.1Cox regression analysis for the low-risk cohort using the previously selected model
7.2.1 Cox regression analysis for the low-risk cohort with model selection using filled-in data
7.1.2 Cox regression analysis for the high-risk cohort using previously selected model
7.2.2 Cox regression analysis for the high-risk cohort with model selection using filled-in data
7.1.3 Discussion of the analysis using previously selected models filling-in the missing data
7.3 Bayesian PLANN analysis for the high-risk cohort using the variables selected by Cox regression
7.4 Investigate the applicability of Nominal Logistic Regression for missing data prediction

7 Effectiveness of predicting missing data using logistic regression

In section (3.2.2), nominal logistic regression was proposed to predict missing data from a set of complete variables, using feed forward variable selection.

Within this chapter, the effectiveness of the predicted values is evaluated. The process starts by modelling each of the cohorts using Cox regression. Variable interactions are not considered at this stage. The substantial improvement is sought for accuracy of survivorship prediction, differentiation between the survival of prognostic groups, and characteristic attribute profiles. Finally, the Cox regression analysis was benchmarked with the Bayesian PLANN model using the high-risk cohort, where interactions between predictor variables have caused difficulties for Cox regression.

7.1 Cox regression analysis of filled-in missing data using previously selected models

7.1.1 Cox regression analysis of the low-risk cohort with missing data filled-in using nominal logistic regression

The 4 variables listed in section (3.2.2) were the variables that contained large amount of missing values. Some of the other variables also contain missing data, but only a small fraction, these cases (77 cases) were discarded, in order to be used to predict the 4 variables contained missing values, leaving 1473 completed cases.

Since *pathological size* is a clinical criterion for patient cohort allocation, predicting it changes the composition of the low and high-risk cohorts, as follows

	Before filling-in missing data	**After filling-in missing data**
Low-risk cohort	917	1070
High-risk cohort	633	403

Table (7.1): Patients allocation of each cohort before and after filling-in missing data.

The model used for the initial analysis was the Cox selected model for the original low-risk cohort, using 5-fold cross validation. Figure (7.1) displays the ranked prognostic indexes and the 3 mortality risk groups that the log-rank test partitioned, containing contained 215, 645 and 210 patients respectively. The distribution of the prognostic indexes has shifted to lower values when compared with the original results in figure (4.9). Figure (7.2) shows the Cox predicted mean survivorship for each prognostic group with the corresponding Kaplan-Meier estimate of the survival functions. The survivorship of the highest and lowest survival group at 5 years is 0.9 and 0.52.

The attribute histograms for the prognostic groups are displayed in figure (7.3). The survival prediction for the prognostic groups is a slightly better match of the Kaplan-Meier curves than the previous Cox results where the missing data were treated as separate categories. The confidence intervals calculated by Kaplan-Meier estimation for each prognostic group are smaller and the attribute histograms also show better profiling, except for *histology*, where the profile is less specific than previously.

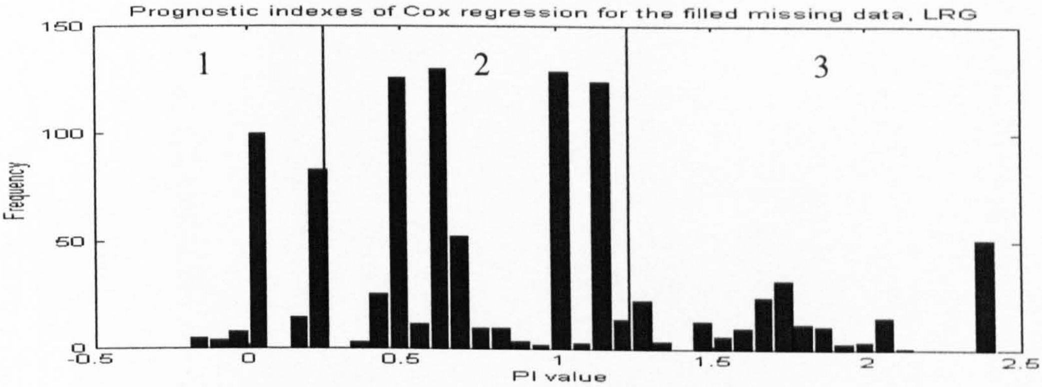
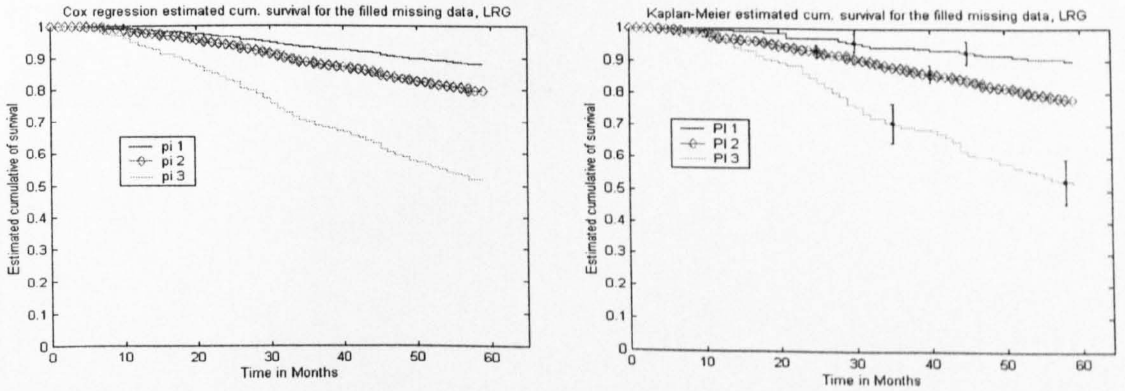


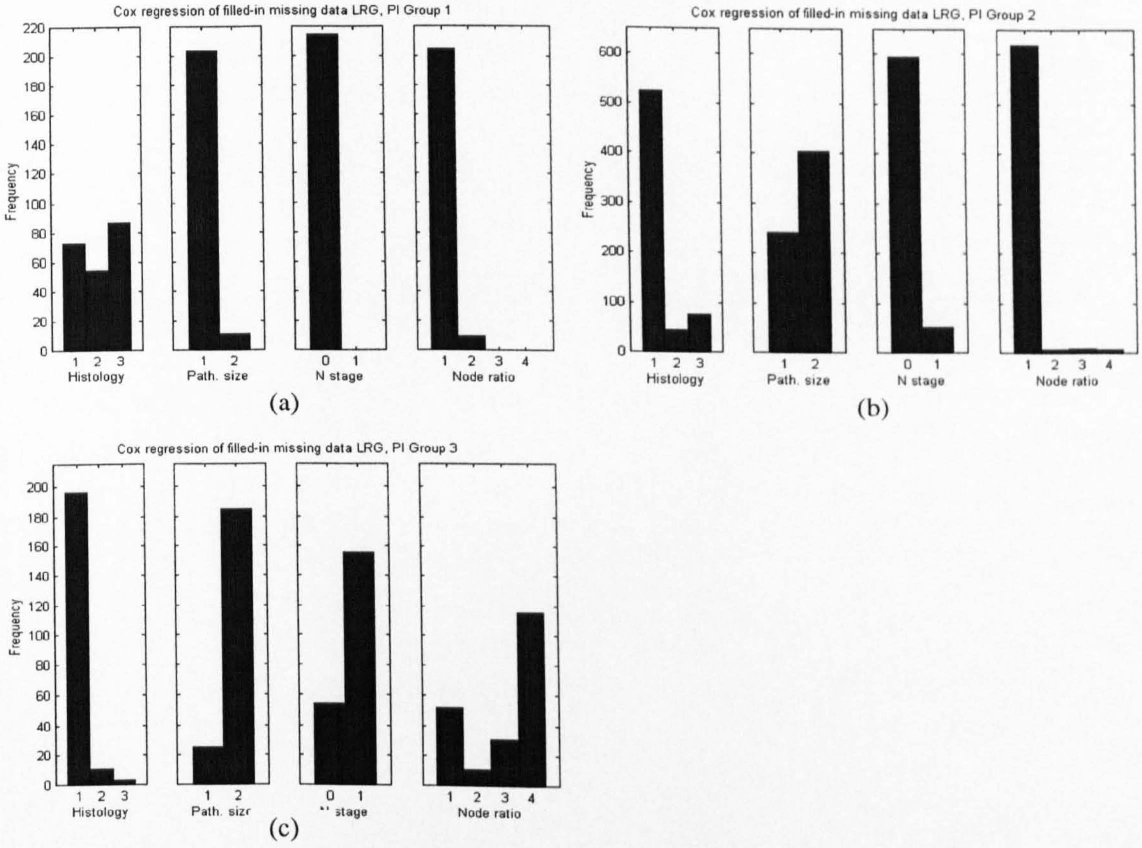
Figure (7.1): Cox calculated prognostic indexes for the filled-in low-risk cohort and log-rank test partitioned prognostic groups.



(a)

(b)

Figure (7.2): (a) The Cox the Cox predicted mean survivorship for prognostic groups with (b) the corresponding Kaplan-Meier estimate of the survival functions.



(a)

(b)

(c)

Figure (7.3): The attribute histograms of prognostic groups for the filled-in low-risk cohort.

The original 4 prognostic groups were merged into 3 groups. Table (7.2) displays the allocation of sujects from the original prognostic group into the 3 new groups. In the prognostic group 1 of the original low-risk cohort, 80 records of *nodes ratio* were labelled as missing and were predicted as category 1 (<=20% of positive nodes from removal). There are 125 *nodes ratio* missing values in the prognostic group 2 and 3 records were deleted in the prediction process, 118 of the remaining missing values were predicted as category 1 and the rest were predicted as category 2 (20%-30%). A total of 258 records were predicted out of the original 268 *nodes ratio* missing values in the prognostic group 2, the number of records predicted to be *nodes ratio* category 1, 2, 3 and 4 are 248, 4,1 and 5 respectively. Finally in the prognostic group 4, 17 and 14 records were predicted as category 1 and 4 respectively, 1 record was deleted. The following table (7.3) summarises the attribute details of the variables in the model and the allocation of the predicted records of *nodes ratio* in the filled-in low-risk cohort analysis, The reallocation of unknown *nodes ratio* was followed a specific pattern.

Original Prognostic Group	**Prognostic Group after Prediction**	**Number of records**
1	1	75
1	2	50
1	3	1
2	1	70
2	2	105
2	3	6
3	1	2
3	2	390
3	3	72
4	1	0
4	2	1
4	3	110
Total:		**882**

Table (7.2): Patient allocation to prognostic groups after filling-in missing data.

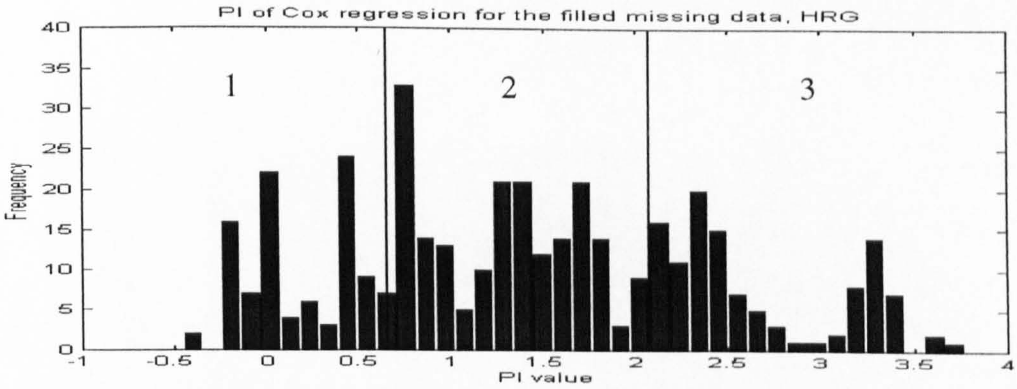
Prognostic group of the original low-risk cohort analysis	**Corresponding prognostic group in the analysis of the filled-in low-risk cohort**	***Histology***	***Pathological size***	***Node Stage***	***Nodes Ratio***
1(80)	1(51)	2	1	0	0
		3	1	0	1
	2(29)	3	1	1	1
		3	2	0	1
2(122)	1(60)	1	1	0	1,2
		2	2	0	1,2
	2(62)	1	1	0	1
		2	1	1	1
		2	2	0	1
		3	2	0,1	1
3(258)	1(1)	1	1	0	2
	2(248)	1	1,2	0	1
	3(9)	1	1	1	1,2
		2	2	0	1
4(31)	3(31)	1	2	1	1,4

* The figure in the bracket indicates the number of cases that *nodes ratio* have been predicted for the associated prognostic group.

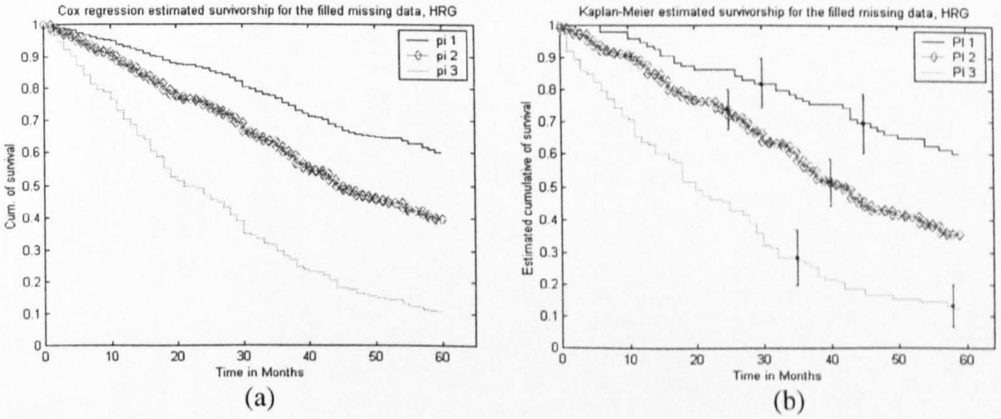
Table (7.3): The allocation of missing *nodes ratio* in each prognostic group before and after prediction and their corresponding variable details.

7.1.2 Cox regression analysis of filled-in missing data for the high-risk cohort

The filled-in high-risk cohort now contains only 403 records since some of the filled-in values for *pathological size* have placed patients into the low-risk cohort. A total of 188 cases were transferred to the filled low-risk cohort after following the clinical cohort separation criteria applied to the predictions of *pathological size*. Within these records, 89 and 99 records were filled with values 1 and 2 respectively. Furthermore, 68, 99 and 21 out of 188 records were allocated into prognostic group 1,2 and 3 by the prognostic index for the low-risk cohort with the missing data filled-in. The predictive modelling for the filled high-risk cohort was carried out with 3-fold cross validation using the previously selected Cox model for the high-risk cohort, without interactions term. The patients were partitioned into 3 mortality risk groups with 105, 167 and 131 patients, respectively, as shown in figure (7.4). A similar proportion of patients was allocated to each prognostic group as for the previous Cox model for this cohort. The results in figure (7.5) also show a better prediction of the survivorship function for each prognostic group. The mean survivorship of the 3 groups at 5 years is 0.60, 0.35 and 0.12. The attribute histograms, figure (7.6), show similar profiles to using missing values as separate attributes except for *pathological size* and *nodes ratio*, which do not show well differentiated profiles.

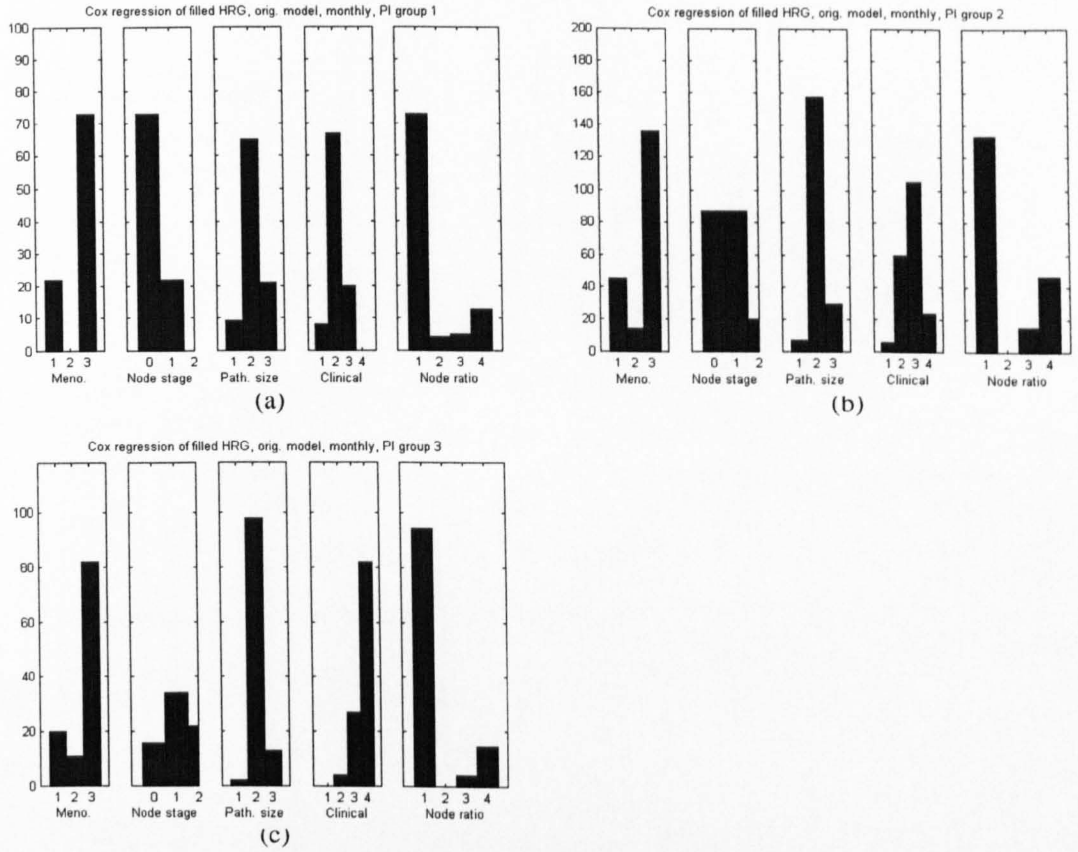


Figure(7.4): The partitioned prognostic groups for the filled-in high-risk cohort using the Cox regression.



(a) (b)

Figure (7.5): (a) The Cox predicted survivorship for prognostic groups of filled-in high-risk cohort and (b) the corresponding Kaplan-Meier estimate of the survivor functions.



(a)

(b)

(c)

Figure (7.6): The attribute histograms for prognostic groups of filled-in high-risk cohort.

Table (7.4) displays the location of patients in the original prognostic groups and after the filling-in of *pathological size* and *nodes ratio*. In the Cox selected original model, *pathological size* and *nodes ratio* contained missing values is 196 and 257 cases, respectively. Table (7.5) and (7.6) summarise the number of records of each predicted value and their position in the new prognostic groups. A total of 6 missing values of *pathological size* were filled with value 1 and 190 records were filled with value 2. As expected, all the missing values of *nodes ratio* in this cohort were filled with value 1.

Original prognostic group	Prognostic group after filling-in the missing value	Number of records
1	1	47
1	2	1
1	3	0
2	1	43
2	2	146
2	3	2
3	1	5
3	2	48
3	3	111
	Total:	**403**

Table (7.4): The patients allocation to prognostic groups after filling-in the missing data.

Predicted value of *pathological size*	Prognostic group after filling-in the missing value	Number of records
1	1	1
1	2	4
1	3	1
2	1	27
2	2	83
2	3	80
	Total:	**196**

Table (7.5): The predicted value of *pathological size* and their location in prognostic groups.

Predicted value of *nodes ratio*	Prognostic Group after filling-in the missing value	Number of records
1	1	50
1	2	117
1	3	90
	Total:	**257**

Table (7.6): The predicted value of *nodes ratio* and their location in prognostic groups.

7.1.3 Discussion of the filled-in missing data analysis using Cox regression

Patients in the low-risk cohort were merged into 3 prognostic groups from the original 4 groups, which leads to allocation of patients differently to prognostic groups from the previous result. Hence, the attribute profiling looks more specific than the previous result in the low-risk cohort analysis. The different allocation of patients and the better attribute profiling are also repeated in the filled-in high-risk cohort analysis. In general, the accuracy of survival prediction was improved for the two sets of result and the filled-in values have not affected the distribution of the attributes in each prognostic group. However, some of the variables are not showing clearly differentiated attribute profiles. This possibly indicates that some of the variables in the models are no longer relevant to the data, hence, searching for a new optimal model for each of cohort is needed and summarised in section (7.2). Or else, it shows that interactions between variables are becoming more important, reported in section (6.5).

7.2 Independent model selection for the filled-in data

7.2.1 Model selection for the filled-in low-risk cohort

The Cox selected model for the low-risk cohort with missing data filled-in is similar to the Cox selected model for the low-risk cohort where the missing data were treated as separate category. Variable *age group* was selected additionally and *nodes ratio* was replaced by the *number of nodes involved,* which is a related variable. The new model contains *node stage, histology, pathological size* and *number of nodes involved* as well as *age group.* This was selected by forward stepwise elimination without variable interactions. Table (7.7) summarises the log likelihood and AIC values as each variable is entered into the model. The *age group* was the last variable entering the model and the reduction of AIC value from the last model was also the smallest. If there is a need of reducing the number of variables in the model, the *age group* would be the potential candidate.

Variables entering the model	$-2\log\hat{L}$ **values**	**Degrees of freedom**	**AIC value**
Number of nodes involved	3421.25	3	3430.25
Node stage	3404.01	4	3416.01
Pathological size	3387.233	5	3402.233
Histology	3378.392	7	3399.392
Age group	3371.784	9	3398.784

Table (7.7): The log likelihood and AIC value of each variable entering to the model which is selected for the completed low-risk cohort.

The prediction of survivorship was carried out with Cox regression using a 5-fold cross validation. As a result, a total of 4 prognostic groups are partitioned as showed in figure (7.7). Figure (7.8) displays the Cox predicted mean survivorship for prognostic groups with the corresponding Kaplan-Meier estimate survivor function and also, the attribute histograms for prognostic groups are shown in figure (7.9).

The predictions from the newly selected model are almost identical to those predicted with missing values treated as separate attributes. By including the *age group* in the model, one more prognostic group was partitioned, compared with figure (7.1). The mean survivorship of prognostic groups at 5 years is 0.9, 0.73, 0.75 and 0.4. The attribute profiles also show that the attribute profiling for each variable is more specific, when compared with figure (7.3).

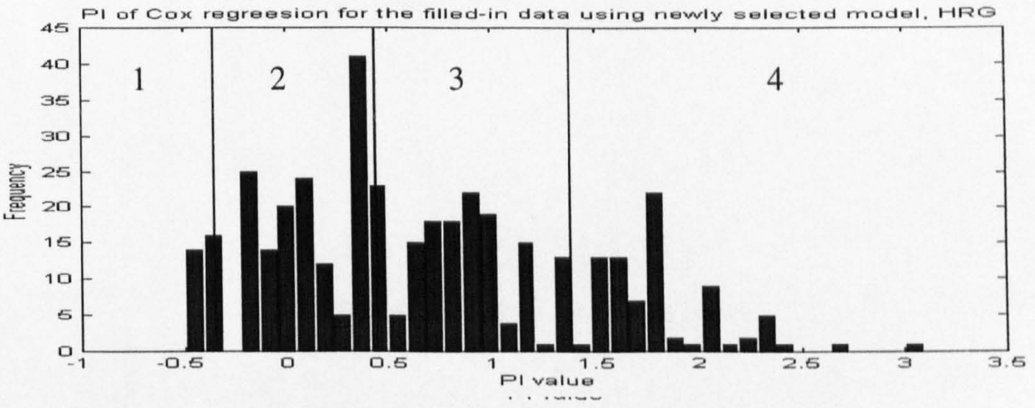
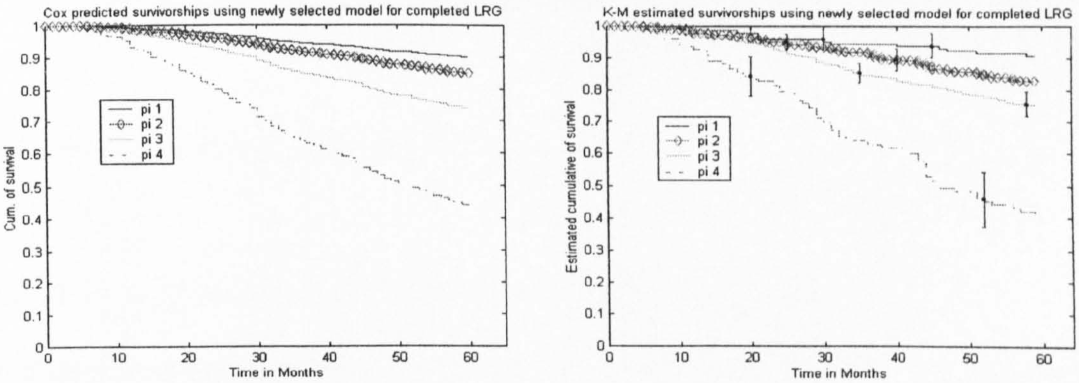


Figure (7.7): The partitioned prognostic groups by the Cox regression using the newly selected model for the completed low-risk cohort.



(a) (b)

Figure (7.8): The Cox predicted survivorship for the prognostic groups of completed low-risk cohort using the newly selected model and the corresponding Kaplan-Meier estimate survivor function.

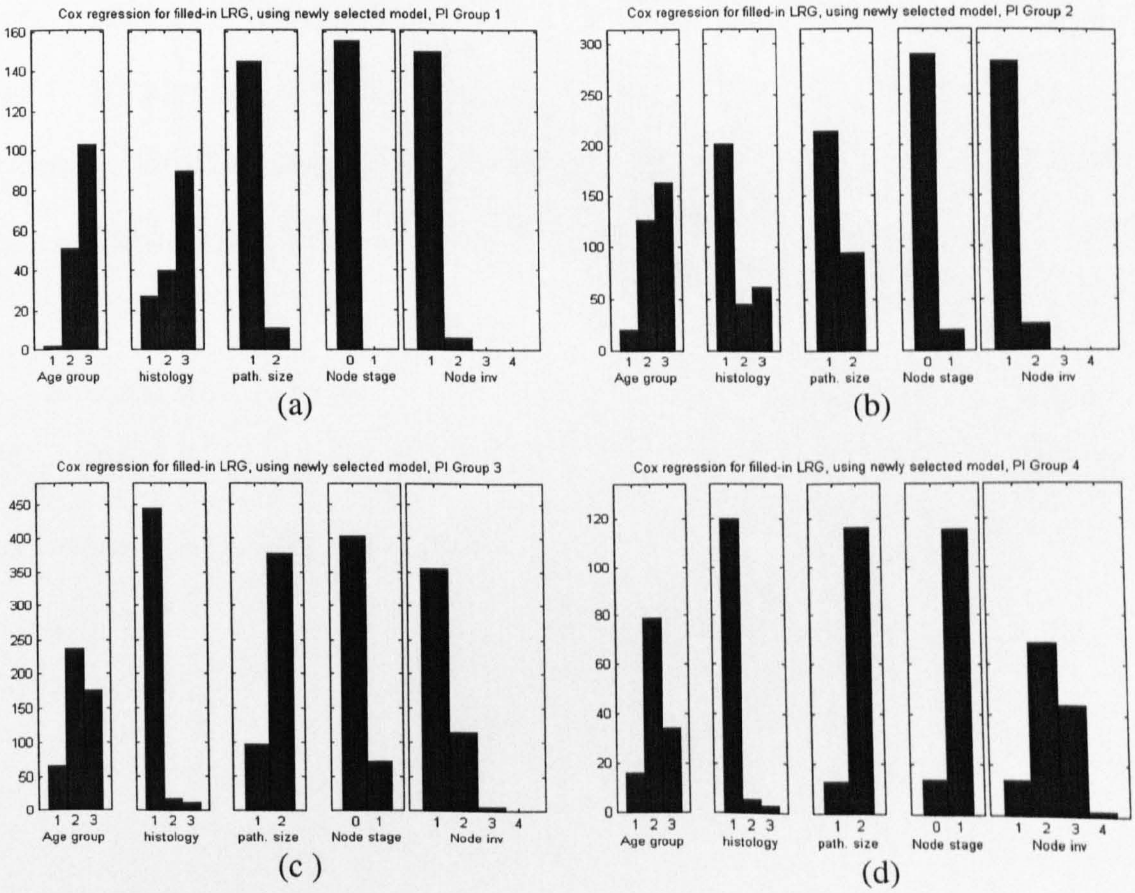


Figure (7.9): The attribute histograms of the prognostic groups using the newly selected model for the completed low-risk cohort.

7.2.2 Model selection for the completed high-risk cohort

7.2.2.1 The effect of sample size in the high-risk cohort

Figure (7.10-7.12) displays three different sets of survival curves of *clinical stage* from three different sample conditions for the high-risk cohort. Figure (7.10) discards all the cases with *pathological size* value missing are discarded, that is assuming that the missing mechanism is at random and therefore the data distribution remains unaffected after missing values are removed. Figure (7.11) includes the cases where treating the *pathological size* missing values as a separate attribute, and figure (7.12) has the missing *pathological size* missing values predicted. Clearly, the survival curves of figure (7.11) are very different from the other two. From the highest to lowest survivorship, the *clinical stage* categories is in an order of 1, 2, 3 and 4, whereas the order was shown as 2, 1, 3 and 4 in the other two figures.

It is concluded that representing missing values of *pathological size* as a separate attribute, results in the most differentiation between different prognostic risk groups, when compared to two alternative strategies.

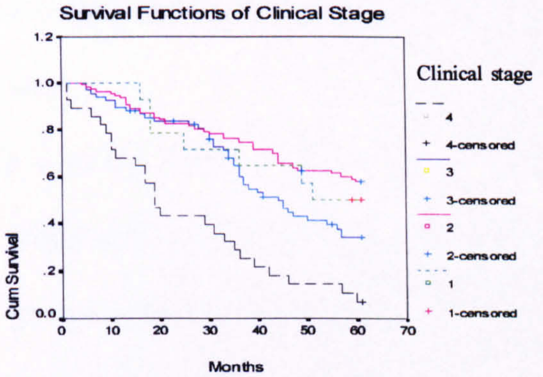


Figure (7.10): The survival function of *clinical stage* from the high-risk cohort where the *pathological size* labelled as missing, were left out, only contained 209 records.

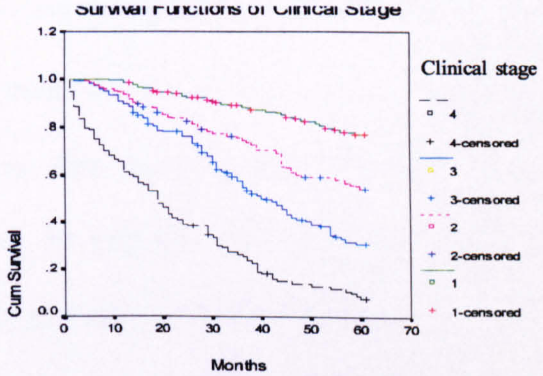


Figure (7.11): The survival function of *clinical stage* from the high-risk cohort, coding *pathological size* missing values as a separate attribute, contained 633 records.

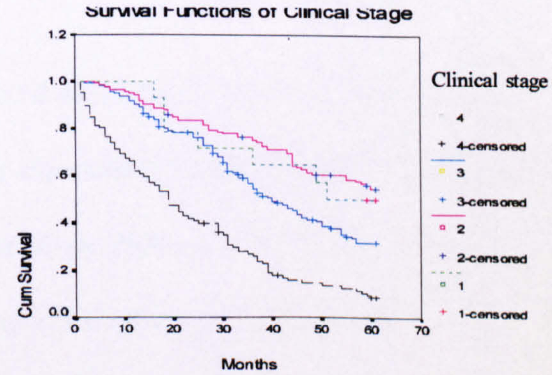


Figure (7.12): The survival function of *clinical stage* from the completed high-risk cohort where the missing *pathological size* has been predicted, contained 403 records.

7.2.2.2 Model selection for the completed high-risk cohort

The Cox selected model for the high-risk cohort estimating the missing values contained only 3 variables, namely *age group, node stage* and *clinical stage*. This was selected by forward stepwise elimination procedure and variable interactions were not considered. None of these variables contain missing data.

Since the sample size is smaller than the other data set, the analysis is completed with a 4-fold cross validation in order to have significant number of samples for training and the results are displayed in figure (7.13) - (7.15) using the newly selected model. The results are shown in the order of prognostic groups partitioning, the Cox predicted survivorship with the Kaplan-Meier estimate of the survivor functions for each of the prognostic group and the corresponding attributes histograms. Note that four different prognostic groups were identified, one more than previously. Prognostic group 1, the highest survival group, only contains 30 patients, where the Kaplan-Meier estimation shows the mean survivorship is kept constant at the value of 1 for the first 19 months after surgery, then starts to decline gradually to the value of 0.83, remaining constant from month 39 onwards. The remaining groups contains 164, 130 and 79 patients respectively and the corresponding survivorship after 5 years is 0.44 0.27 and 0.045.

The predicted survivorships for all of the groups are consistent with the Kaplan-Meier confidence intervals. The highest survival group also shows a distinctive characteristic from the attribute profiles, where the attributes are concentrated on *age group* 3, *node stage* 0 and *clinical stage* 2. Then the *age group* changes from attribute 3 to loosely spread between 1 to 3, while *clinical stage* gradually moves from 2 to a sequence of 1,3 and 4. Finally the *node stage* moves from 0 to 3 over the rest of prognostic groups. The variables *clinical stage* and *node stage* show clear differentiated attribute profiles from prognostic group 3 to 4 but the *age group* does not.

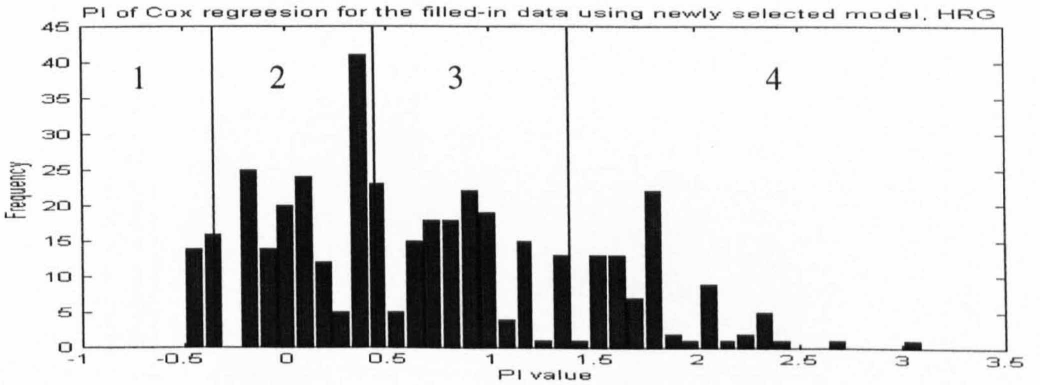
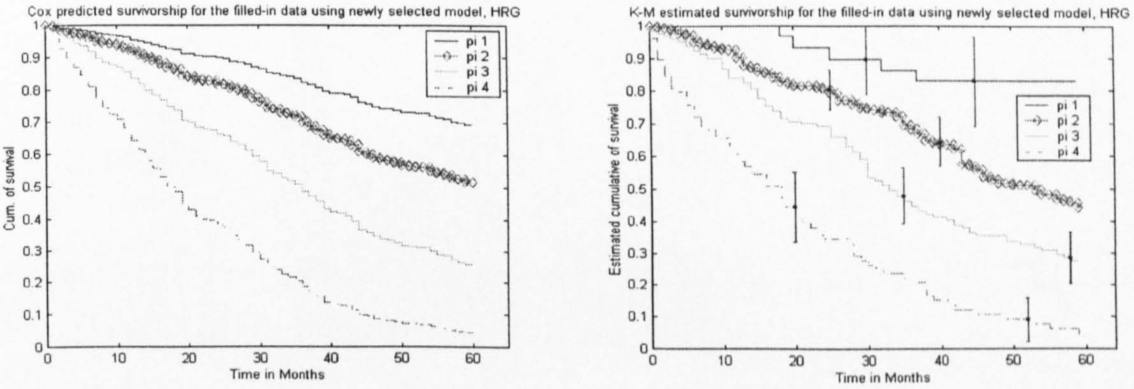


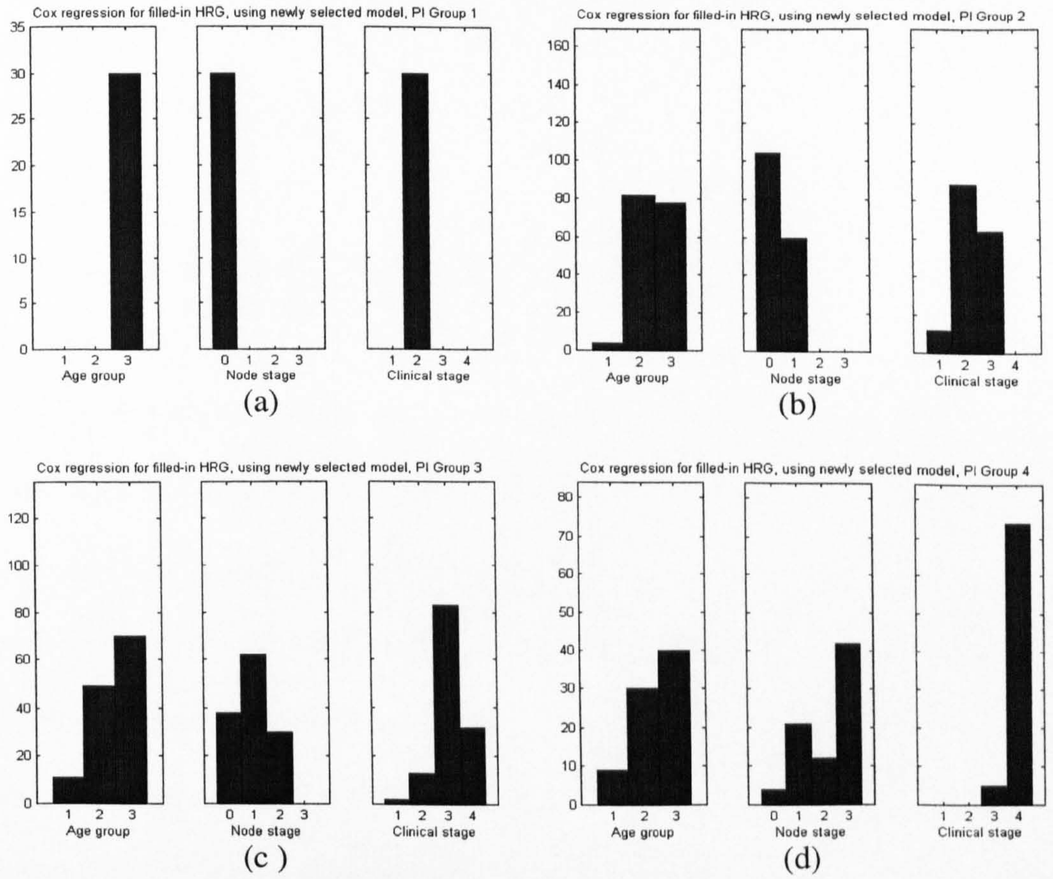
Figure (7.13): The partitioned prognostic groups by the Cox regression using the newly selected model for the completed high-risk cohort.



(a)

(b)

Figure (7.14): The Cox predicted survivorship for the prognostic groups of completed high-risk cohort using the newly selected model and the corresponding Kaplan-Meier estimate survivor function.



(a)

(b)

(c)

(d)

Figure (7.15): The attribute histograms of the prognostic groups using the newly selected model for the filled-in high-risk cohort.

7.3 Applying the Bayesian PLANN model to the original high-risk cohort using the Cox selected by filling-in the missing values for the high-risk cohort

The Bayesian PLANN model was applied to the variables selected by Cox modelling of the completed high-risk cohort, aiming to investigate the possibility of variable interactions, and allow performance comparison with Cox regression. The network was evaluated with a 4-fold cross validation to define the prognostic index boundaries for the mortality risk groups. A total of 403 records were considered. The 3 variables selected by the Cox regression from the completed high-risk cohort were transformed into 9 binary input variables in the usual way, which together with the time variable formed the input layer, 18 hidden nodes were used and the single output node represented the conditional probability hazards in particular time intervals. Baseline attributes and grouped ARD technique were also used.

The network was then applied to the 633 records of the original high-risk cohort which contained missing *pathological size*. Figure (7.16) displays the distribution of prognostic indexes from these 633 records and the pre-defined positions where aggregate mortality risk groups. The network predicted survivorship for each mortality risk group together with the corresponding Kaplan-Meier estimate of the survivor functions are displayed in figure (7.17), and the attribute profiles of prognostic groups are shown in figure (7.18).

The highest survival group, prognostic group 1, contains 31 patients and appears to be a specific group of *age group* 3, *nodes stage* 0 and *clinical stage* 2 which is identical to the Cox partitioned prognostic group 1 in figure (7.14a). Prognostic groups 2,3 and 4 contain 384, 142 and 76 patients, respectively, which are different from the Cox partition. However, there is no significant survivorship and attribute profiling difference from the results with Cox regression, suggesting that variable interactions between the 3 selected variables are not significant.

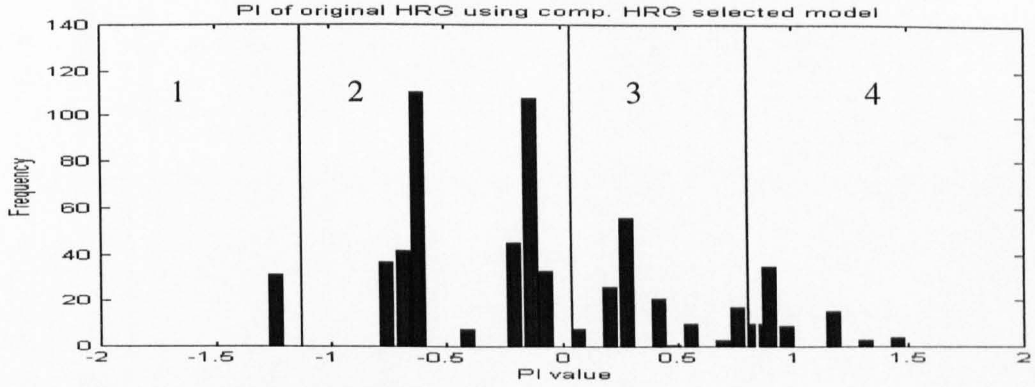
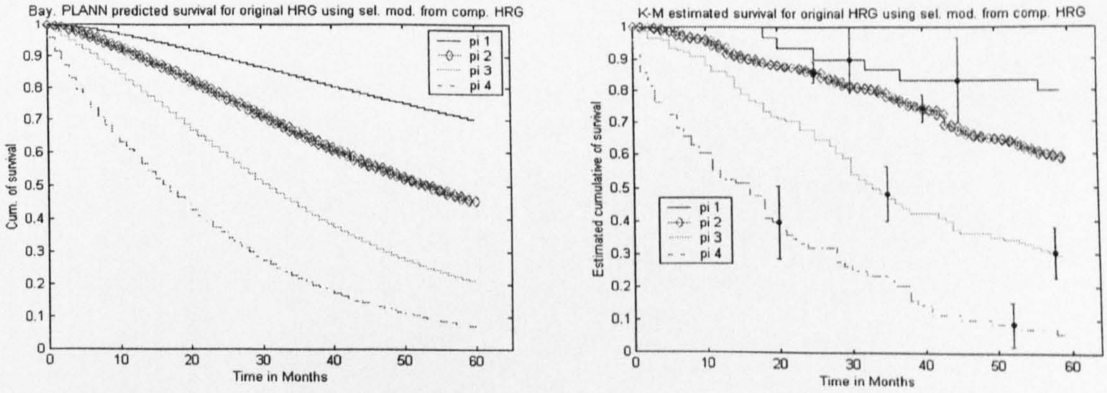
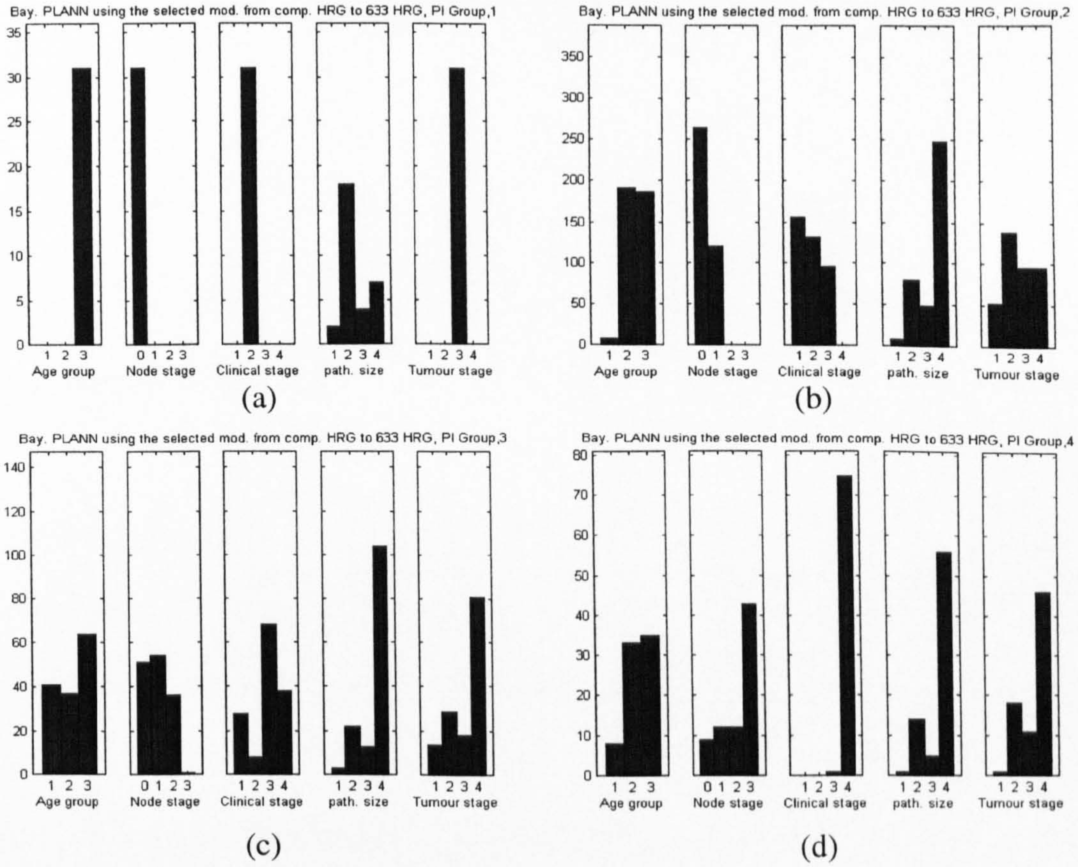


Figure (7.16): The calculated prognostic indexes for the original high-risk cohort which were gathered from network trained by the completed high-risk cohort using the Cox selected model from it.



(a) (b)

Figure (7.17): (a): The network predicted survivorship for the prognostic groups in figure (7.16) and (b) the corresponding Kaplan-Meier estimate of the survivor functions.



(a)

(b)

(c)

(d)

Figure (7.18): The attribute histograms for the prognostic groups in figure (7.19), *pathological size* and *tumour stage* are also displayed to monitor the distribution of the patients with *pathological size* missing.

7.4 Discussion the adaptability of the regression method for missing data prediction

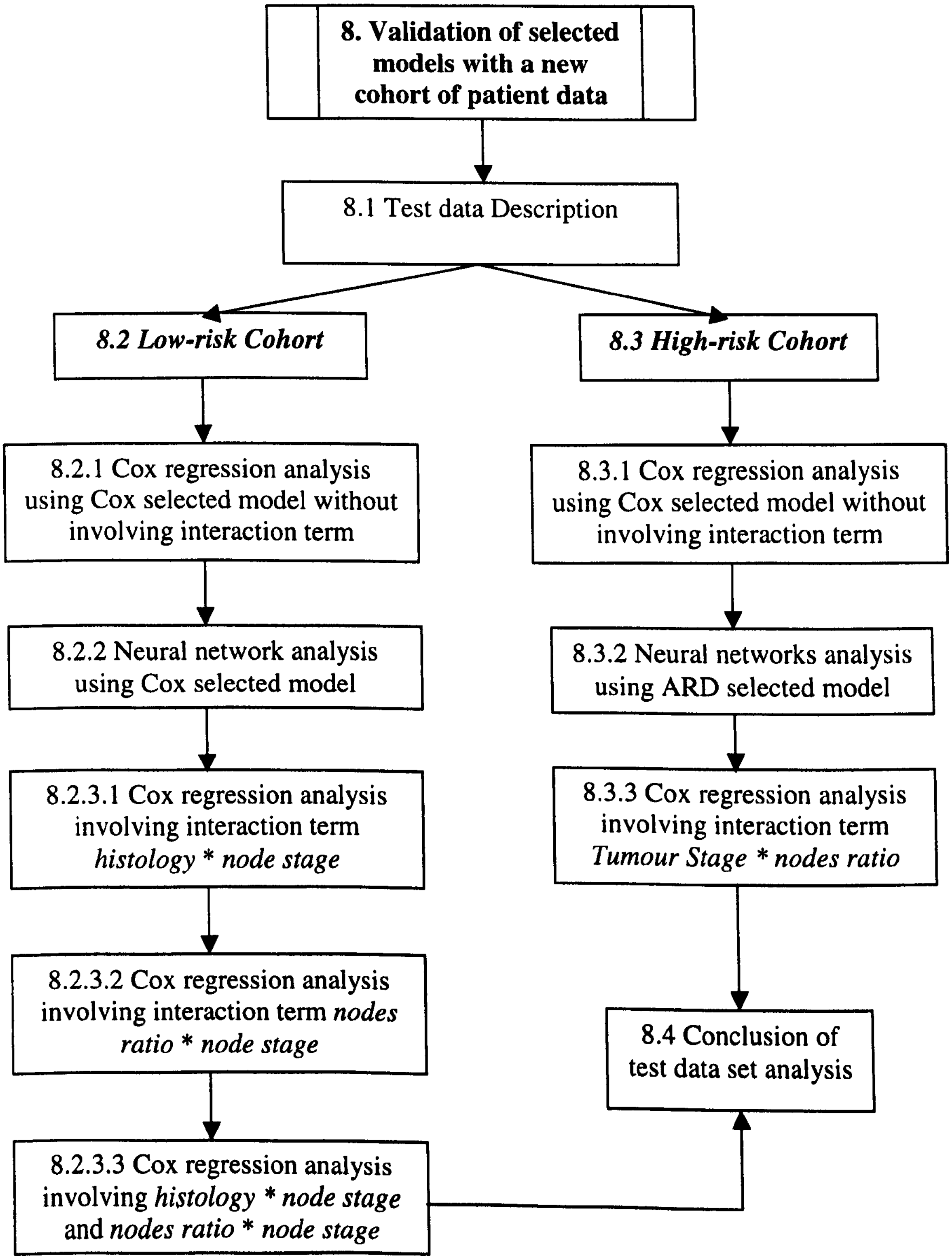
In terms of the improvement to the survival prediction accuracy after filling-in the missing data, the survival predictions have been improved slightly but not significantly.

For the low-risk cohort, the models selected before and after filling-in the missing data are very much the same, except that *age group* was selected additionally and *nodes ratio* was replaced by *number of nodes involved*, considering that the samples size of the two data sets are different. There is one more prognostic group is partitioned from the filled-in low-risk cohort analysis using the model selected from it, when comparing with the results obtain from the original data. The 5 years survivorship of the lowest survival group reduced from 0.52 to 0.4.

The filled-in high-risk cohort contained much less samples, less than 2/3 of the original high-risk cohort. Figure (7.10-7.12) displays the different behaviour of *clinical stage* at the change of sample conditions, excluding the cases with the values of pathological size missing, including these cases as a separate attribute, and having them predicted. The model selected from the completed high-risk cohort only contained 3 variables. Two of them were already in the original model, but *age group* was added to the model. Again, an additional prognostic group was partitioned when using the model selected from the filled-in data and this group contained very few patients with very high survivorship. This group of patients was apparent again when this model was tested by the original data set. As a result, the 5-year survivorship of some of the other groups was raised, which is the result of the existence of the records with *pathological size* missing in the data. These 3 variables are capable of identifying a very high survival group in the high-risk cohort, which the original model could not. Nevertheless, the Bayesian PLANN model analysis confirmed that there is no indication of significant variable interactions.

It is concluded that filling-in the missing values in the data results is a more detailed breakdown of the prognostic risk groups than was possible from the original data set. This was due, in part, to the change in the allocation of records between the low- and high-risk cohort.

Chapter 8


8. Validation of selected models with a new cohort of patient data
8.1 Test data Description
8.2 Low-risk Cohort
8.3 High-risk Cohort
8.2.1 Cox regression analysis using Cox selected model without involving interaction term
8.3.1 Cox regression analysis using Cox selected model without involving interaction term
8.2.2 Neural network analysis using Cox selected model
8.3.2 Neural networks analysis using ARD selected model
8.2.3.1 Cox regression analysis involving interaction term *histology * node stage*
8.3.3 Cox regression analysis involving interaction term *Tumour Stage * nodes ratio*
8.2.3.2 Cox regression analysis involving interaction term *nodes ratio * node stage*
8.4 Conclusion of test data set analysis
8.2.3.3 Cox regression analysis involving *histology * node stage* and *nodes ratio * node stage*

8. Evaluate modelling methods using a prospective test data set

Cox regression and neural networks have been extensively applied in many medical applications. In particular, Cox regression has used for more than 20 years in medical survival analysis. Previous chapters have demonstrated the use and strength of each method.

To investigate the robustness of each method further, the models fitted to a patient cohort recruited during 1983-89 were applied to a second cohort recruited between 1990 - 93. These data acted as a validation set to evaluate the predictive value of the prognostic indexes derived by the Cox regression and the neural network model. For each method, the network weights and the cut-off points for prognostic group partitioning follow previously defined for the first patient cohort.

8.1 Description of the validation data set

The validation set comprises records from1653 new patients. Within these records, 388 were discarded due to missing data, leaving 1265 cases for model validation. The data were divided into low- and high-risk cohorts, following the same separation criteria as used for the design data, resulting in 931 and 334 cases in each group, respectively.

The population distribution of the two data sets is slightly different, as the validation data contains a higher proportion of low risk patients. Originally, there were 59% and 41% of patients allocated to low-risk and high-risk cohort from the entire data set, but for the second cohort these figures become 73.6% and 26.4%, respectively. Moreover, this characteristic also reflects in each of the variables as illustrated in appendix (II). For each of the variables, more patients are under the low-risk attributes than the first data set, the design data.

Missing data were still a feature of the validation data set. Some of the variables even contain a higher portion of missing data than the design data, such as *predominant site* and *histology* in the high-risk cohort. However, the number of records contained missing data in *nodes ratio* has been reduced for both of low- and high-risk cohorts.

8.2 Test data set low-risk cohort analysis

8.2.1 Validating the Cox regression modelling method using low-risk cohort of validation set

Previously, the low-risk cohort was implemented with 5-fold cross validation. All subjects in the low-risk cohort of validation set acts as a test set for each cross validation set of network weights, then five sets of results are collected and averaged as the final result for the low-risk cohort of validation set. The model fitted to the Cox regression is the Cox selected model for the low-risk cohort of design data, no interaction involved. The results for this cohort are displayed in figure (8.1) - (8.3). The Cox calculated prognostic indexes are shown in figure (8.1). The predicted mean survivorship for the prognostic groups is displayed in figure (8.2), together with the corresponding Kaplan-Meier estimate survival functions. The attribute histograms for each prognostic group are displayed in figure (8.3).

The Cox regression produces similar kind of survival prediction and attribute profiles as for the low-risk cohort of design data. However, the Kaplan-Meier estimates confirm that the survivorship of the low-risk cohort of the validation set is better than the Cox estimated for each prognostic group. The Cox estimation for each group contains around 0.1 error over 60 months on average. This is discussed further at the end of this chapter.

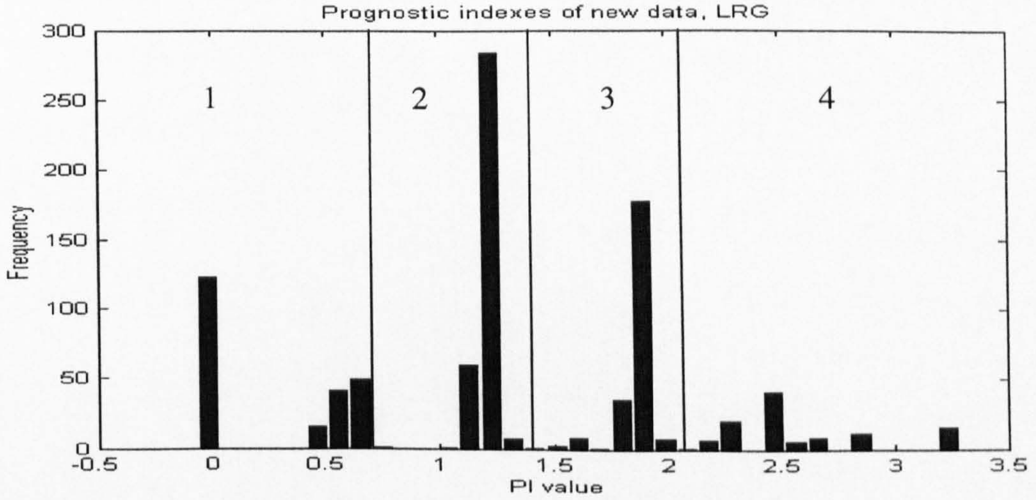


Figure (8.1): Cox regression calculated prognostic indexes for the low-risk cohort of validation set using the Cox regression selected variables for the original low-risk cohort, without involving interaction term.

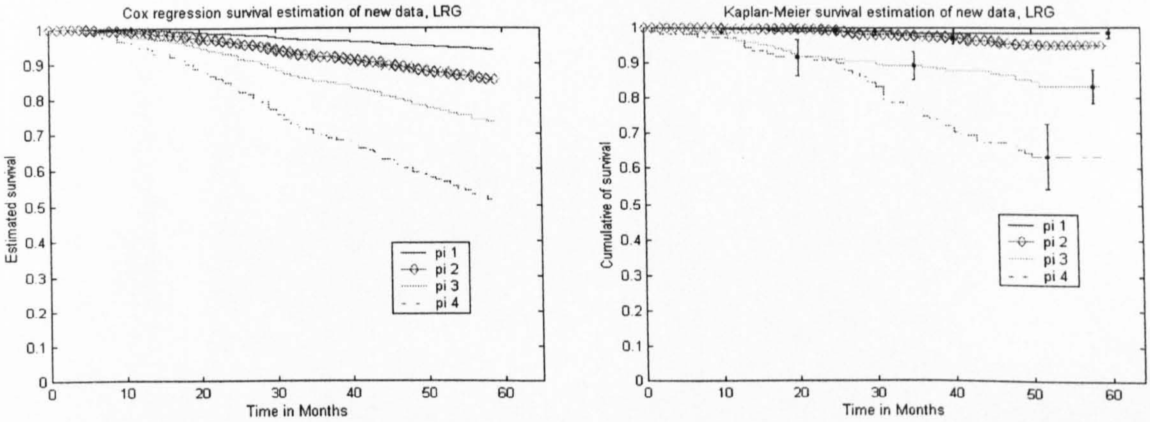


Figure (8.2): The Cox regression predicted mean survivorship for prognostic groups, and the corresponding Kaplan-Meier estimate survivor functions. The Kaplan-Meier estimates confirm that the survivorship of the low-risk cohort of the validation set is better than Cox estimated.

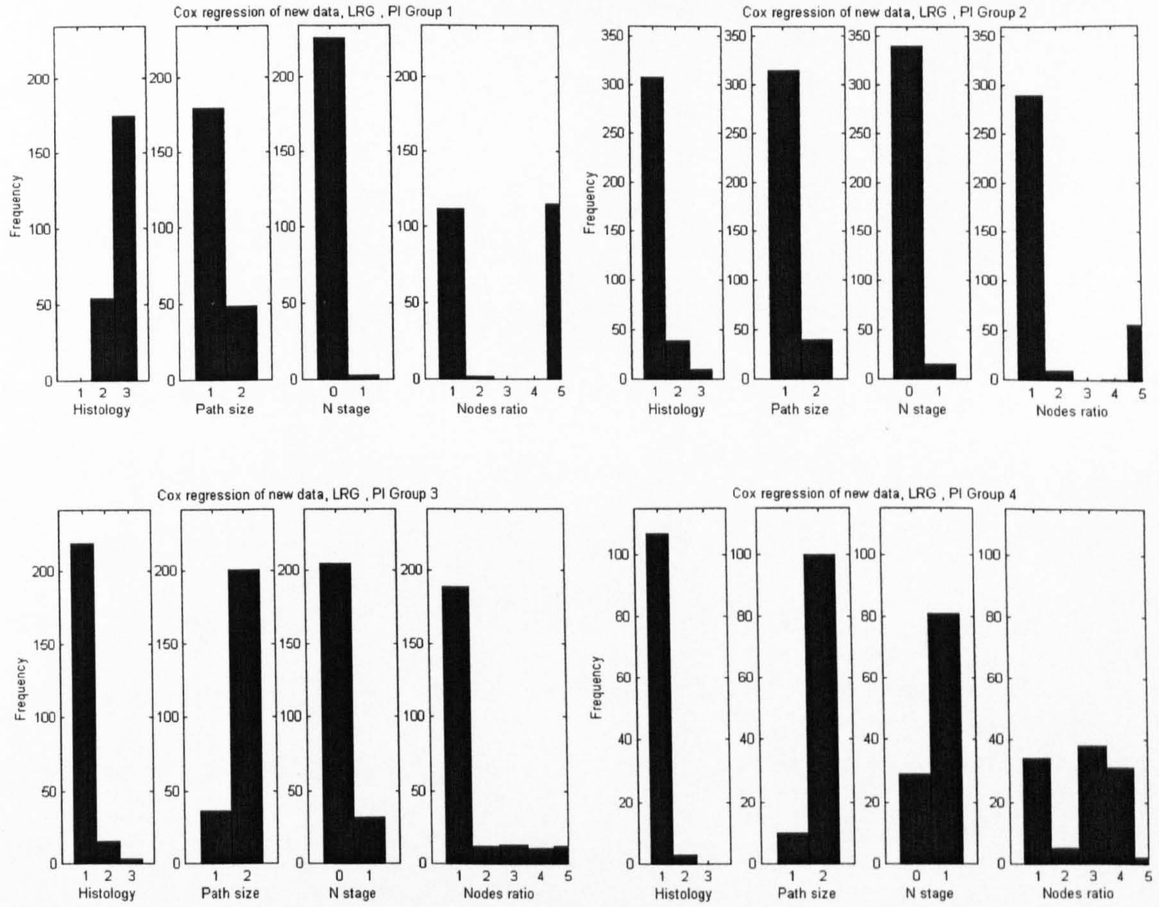


Figure (8.3): The attribute profiles for prognostic groups of the low-risk cohort of the validation set using the Cox regression. The results show no distinguishable difference from the results for the low-risk cohort of design data with Cox regression.

8.2.2 Validating the neural networks modelling method using low-risk cohort of validation set

The results for the low-risk cohort of the validation set were generated in the same way as with the Cox regression, by averaging predictions from 5 cross-validation networks. The network outputs were still marginalised toward the averaged hazard of the low-risk cohort of design data and the same prognostic index intervals were used for prognostic risk groups. The results are presented in figure (8.4) - (8.6), showing the partitioned prognostic groups using the intervals defined for the low-risk cohort of design data, the mean survivorship predicted by the network for each prognostic group and the corresponding Kaplan-Meier estimates of the survivorship functions, finally, the attribute histograms.

The network predicted survivorship for the different prognostic groups is similar to the result for the low-risk cohort of design data using same approach in figure (5.19), and the corresponding observed survivorship has showed a better result except group 4, of which the observed survival rate is poorer than the predicted. Moreover, the attribute profiles show no distinguishable difference from previous result, figure (8.6).

Results for the test data set can also be obtained by modelling the entire design data and tests by the validation set. Results from both approaches show no significant differences as illustrated in figure (8.7). The first approach does not require retraining with the complete data set, which is expensive computationally and maintains the consistency and fairness for result comparison. Therefore, all the results generated for the validation data set is completed by the first approach.

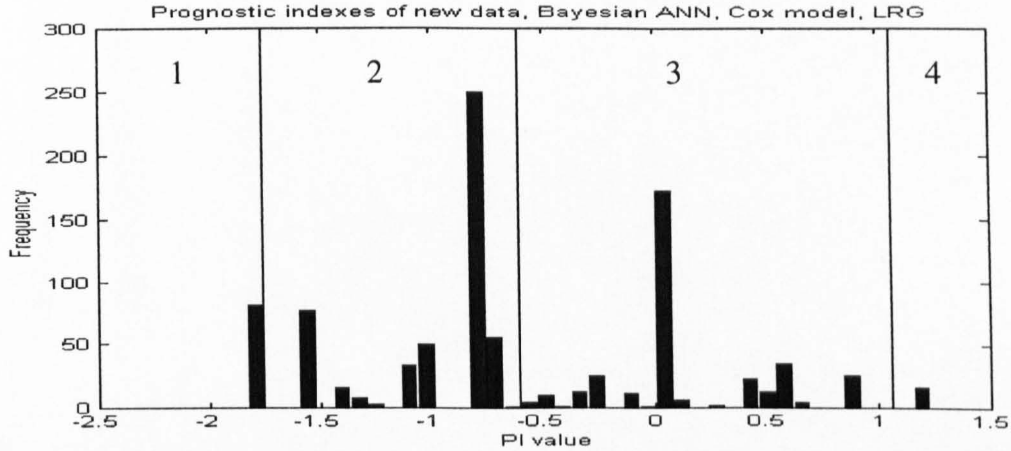


Figure (8.4): The neural network calculated prognostic indexes for the low-risk cohort of validation set.

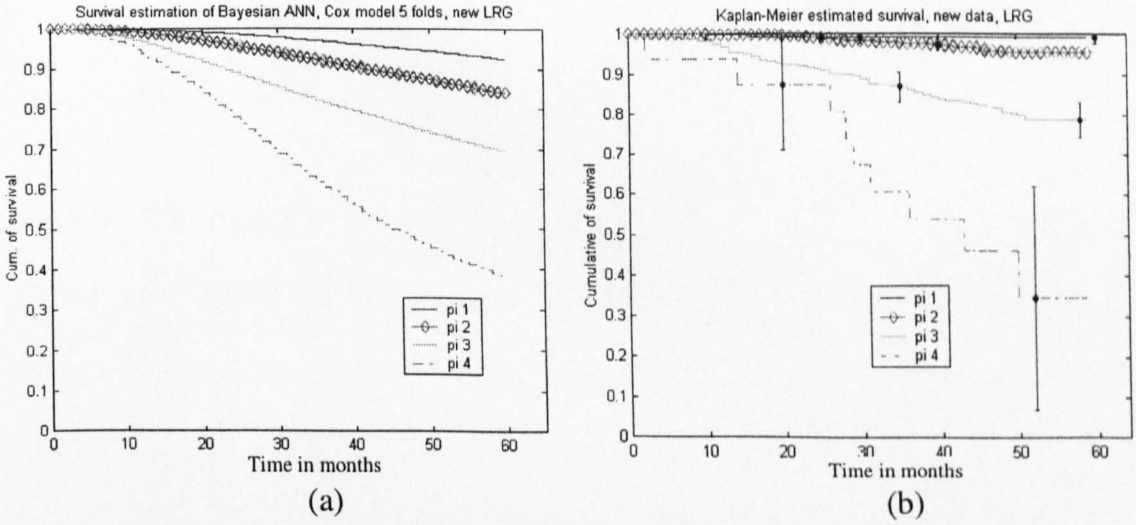
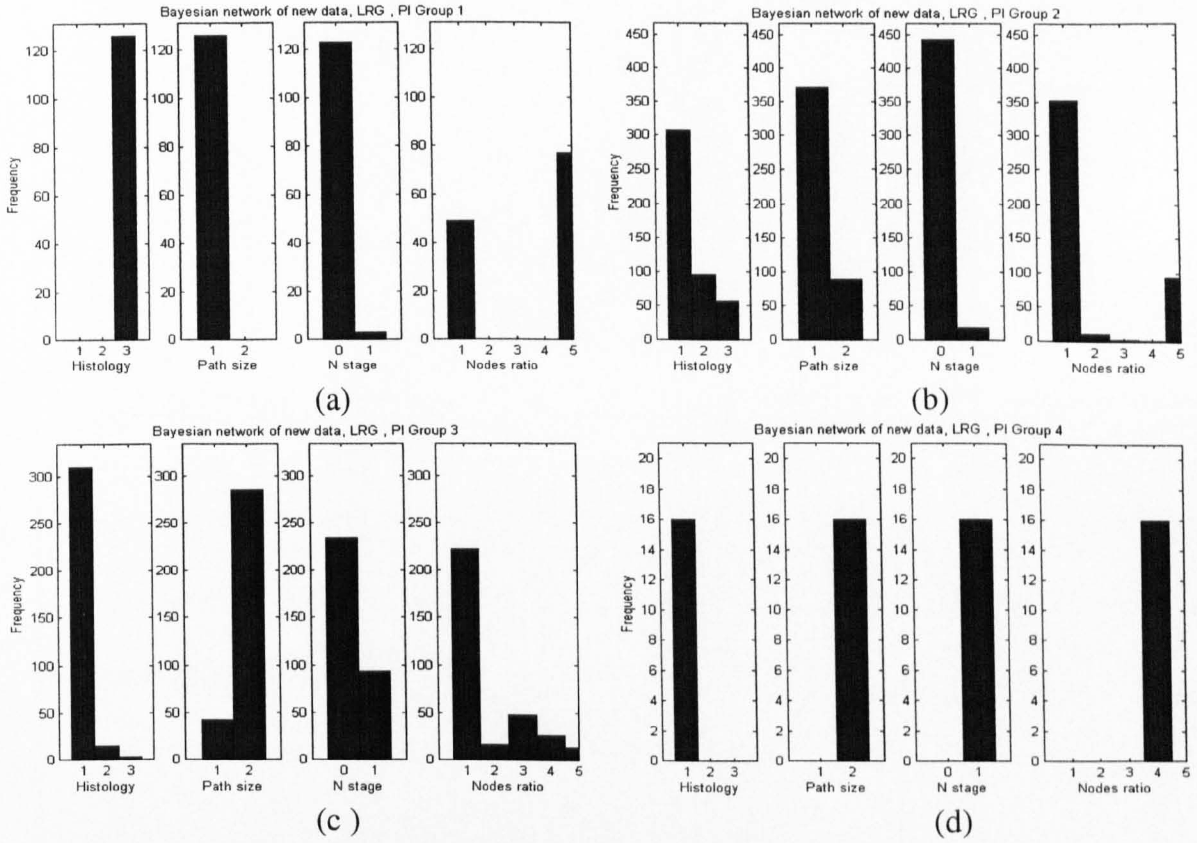


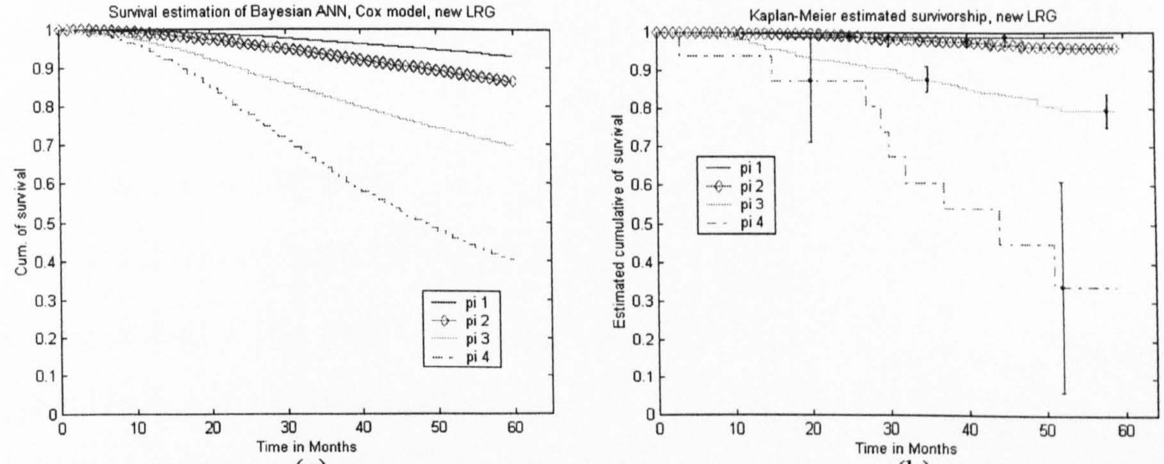
Figure (8.5): (a) The neural network predicted survivorship the pre-defined prognostic groups and (b) the corresponding Kaplan-Meier estimate of the survivor functions.



(a) (b)

(c) (d)

Figure(8.6): The variable histograms for prognostic groups of low-risk cohort of validation set using the neural network. The results show no distinguishable difference from the results for low-risk cohort of design data on the same approach.



(a) (b)

Figure (8.7): The network was trained with the design data and tested by the test set. (a): The network predicted survivorship for the pre-defined prognostic groups and (b): the corresponding Kaplan-Meier estimate of the survivor functions, resulting that no difference was made from the combined results for 5 cross validation sets, figure (8.5).

8.2.3 Validating the Cox regression method using the low-risk cohort of validation set involving interaction term

The following is the analysis of Cox regression for the low-risk cohort of validation set involving different interaction terms. Details of the interaction terms and their effect on the survivorship of each group are listed in table (8.1). The predicted survivorship is similar with the results of the design data.

Interaction term	**Effect**
*Histology * node stage*	The highest survival group in the low-risk cohort has a survival rate > 0.95.
*nodes ratio * node stage*	The lowest survival group in the low-risk cohort has survival rate < 0.3.
*Histology * node stage*, together with *nodes ratio * node stage*	Lost the capability to accurately identify the lowest and the highest survival group.

Table (8.1): Identified interaction terms for the low-risk cohort of design data and their effect on the group survivorship.

8.2.3.1 Inclusion of a pairwise interaction involving *histology* and *node stage*

The low-risk cohort of validation set is fitted with a model consisting of *pathological size, histology, nodes ratio* and *histology * node stage*, shown for the design data in section (6.5.1) and Cox results for the validation data are summarised in figures (8.8) - (8.10), showing that the specification of two groups f population is lost, group 1 and group 4. They are corresponding to the highest and the lowest survival group in the low-risk cohort of design data.

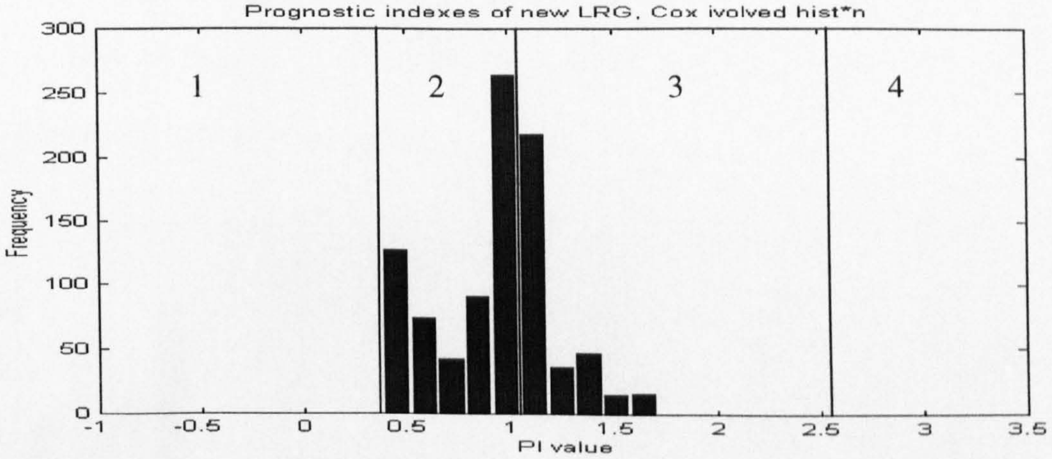
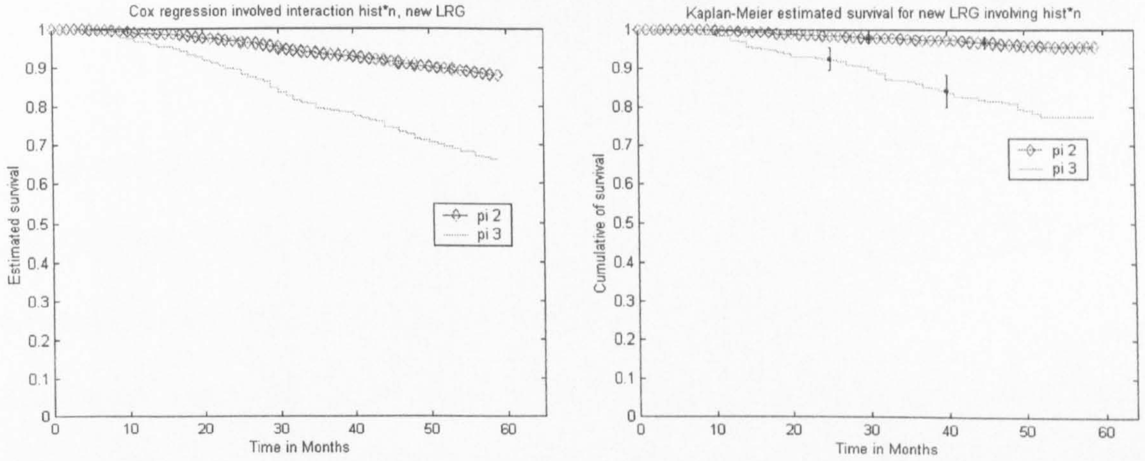
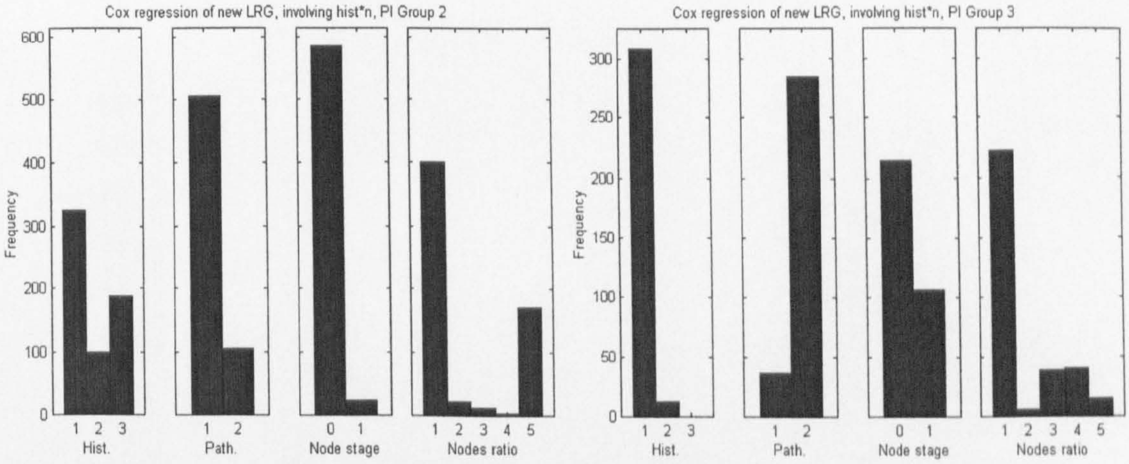


Figure (8.8): The Cox regression calculated prognostic groups involving interaction term, *histology* and *node stage,* only two prognostic groups are recorded.



(a) (b)

Figure (8.9): The Cox regression predicted survivorship for the 2 prognostic groups involving interaction term, *histology* and *node stage,* and the corresponding Kaplan-Meier estimate of the survivor functions. The observed survivorship for both groups is better than predicted by Cox regression.



(b) (c)

Figure (8.10): The attribute histograms for the prognostic groups of Cox regression involving interaction term. The results show no distinguishable difference from the low-risk cohort of design data result involving interaction term, *histology* and *node stage,* on the same approach.

8.2.3.2 Inclusion of a pairwise interaction involving *nodes ratio* and *node stage*

A model comprising *node stage, pathological size, histology* and *nodes ratio * node stage* identifies the lowest survival group in the low-risk cohort, of which the survivorship is below 0.3 as shown in the Cox's prediction for the low-risk cohort of the validation set, figure (8.12). This suggests that the interaction between *nodes ratio * node stage* is important to the group with very low survival. The Cox prediction and attribute profiles show no different from previous results using same model as in section (6.5.2), except that the observed survivorship for each patient group has improved compared to the model predictions in figure (8.12). Clearly, group 1 and 2 are brought closer together towards a probability of 1.

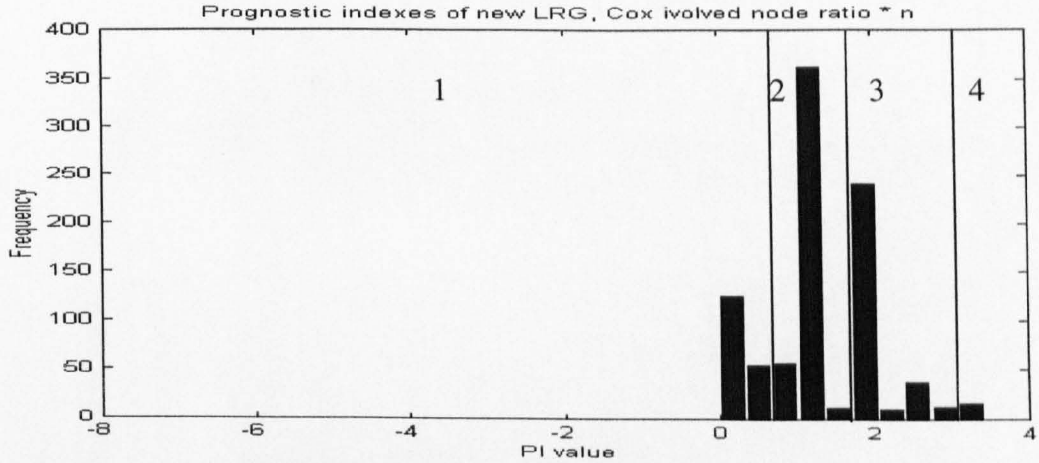
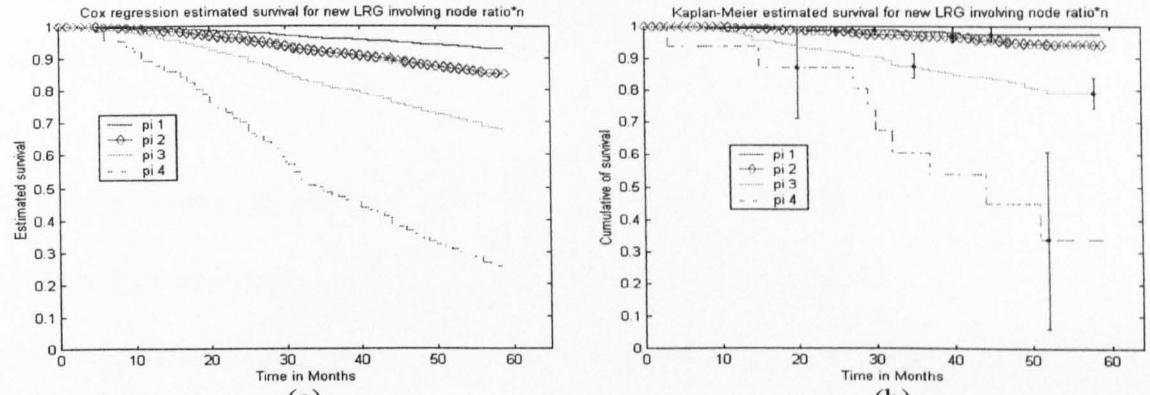
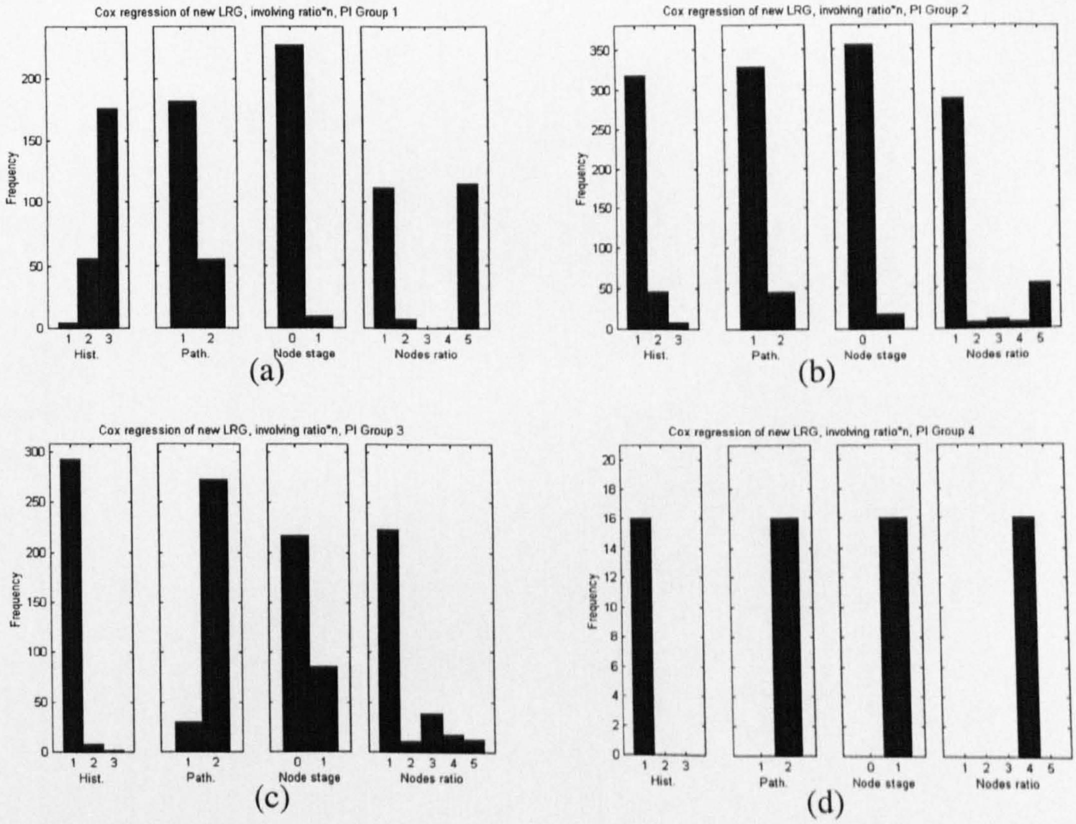


Figure (8.11): The Cox regression partitioned prognostic groups involving interaction term *nodes ratio* and *node stage.*



(a) (b)

Figure (8.12): The Cox regression predicted survivorship involving interaction term *nodes ratio* and *node stage* for the prognostic groups of low-risk cohort of validation set and their corresponding Kaplan-Meier estimate of the survivor functions.



(a) (b) (c) (d)

Figure (8.13): The attribute histograms for the prognostic groups of Cox regression involving interaction term nodes ratio and node stage. The results show no distinguishable difference from the previous result on the same approach.

8.2.3.3 Inclusion of interaction terms *histology * node stage* together with *nodes ratio * node stage*

The model containing two interaction terms comparises *histology, pathological size, histology * node stage* and *nodes ratio * node stage*. Previous results have shown the individual characteristics of these two interaction terms cannot be merged by bringing them together into a single model, section (6.5.3). This applies also to the new data set, as the results show no significant difference from previous results in section (6.5.3), including the attribute profiles. The observed survivorship for each patient group is again higher than for the design data.

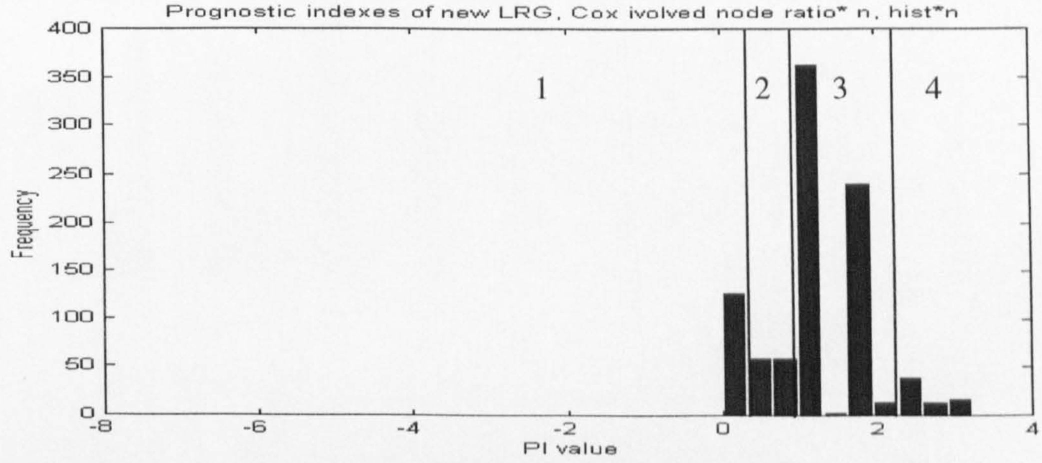
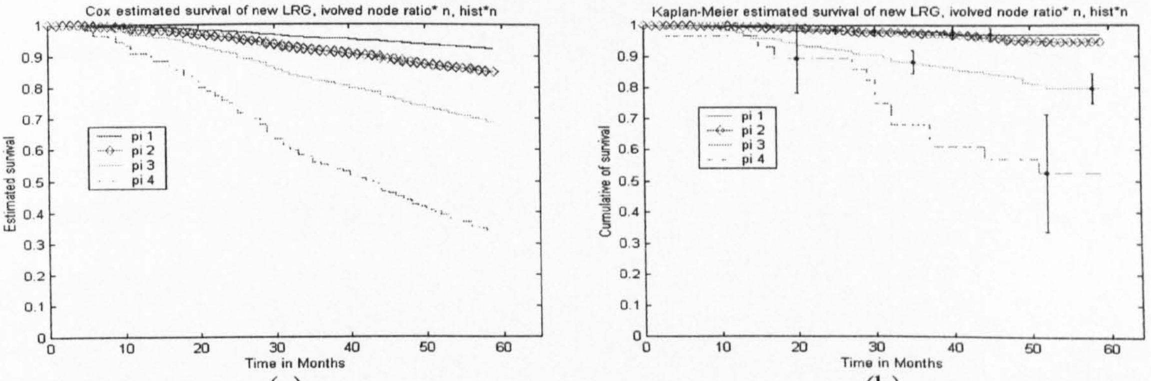
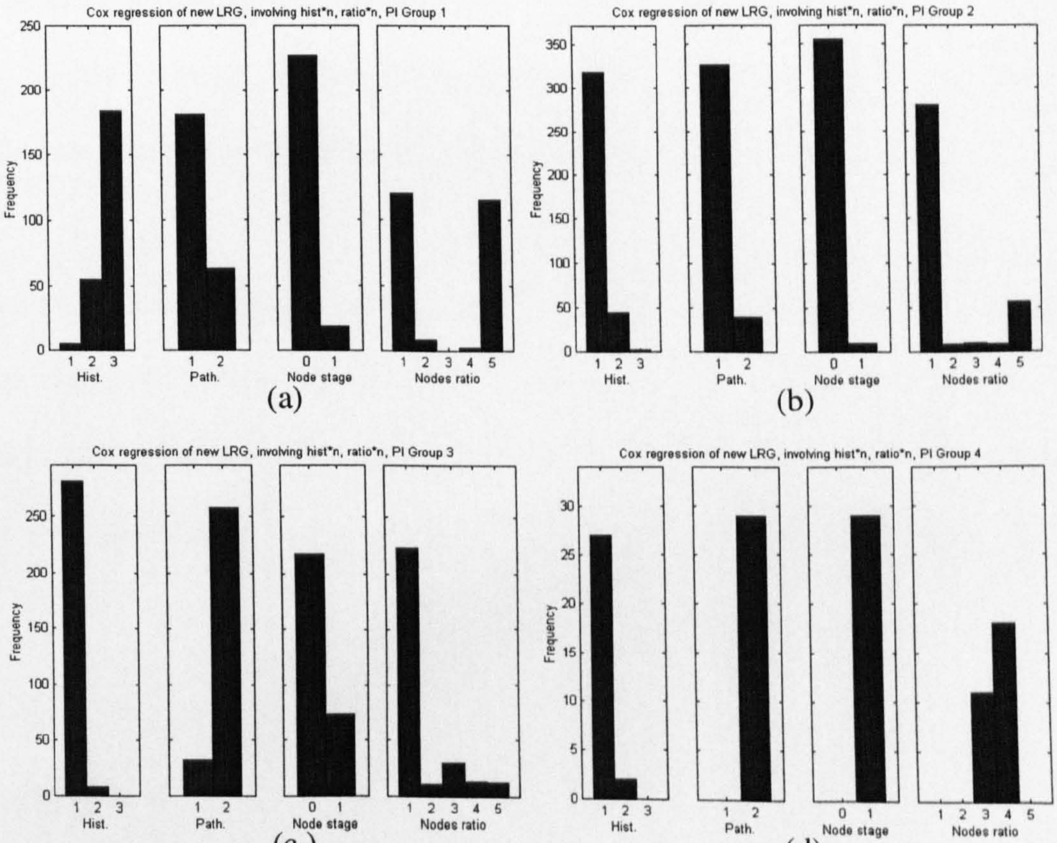


Figure (8.14): The Cox regression partitioned prognostic groups involving interaction terms *histology* and *node stage*, *nodes ratio* and *node stage*



(a) (b)

Figure (8.15): The Cox regression predicted survivorship for the prognostic groups involving interaction terms *histology* and *node stage*, *nodes ratio* and *node stage* and the corresponding Kaplan-Meier estimate of the survivor functions. The observed survivorship of the four prognostic groups is better than estimated.



(a) (b)

(c) (d)

Figure (8.16): The attribute histograms for the prognostic groups of Cox regression involving interaction terms *histology* and *node stage*, *nodes ratio* and *node stage*. The results show no distinguishable difference from the previous result on the same approach.

8.3 validation data set high-risk cohort analysis

8.3.1 Validating the Cox regression modelling method using the high-risk cohort of validation data set

The Cox regression results for the high-risk cohort of validation set were averaged over the 3 cross-validation sets with a model *comprising menopausal status, predominant site, tumor stage, node stage, histology* and n*odes ratio*. The results are presented as the follows: figure (8.17) illustrates the distribution of prognostic indexes, the Cox regression predicted survivorship for each prognostic groups are displayed in figure (8.18) together with the corresponding Kaplan-Meier estimates survival function, and the attribute histograms for prognostic groups are displayed in figure (8.19).

The predicted survival rates remain consistent with those expected from the design data, but the highest risk group now shows an increase in 5-year survival to around 0.3. This indicates that there may have been a significant improvement in the effectiveness of care for this patient group.

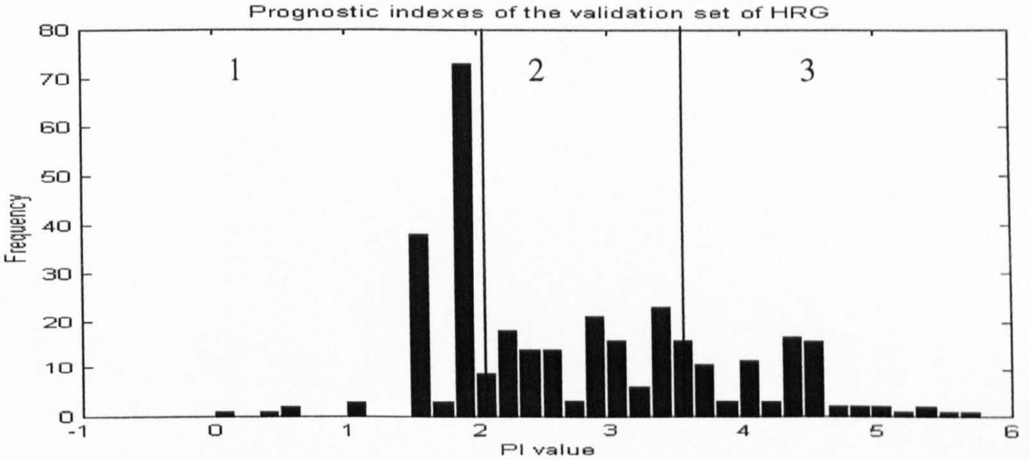
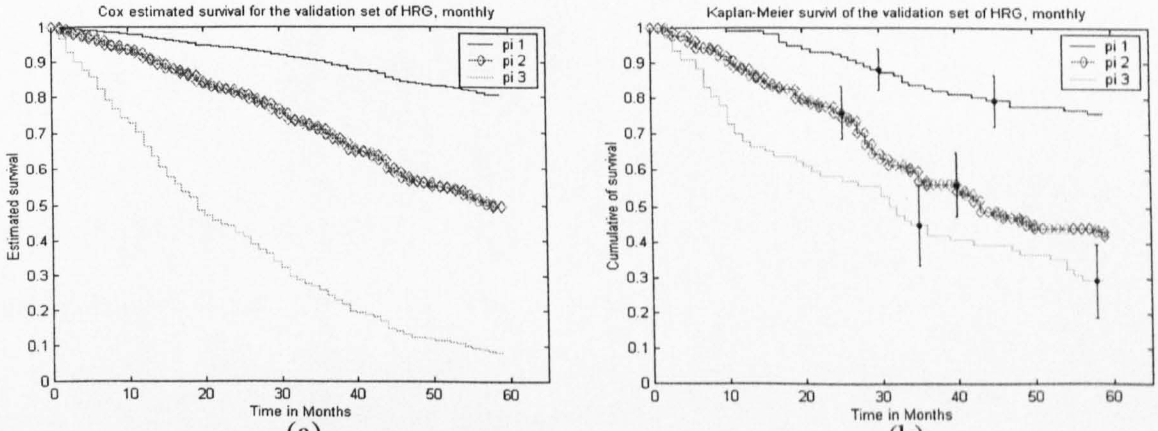
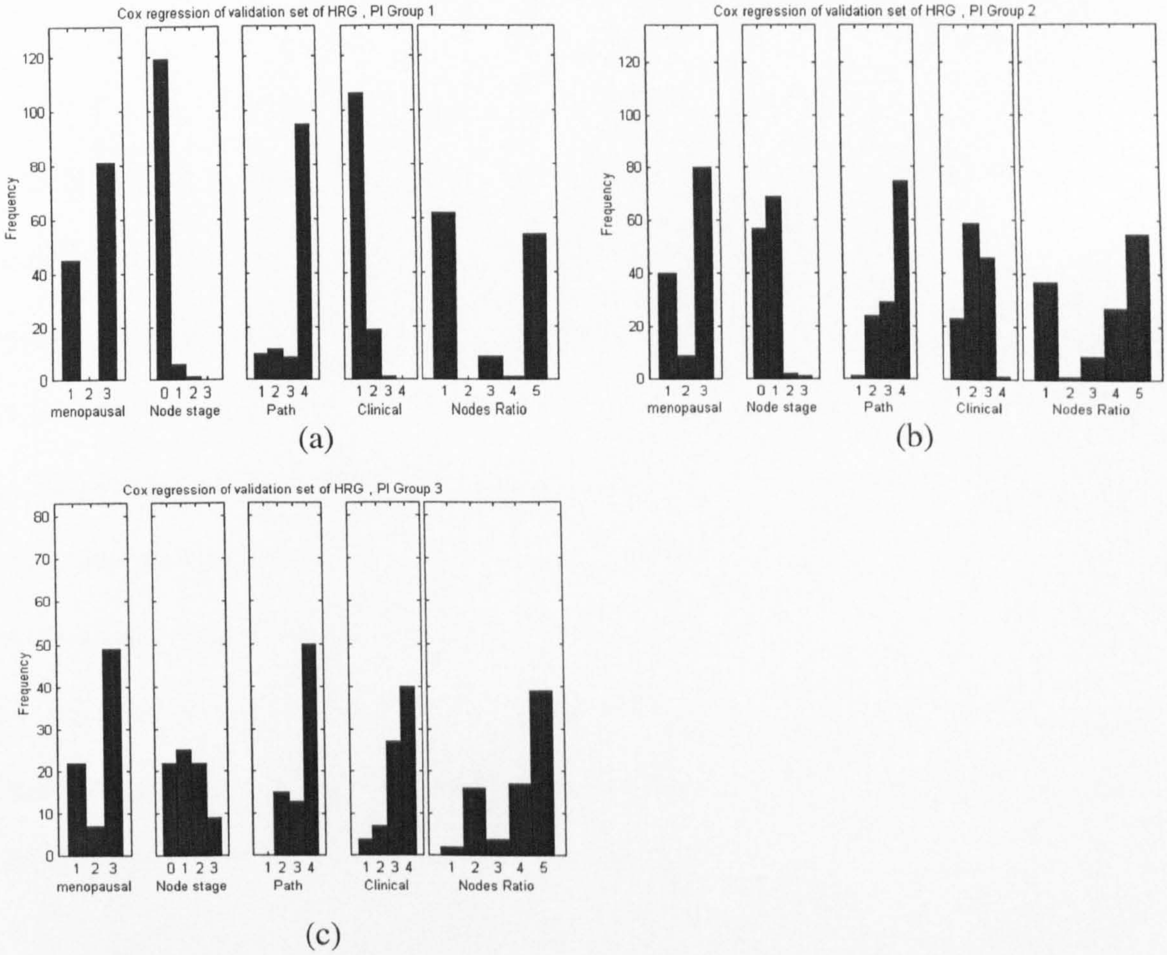


Figure (8.17): The Cox regression calculated prognostic indexes and divided prognostic groups for the high-risk cohort of validation set.



(a) (b)

Figure (8.18): The Cox regression predicted survivorship for prognostic groups and the corresponding Kaplan-Meier estimate of the survivor functions. The survival prediction for prognostic group 3 contains 0.2 error.



(a)

(b)

(c)

Figure (8.19): The attribute histograms for the prognostic groups. The results show no distinguishable difference from the results on the same approach.

8.3.2 Validating the neural network model with the high-risk cohort of validation set

A similar study to that carried out for Cox regression, was performed also with the PLANN model. Figure (8.20) shows distribution of the prognostic indexes and its partition into different groups using the same intervals as are shown in figure (6.1). Figure (8.21) illustrates the network predicted survivorship for prognostic groups and the Kaplan-Meier estimate of the survivor functions. Finally, the attribute profiles for the prognostic groups are displayed in figure (8.22). The results with the ARD selected model are presented in figure (8.23) - figure (8.25) in the same order. Both sets of results show similar survival predictions as for the high-risk cohort in the design data and no distinguishable difference is observed in the attribute profiles. As for Cox regression, there is a noticeable improvement in survival for the group at highest mortality risk.

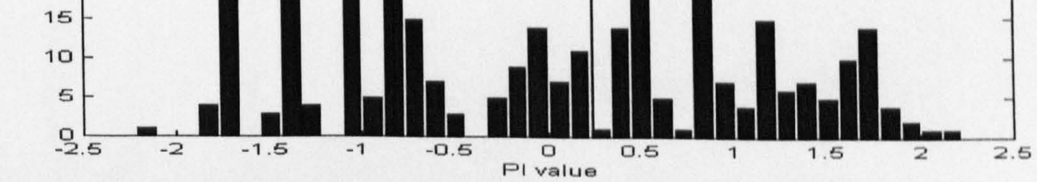


Figure (8.20): The neural networks calculated prognostic indexes and partitioned prognostic groups using the Cox selected model for the original high-risk cohort

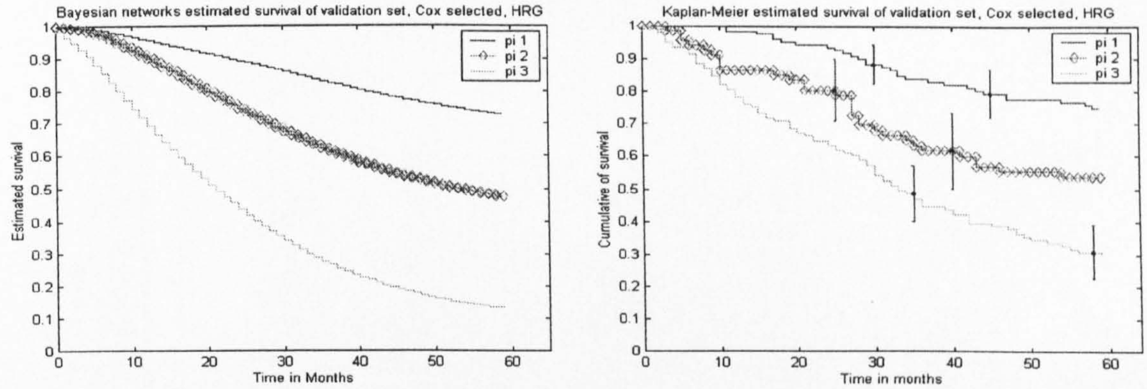


Figure (8.21): The neural networks predicted survivorship for the prognostic groups using the Cox selected model and the corresponding Kaplan-Meier estimate of the survivor functions. Only the performance of prognostic group 3 is not met the expectation as the other groups.

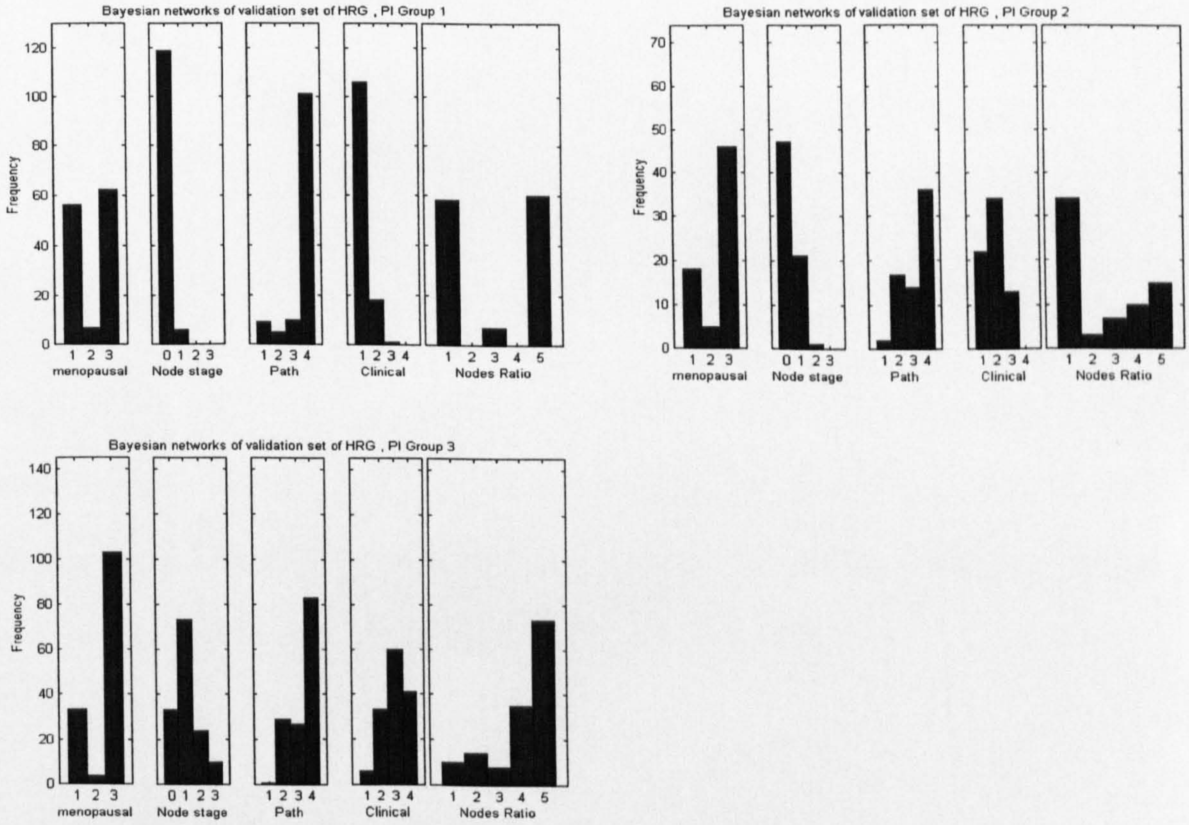


Figure (8.22): The attribute histograms for the prognostic groups using Cox selected model. The results show no distinguishable difference from the results for high-risk cohort of design data on the same approach.

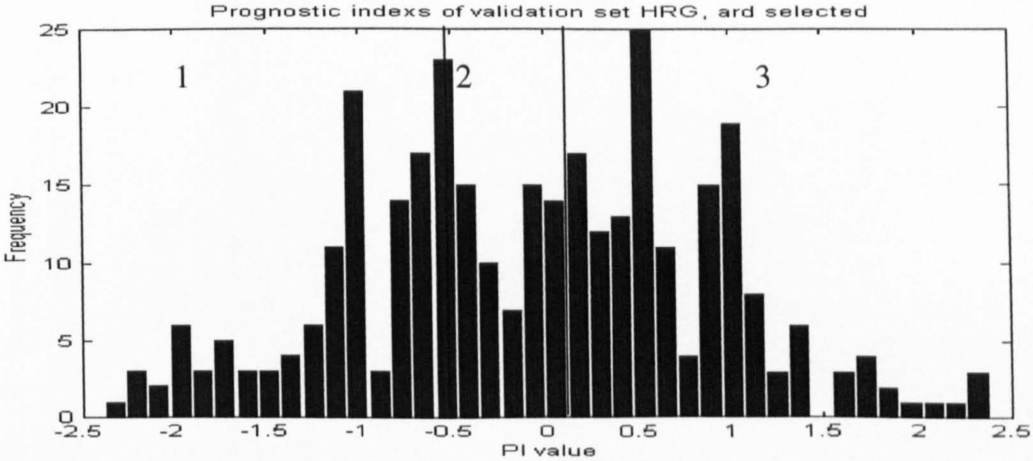


Figure (8.23): The neural networks calculated prognostic indexes and prognostic groups using the ARD selected model for the original high-risk cohort.

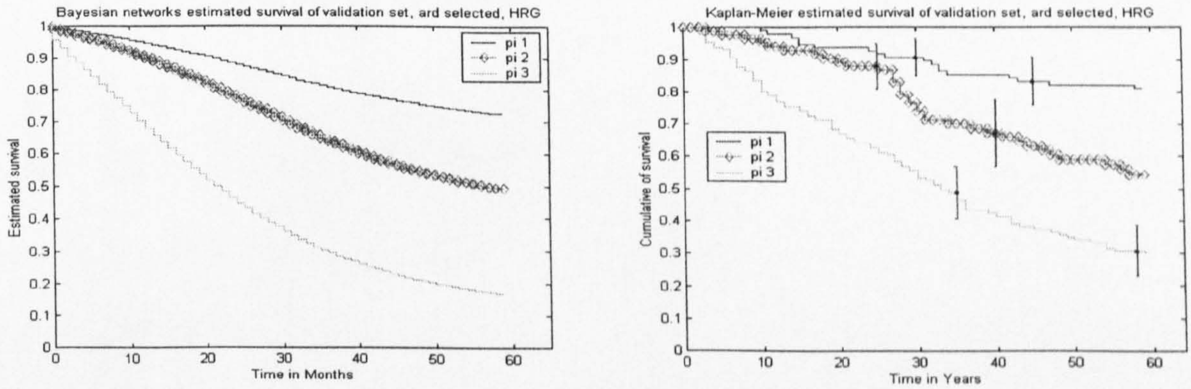


Figure (8.24): The neural networks predicted survivorship for the prognostic groups using the ARD selected model and the corresponding Kaplan-Meier estimate of the survivor functions. Both of the prognostic group 1 and 3 have shown a higher survivorship than the neural networks predicted.

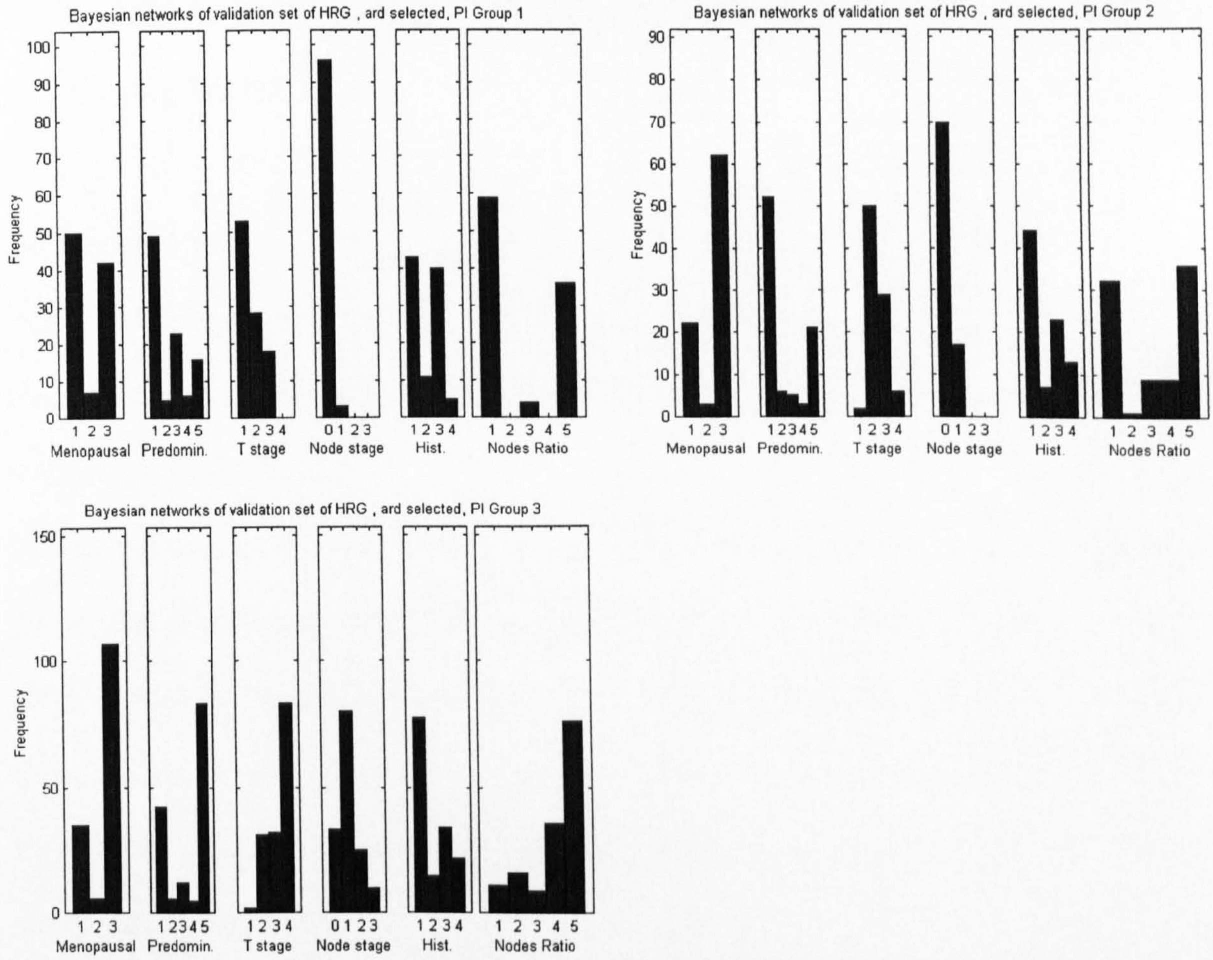


Figure (8.25): The attribute histograms for the prognostic groups of neural networks using the ARD selected model. The results show no distinguishable difference from the result for the high-risk cohort of design data on the same approach.

8.3.3 Validating the Cox regression method using the high-risk cohort of validation set involving interaction term

As shown in section (6.3), the best model for the high risk cohort consists of *clinical stage,* and the pairwise interaction *nodes ratio * tumour stage*. The results of the high-risk cohort of validation set fitted to this model are displayed in figures (8.26) - (8.28), combining the predictions from 3–fold cross validation. The Cox predicted survivorship for each prognostic group is as accurate as the Kaplan-Meier estimates and similar proportion of patients allocated to prognostic groups to the previous results in section (6.3). However, the results show that the predicted survivor function for prognostic groups 1 and 3 are different from the results shown in figure (6.8a). The predicted survival for prognostic group 1 has reduced from 0.78 to 0.66 and for group 3 it has increased from 0.08 to 0.3..

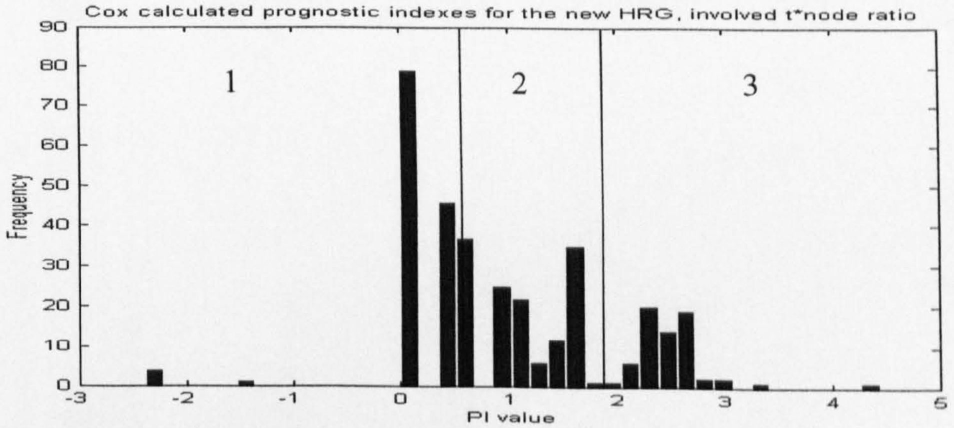


Figure (8.26): The Cox regression partitioned prognostic groups involving interaction term *tumour stage * nodes ratio*. The three prognostic groups contained 130, 138 and 66 patients respectively.

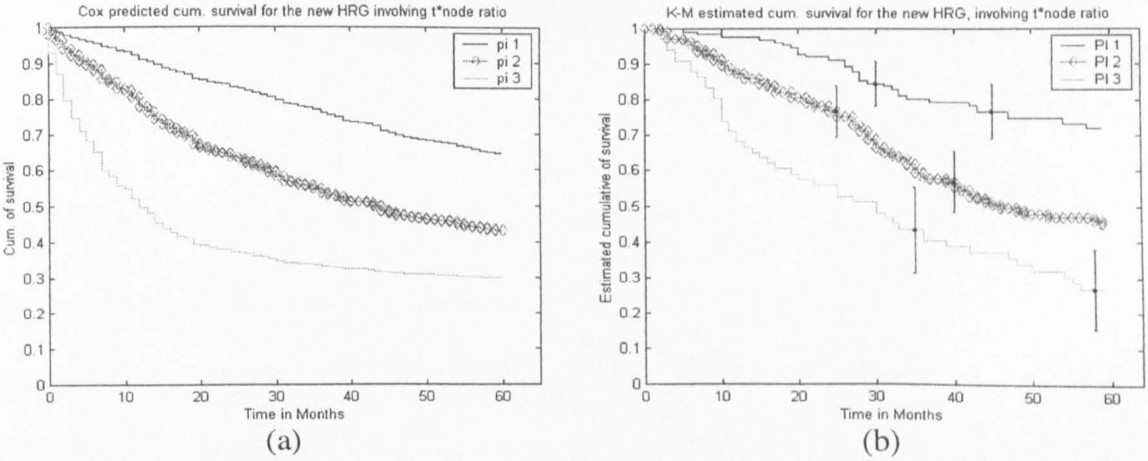


Figure (8.27): The Cox regression estimated survival curves for each of the prognostic groups involving interaction terms *nodes ratio * tumour stage* and their corresponding Kaplan-Meier estimated survival curves.

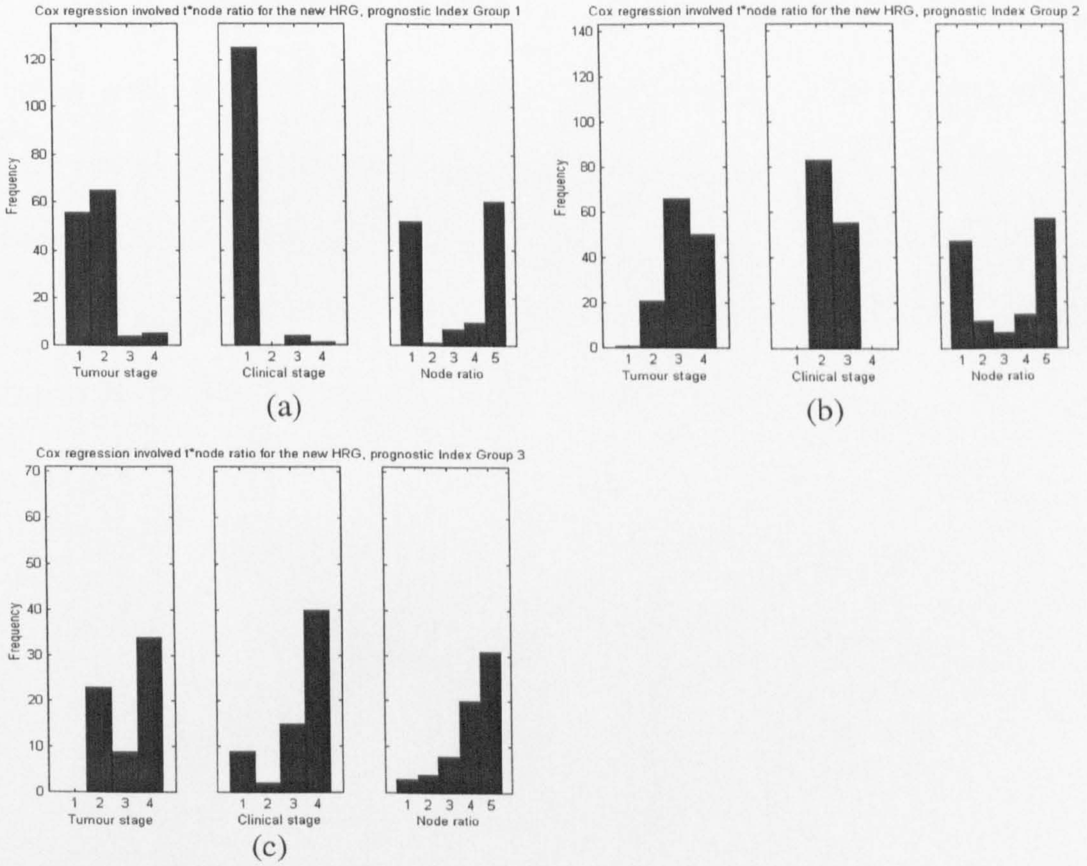
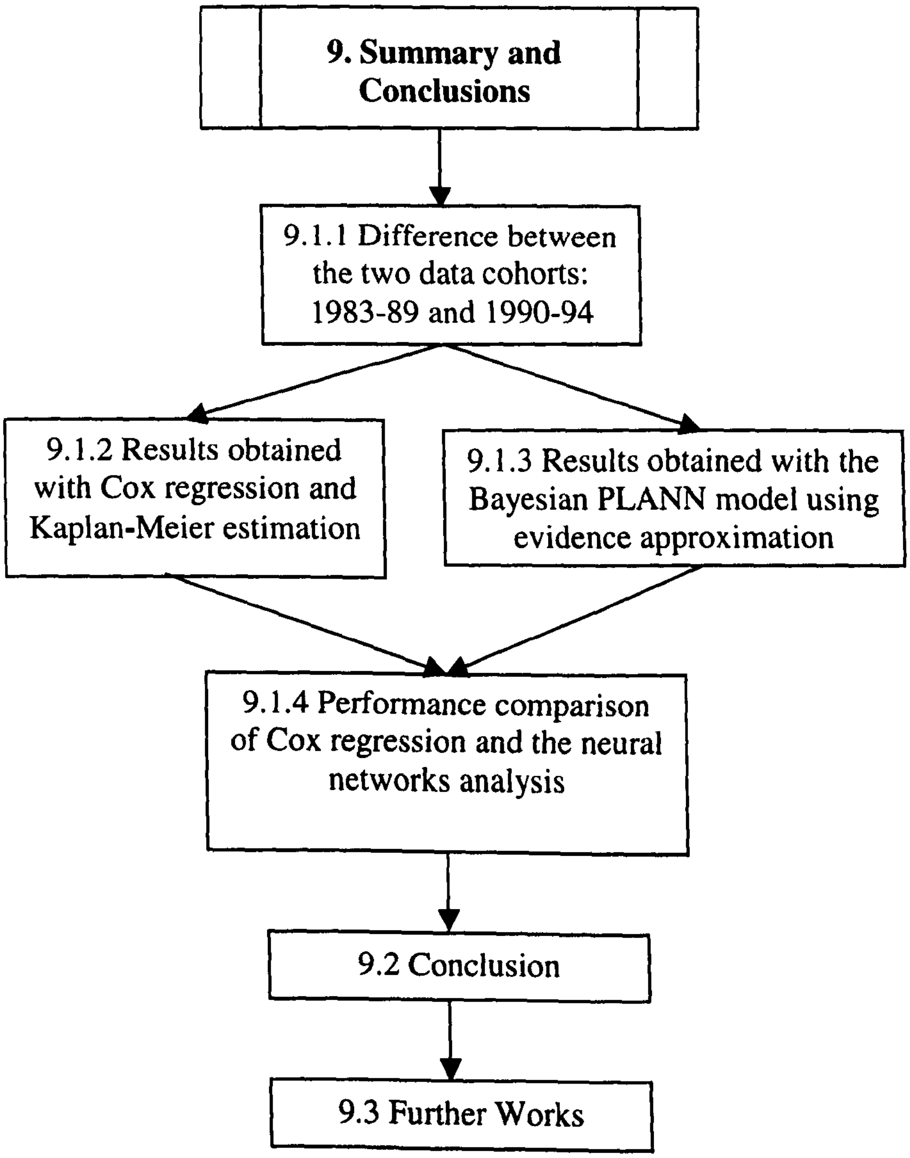


Figure (8.28): The attribute histograms for the prognostic groups of Cox regression involving interaction terms *tumour stage * nodes ratio.*

8.4 Discussion of test data set analysis

From the results for the validation data, it can be concluded that the overall survivorship improved since the previous cohort and the survival prediction by different modelling methods for the test data set is similar with the prediction made for the design data. The results also show change of population where the predicted survivorship is similar with the Kaplan-Meier estimate but they are different from the results for the design data, as shown in the high-risk cohort interaction analysis in section (8.3.3). In both models, the Cox regression and PLANN, the attribute profiles have not shown significant differences between the training and validation cohorts, with the exception of the high-risk cohort with a pairwise interaction term. The survivorship of breast cancer for the high-risk patients improved since the early of 90's at least by 0.2. On the other hand, the survivorship for the patient groups in the low-risk cohort has also improved but not as much as for the high-risk cohort. Moreover, the two sets of data have shown a different population distribution with fewer patients at the high risk.

Chapter 9


9. Summary and Conclusions
9.1.1 Difference between the two data cohorts: 1983-89 and 1990-94
9.1.2 Results obtained with Cox regression and Kaplan-Meier estimation
9.1.3 Results obtained with the Bayesian PLANN model using evidence approximation
9.1.4 Performance comparison of Cox regression and the neural networks analysis
9.2 Conclusion
9.3 Further Works

9.1 Summary

9.1.1 The changes of breast cancer prognosis within 10 years

Two sets of breast cancer data were considered in this thesis. They are cohorts of patients recruited by Christie Hospital during 1983 to 1989, and 1990 to 1993 each time with five years of follow-up. The attribute distributions in the two sets of data are different. There are fewer high-risk patients in the second cohort, and their 5-years survivorship is improved compared with same prognostic groups in the earlier cohort. This may reflect improvements in patient care, summarised in chapter (8).

Missing data are unavoidable, and this is present also in the second data set, even though the quality of the data provided is excellent. If there is only a relatively small amount of missing data in the entire data set, then those records can be simply discarded. Otherwise, they need to be handled carefully. Using the mean value of the variable is one of the commonly used methods to fill the missing values for continuous data. However, the situation becomes complicated in the case of categorical data. Within this thesis, we suggested using Nominal Logistic Regression to predict missing values, which required the identification a set of complete predictor variables for each variable with missing data. However, the values predicted by logistic regression may not be free of bias and there is no significant change to the survival predictions. Ripley (1998) also reported that filling in missing data using regression or other methods may not result in significant improvements to the data analysis. Therefore, treating the missing data as a separate category within the variable is a safe and efficient way to handle categorical missing data.

9.1.2 Results obtained with Cox regression and Kaplan-Meier estimation

9.1.2.1 Kaplan-Meier estimation

Kaplan-Meier estimation is a non-parametric model of survivorship. The role of this estimation in the thesis is to describe the survivorship of patients in different prognostic groups generated by the Cox regression or the neural networks, and hence to ascertain the accuracy of the survival predictions made by each modelling method. A 95% confidence interval is calculated for each survival curve.

9.1.2.2 Cox regression

Cox regression has been the method of choice in medical survival modelling since it was first proposed in 1972. The robustness, flexibility and commercial availability of software are some of the factors that contribute to the popularity of this method. Within this thesis, Cox regression was used both for variable selection and as a direct modelling methodology. The neural networks approach confirmed that the data are only slightly off the proportional hazard assumption. Cox models selected with the AIC criteria were still capable of producing good differentiation and accurate survival prediction for prognostic groups by mortality risk.

9.1.3 Neural network modelling with PLANN

In this thesis, the use of a Bayesian framework to regularise the Partial Logistic Artificial neural network with in the evidence approximation is demonstrated to be able to model censored survival data accurately on a monthly basis. The marginalisation procedure has to be improved because the output variable is not balanced between class labels, which involves a modification to the cost function, and the gradient and the Hessian calculations. This moderates the network towards the best unconditional estimate of the output which for us, is the mean hazard. Then Bayes' theorem is used to refer the estimates of the predicted hazard back to the true priors. Categorical data also has be handled differently in the PLANN network, by assigning one of the attributes as the baseline. Then the rest of the attributes corresponding to each variable share the same value of regularisation coefficient, and ARD is used for variable selection. Variable interactions can be naturally mapped within the network structure, but the explicit relationship between variables is difficult to trace.

The potential of the Bayesian regularisation framework applied to PLANN was explored in this thesis and it was concluded that the network performance in prognostic group differentiation and survival prediction is comparable to that of Cox regression, having the further advantage that:

- The proportionality of the hazards need not be observed.
- The network output is a smoothed hazard over time.

Since the Bayesian PLANN model is capable of handling non-linearity in the data, it is further capable of:

- Handling arbitrous interactions.
- Handling non-linear covariate time dependencies.
- Supporting variable selection, using ARD.

9.1.4 Performance comparison of Cox regression and the neural networks analysis

The neural network model and Cox regression separate patients into prognostic groups differently, as summarised in appendix (III). The neural network prognostic groups whose attribute profiles are more specific than the Cox regression without variable interactions. The ARD technique can be used for model selection, which has been demonstrated in the analysis of high-risk cohort, summarised in chapter (6) in which the selected variables implicitly take into account of variable interactions.

In the low-risk cohort, PLANN provided candidate terms for pairwise interactions, from which Cox regression found two pairs that contribute to two different prognostic groups, as described in section (6.5). However, these two pairs of interacting variables work against each other in Cox regression, but lead to better prognostic group separation if modelled with PLANN.

However, training the neural network for such large data set is computational time consuming and the ARD model selection process is also not straightforward. The weight decay hyper-parameters computed for each variable are not consistent when the network was trained repeatedly with different initial conditions, causing changes to their rank order. In this thesis, the network was trained with all available input variables at the beginning, then gradually eliminated the variables with the largest value of the weight decay parameter, alpha, until no more variables can be discarded without serious reduction in performance. There is no clear guidance to assist in the use of ARD for model selection. Moreover, the network predictions become less accurate when the data uncertainty is large, since they are marginalised towards averaged hazard of the data.

For its part, Cox regression is widely available in commercial and it is easy to use. Also the demand of computational time is limited. As the analysis of the validation data set in chapter (8) showed, the Cox regression is not much affected by the data uncertainty and produces good estimation of survivorship for each prognostic group. Moreover, it captures the shape of a survivor function over time in better detail.

The Cox regression performed well even when the proportional hazards assumption is not strictly observed and showed similar results when the method was tested by the validation data set, which has demonstrated the robustness of this approach.

However, the Cox regression in variable interactions must be pre-specified, but it is difficult to include all the combination of variable interactions for model selection when many variables are present.

Differentiation has been observed for the models with interaction terms for both cohorts in the design data, in which the prognostic groups are less overlapped and the accuracy of survival predictions for each group is considerably more accurate. On the other hand, the overall improvement of survivorship is apparent for all groups in the low-risk cohort of validation data set. However, there is no evidence suggested a systematic improvement in the high-risk cohort for any prognostic groups. Particularly for the lowest survival group, identified either by the models with or without interaction terms, has shown a clear survivorship improvement for each model but not the case for the other groups. Only the model with interaction term is capable of providing an accurate survival prediction for the lowest survival group while the other models suggest a lower survivorship should be for this group according to the covariate values.

9.2 Conclusions

Smith (2000) points out that an increasing number of patients do not benefit from systemic therapy, when systemic therapy is only offered to patients with tumours larger than 1cm in diameter. With the widespread use of mammographic screening programs, the average tumour is now in the 1.5-cm range at diagnosis. Throughout this thesis, the survivorship of prognostic groups in each cohort modelled in detail, in which a high risk patient group in the low-risk cohort was identified. A low risk patient group in the high-risk cohort was also identified, who might have gone through the rigors treatments unnecessarily.

In terms of the development of the neural network methodology for censored survival data, in chapter (5), the PLANN model was extended with regularisation within a Bayesian approximation for the hyperparameters. This gives an automatic determination of suitable values for the regularisation parameters requiring adjustments only to the number of nodes in the hidden layer. It results in smooth estimates of the discrete time hazard and allows for non-proportionality and non-linear interactions between covariates. In order to handle the categorical data more effectively, the ARD technique was modified to suit the data structure, in which several inputs corresponding to same variable share same value of alpha hyperparameter, and also the baseline attributes referral. The target distribution is very unbalanced, which requires a modification to the training algorithms and to the estimation of the conditional hazard with the result that the network outputs are marginalised towards the data averaged hazard. The use of ARD technique for soft pruning are also demonstrated, which is useful to determine a parsimonious neural network model.

In terms of the data analysis technique, we proposed dividing patients into prognostic groups using the log-rank test to group the calculated prognostic indexes into mortality risk group. This was interpreted by displaying the attribute profiles for each prognostic group using the selected variables. The extended Bayesian PLANN model and Cox regression were optimised by a monthly analysis of 5 years for two cohorts of patients, defined by clinical staging to be low- and high-risk cohort. In each cohort, prognostic indexes for mortality risk groups are formulated. For each method, the mean survivorship for each prognostic group is estimated and compared with the Kaplan-Meier estimate derived from the observed survival of those patients. It is summarised in chapters (4) and (5). Using PLANN to identify candidate pairwise interactions to include in a Cox regression model, a term involving *nodes ratio* * *tumour stage* was found to be useful in determining a specific high mortality group within the high-risk cohort, and for the low-risk cohort, two pairwise of interaction terms are also found, each corresponding to a high mortality group and a low mortality group separately. It is reported in chapter (6).

A second cohort of patients was used to validate the methodologies and their corresponding results for the first data set, which is summarised in chapter (8). Results showed that the population characteristics of two sets of data are different, for instance the two groups of patients identified by the interaction term *histology* * *node stage* from the low-risk cohort is no longer present in the second data set, appendix (III). A comparison of the model prediction with the Kaplan-Meier estimates shows an improvement in survival for the validation data set.

It was also proposed to use nominal logistic regression to predict the categorical missing data. However, the efforts have not been rewarded with significant improvement to the analysis. Therefore, treating the missing data as a separate category is the sage and most efficient way to handle categorical missing data, as summarised in chapter (7).

9.3 Further Work

The study could be extended to ten years allowing a more detailed study of possible deviations from the proportional hazards assumption over longer periods of time.

Accurate survival estimation would provide useful information to enable the clinicians and patients to make better informal discussion and decision regarding treatments and surgery.

9.4 List of publications

Lisboa P. J .G., Wong H, Harris P, Kirby S. P. J. and Swindell R. (1998) Survival of Breast Cancer Patients Following Surgery: A Detailed Assessment of the Multi-Layer Perception and Cox's Proportional Hazard Model. *International Joint Conference on Neural Networks.*

Wong H, Harris P, Lisboa P. J .G., Kirby S. P. J. and Swindell R. (1999) Dealing with Censorship in Neural Network Models, *International Joint Conference on Neural Networks.*

Lisboa P. J. G. and Wong H. (2001) Are Neural Networks Best Used to Help Logistic Regression? An Example from Breast Cancer Survival Analysis. *International Joint Conference on Neural Networks.*

Lisboa P. J. G., Vellido A. and Wong H. (2000) Bias Reduction in Skewed Binary Classification with Bayesian Neural Networks. *Neural Networks*, **13**, 407-410.

Wong H., Lisboa P.J.G., Harris P. and Swindell R. (2001) A Bayesian Neural Network Approach for Modelling Censored Data with an Application to Prognosis after Surgery for Breast Cancer. *Artificial Intelligence of Medicine*, (submitted).

Lisboa P. J. G., Vellido A. and Wong H. (2000) Outstanding Issues for Clinical Decision Support with Neural Networks. *Artificial Neural Networks in Medicine and Biology Conference.*

Lisboa P. J .G., Wong H, Harris P and Swindell R. (2001) A Retrospective Study of Breast Cancer Prognosis Using Artificial Neural Networks. *Neural Networks and Expert Systems in Medicine and Healthcare Conference.*

天接雲濤連曉霧，星河欲轉千帆舞。

彷彿夢魂歸帝所，聞天語，殷勤問我歸何處。

我報路長嗟日暮，學詩漫有驚人句。

九萬里風鵬正舉。風休住，蓬舟吹取三山去。

李清照 (born in 1084)

Acknowledgements

The author wish to thank the following persons for their continuing support and contribution for the completion of this thesis:

Paulo Lisboa, for his continuous patience and guidance after such many years of study, otherwise this thesis will never be finished.

Pete Harris, for his thoughtful advice, is always valuable.

Ric Swindell, Christie Hospital in Manchester, for making available of the breast cancer data.

Alfredo Vellido, for being my personal neural networks literature encyclopedia at any time and plenty of '*cooking ideas*'.

Barbara Lee, for her kindness and patience, listening for my endless moaning.

Of course, also Alan Seatwo, for his moral support and believing in my ability.

References

Altman D. G. and Andersen P .K. (1989) Bootstrap Investigation of the Stability of a Cox regression Model, *Statistics in Medicine*, **8**, 771-783

Altman D. G. and Lyman G. H. (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res. Treat.*, **52**, (1-3), 289-303

Anderson P. K. (1982) Testing Goodness of Fit of Cox's Regression and Life Model, *Biometrics*, **28**, 67-77

Arjas E. (1988) A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model, *Journal of the American statistical Association*, **83**, 401, 204-212.

Akaike H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. *Proc. 2nd Int. Symp. On Information Theory,* 267-281, Budapest: Akaemia Kiado.

Bennett S. (1983) Log-logistic regression models for survival data. *Applied Statistics*, **32**, 165-171.

Biganzoli E., Boracchi P., Mariani L. and Marubini E. (1998) Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach, *Statist. Med.*, **17**, 1169-1186.

Bishop C.M., (1995) *Neural network for pattern recognition,* Clarendon Press, Oxford

Brown S.F., Branford A.J. and Moran W. (1997) On the Use of Artificial Neural Networks for the Analysis of Survival Data. *IEEE Trans. Neural Netw.*, **8**, (5): 1072-1077.

Burke H. B., Goodman P. H., Rosen D. B., Henson D. E., Weinstein J. N., Harrell Jr., F. E., Marks J. R., Winchester D. P. and Bostwick D. G. (1997) Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction. *Cancer* **79**, (4): 857-862.

Burke H. B., Bosen D. B. and Goodman P. H. (1995) Comparing the Prediction Accuracy of Artificial Neural Networks and Other Statistical Models for Breast Cancer Survival. *In Advances in neural Information Processing System 7. Proceedings of the 1994 Conference*, 1063-1067, MA. MIT Press

Chen Y.Y. and Schnitt S. J. (1998) Prognostic Factors for Patients with Breast Cancer 1cm and Smaller. *Breast Cancer Res. Treat.*, **51**, (3), 209-225

Christensen E. (1987) Multivariate survival analysis using Cox's regression model, *Hepatology,* **7**, (6), 1346-1987.

Collett D. (1994), *Modelling Survival Data in medical Research,* Chapman & Hall, London.

Cox D. (1972) Regression models and life tables, *Journal of the Royal Statistical Society,* B, **74**, 187-220.

De Laurentiis M. and Ravdin P. M. (1994) A technique for Using Neural Network Analysis to Perform Survival Analysis of Censored Data. *Cancer Letters,* **77,** 127-138.

De Laurentiis M. and Ravdin P. M. (1994) Survival Analysis of Censored Data: Neural Network Analysis Detection of Complex Interactions between Variables. *Breast Canc. Res. Treat.*, **32**, 113-118.

Efron B. (1988) Logistic Regression, Survival Analysis, and the kaplan-Meier Curve, *Journal of the American Statistical Association*, **83**, 402, 414-425.

Faraggi D., Simon R., Yaskil E. and Kramar (1997) A. Bayesian Neural Network Models for Censored Aata. *Biometrica J,* **5**:519-532.

Fleming L. S. and Harrington D. P. (1991) Counting Processes and Survival Analysis, Wiley, New York

Gill R. (1987) A Simple Test for the Proportional Hazards Assumption, *Biometrika*, **72**, 2, 289-300

Gill R. and Schumacher M. (1987) A Simple Test of the Proportional Hazards Assumption. *Biometrika*, **74**, 2, 289-300

Gore S. M., Pocock S. J. and Kerr G. R. (1984) Regression Models and Non-proportional Hazards in the Analysis of Breast Cancer Survival, *Appl. Statist. 33*, 2, 176-195.

Groves D.J., Smye S.W., Kinsey S.E., Richards S.M., Chessells J.M., Eden O.B. and Bailey C.C. (1999) A comparison of Proportional hazards model and neural networks for risk stratification in cases of acute lymphoblastic leukaemia in children, *Neural Comput. & Applic.*, **8**, 257-264.

Hanley J. A. (1989) Receiver Operating Characteristic (ROC) methodology: the state of the art. *Critical Reviews in diagnostic Imaging*, **29**, 307-335.

Henderson I. C. and Patek A. J. (1998) The Relationship Between Prognostic and Predictive Factors in the Management of Breast Cancer. *Breast Cancer Res. Treat.*, **52**, (1-3), 261-288

Highleyman W.H. (1962) The design and analysis of pattern recognition experiments *Journal of Bell System technology*, 41, 723-744

Hopfield J. J. (1987) Learning Algorithms and Probability Distributions in Feed-forward and Feed-back Networks. *Proceeding of the national Academy of Sciences*, **84**, 8429-8433.

Kaplan E. L. and Meier P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481.

Kappen H. J. and Neijt J. P. (1993) Neural Network Analysis to Predict Treatment Outcome. *Annals of Oncology*, **4**, 31-34

Kay R. (1977) Proportional Hazard Regression Models and the Analysis of Censored Survival Data, *Appl. Statist*, **26**, 3, 227-237

Lagakos S. W. (1980) The Graphical Evaluation of Explanatory Variables in Proportional Hazard Regression Models, *Biometrika*, **68**, 1, 93-98

Lei L. J. (1992) The Accelerated Failure Time Model: A Useful Alternative to the Cox regression Model in Survival Analysis, *Statistics in Medicine*, **11**, 1871-1879

Liestøl, K., Andersen P. K. and Andersen, U. (1994) Survival Analysis andNneural Nets. *Stat. Med.*, **13,** 1189-1200.

Lin D. Y. and Wei L. J. (1991) Goodness-of-fit Tests for the General Cox regression Model, Statistical Sinica, **1**,1-17

Lisboa P.J.G., Vellido A. and Wong H. (2000) Bias reduction in skewed binary classification with Bayesian neural networks, *Neural Networks*, **13**, 407-410

MacKay D.J.C. (1992a) Bayesian interpolation in Neural Computation, **4**, (3), 415-447.

MacKay D. J. C. (1992b) The evidence framework applied to classification networks, *Neural Computation*, **4**, (5), 720-736.

Mackay D. J. C. (1994a) Bayesian Methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten (Eds.), *Models of Neural Networks III*, Chapter 6, New York: Springer-Verlag.

MacKay D. J. C. (1994b) Hyperparameters: Optimise or Integrate out? In G. Heidbreder (Ed.), Maximum Entropy and Bayesian Methods, Santa Barbara 1993, Bordrecht: Kluwer.

Mackay D. J. C. (1995) Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks, *Networks: computation in Neural Systems*, **6**, 469-505.

Magee B., Swindell R., Harris M. and Banerjee SS (1996) Prognostic Factors for Breast Recurrence after Conservative Breast Surgery and Radiotherapy: Results from a Randomised Trial, *Radiother. Oncol*, **39**,(3), 223-227

Magee B., Swindell, R., Harris M. and Banerjee S. S. (1996) Prognostic Factors for Breast Recurrence after Conservative Breast Surgery and Radiotherapy: Results from a Randomised Trial, *Radiotherapy and Oncology*, **39**, 223-227

Mantel N. and Haenszel W. (1959) Statistical aspects of the analysis of data form retrospective studies of disease, *Journal of the national Cancer Institute*, **22**, 719-748.

McCready D. R., Chapmen J. A., Hanna W. M., Kahn H. J., Murray D., Fish E. B., Trudeau M. E., Andrulis I. L. and Livkpey H. L. (2000) Factors Affecting Distant Disease-free Survival for Primary Invasive Breast Cancer: use of a Log-normal Survival Mdoel. *Ann. Surg. Oncol.*, **7**,(6), 416-426

McCulloch W. S and Pitts W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bull. Math. Biophys.* **5**, 115.

Møllar M. (1993b) A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural networks*, **6**, (4), 525-533

Nariani L., Coradini D., Biganzoli E., Boracchi P., Marubini E., Pilotti S., Salvadori B., Silvestrini R., Veronesi U., Zucali R. and Rilke F. (1997) Prognostic Factors for Metachronous Contralateral Breast Cancer: A Comparison of the Linear Cox Regression Model and its Artificial Neural Network Extension. *Breast Cancer Research and Treatment*, **44**, 167-178

Nina Pope, Breast Cancer Awareness Campaign (1995), http://hosted.aware.easynet.co.uk/

Ohno-Machado L. A Comparison of Cox Proportional Hazards and Artificial Neural Network Models for Medical Prognosis. *Comput. Biol. Med.* 1997, **27**, (1), 55-65

Ohno-Machado L., Walker M. G. and Musen M. A. (1995) Hierarchical Neural Networks for Survival Analysis. *In MEDINFO95, The Eighth Congress on Medical Informatics.*

Pettitt A. N. and Daud I. B. (1990) Investigating Time Dependence in Cox's Proportional Hazards Model, *Appl. Statist.*, **39**, 3, 313-329

Peto R. and Peto J. (1972) Asymptotically Efficient Rank Invariant Procedures. *Journal of the Royal Statistical Society, A,* **135**, 185-207

Radvin P. M. and Clark G. M. (1992) A practical application of neural network analysis of prediction output of individual breast cancer patients, *Breast Cancer Research and Treatment*, **22**, 285-293

Ravdin P. M., Clark F. M., Hilsenbeck S. G., Owens M. A., Vendely P., Pandian M. R. and McGuire W.L. (1992) A demonstration that breast cancer recurrence can be predicted by neural network analysis, *Breast Cancer Research and Treatment*, **21**, 47-53.

Ripley R. M. (1998) Neural Networks for Breast Cancer Prognosis, thesis of St Cross College, Oxford University

Ripley B. D. and R. M. Ripley , (1998) Neural Networks and Statistical Methods in Survival Analysis, *In Artificial Neural Networks: Prospects for Medicine,* Landes Biosciences

Ripley R. M., Haris A. L. and Tarassenko L. (1998) Neural Network Models for Breast Cancer Prognosis, *Neural Comput. & Applic.*, **7**, 367-375.

Sauerbrei W. and Schumacher M. (1992) A Bootstrap Resampling Procedure for Model Building: Application to the Cox regression Model, *Statistics in Medicine*, **11**, 2093-2109

Schoenfeld B. (1980) Chi-squared Goodness-of-fit Tests for the Proportional Hazards Regression Model. *Biometrika*, **67**, 1, 145-153

Schoenfeld D. (1982) Partial Residuals for the Proportional Hazards Regression Model, *Biometrika*, **69**, 1, 239-241

Smith B. L. (2000) Approaches to Breast Cancer Staging, *The new England Journal of Medicine,* **24**, 580-581

SPSS Base 9.0/user's guide (SPSS Inc. 1999).

Stablein D M., Carter W. H. Jr. and Novak J. W. (1981) Analysis of Survival Data with Nonproportional Hazard Functions, *Controlled Clinical Trials*, **2**, 149-159

Tibshirani R. (1982) A Plain Man's Guide to the Proportional Hazards Model, *Clinical Investigative Medicine*, **5**, 1, 63-68

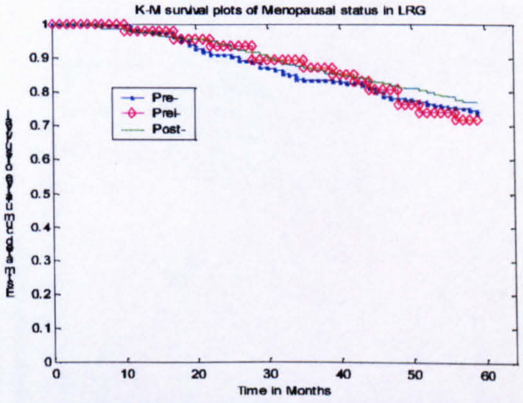
Tarassenko L., Whitehouse R., Gasparini G. and Harris A. L. (1996) Neural Netowrk Predciton of Relapse in Breast Cancer Patients. *Neural Comput & Applic.*, **4**, 105-113

Vonta F. and Karagrigoriou A. (1998) Retrospective Study of Primary Breast Carcinoma. *Conf. in data Sc. Classf. And Related Math.*, 337-341

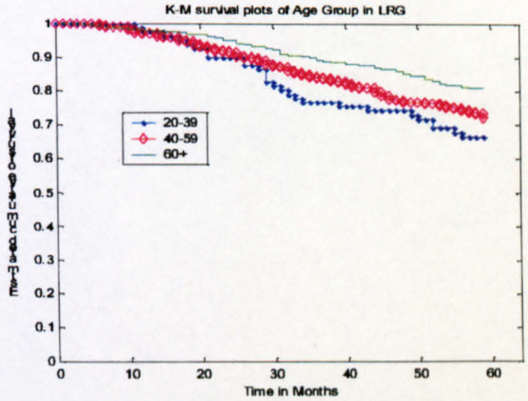
Williams M. R., Hinton C. P., Todd J. H., Morgan D. A. L., Elston C. W. and Blamey R. W. (1985) The Prediction of Local or Regional Recurrence after Simple Mastectomy for Operable Breast Cancer, *Br. J. Surg*, **72**, 721-723

Appendix (I)

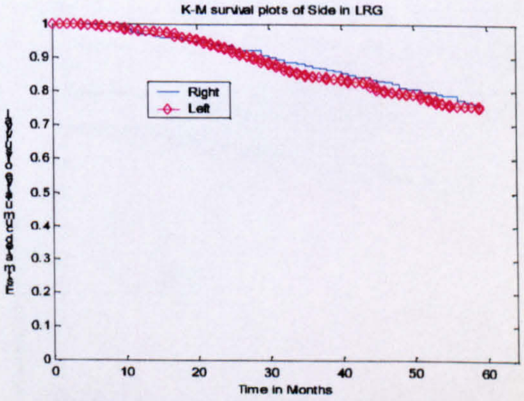
Kaplan-Meier estimate of the survival functions for the low-risk cohort of design data



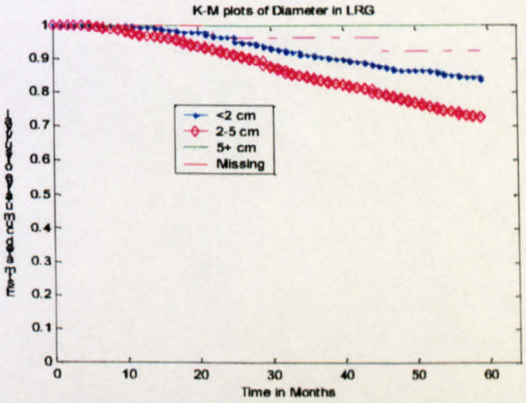
Menopausal Status



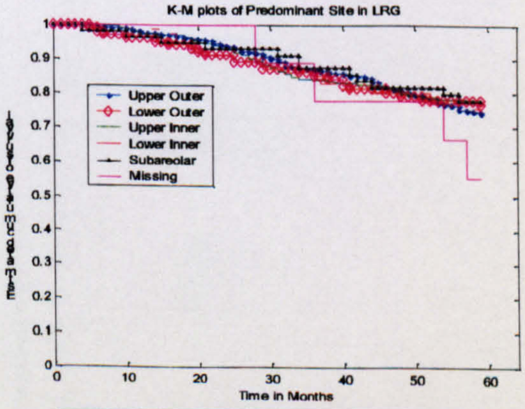
Age Group



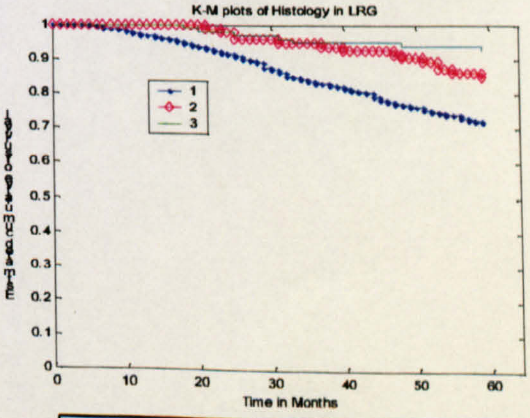
Side



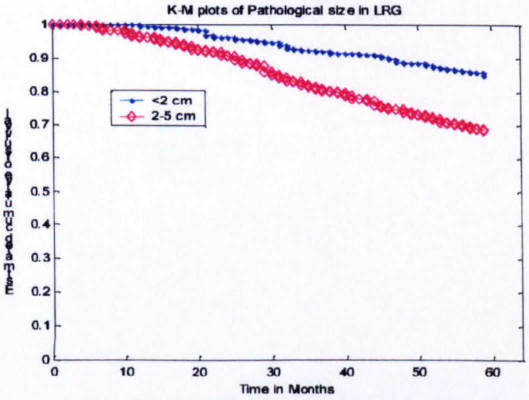
Diameter



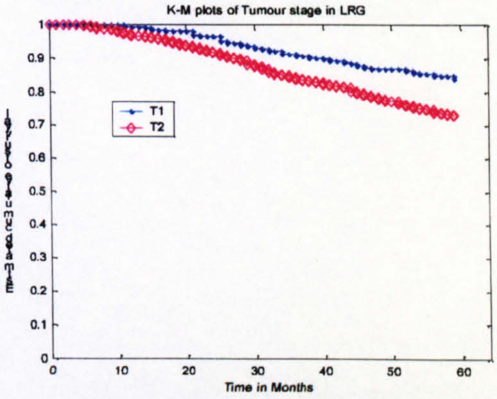
Predominant Site



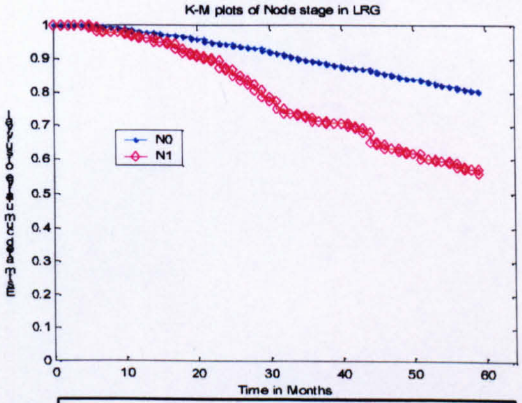
Histology



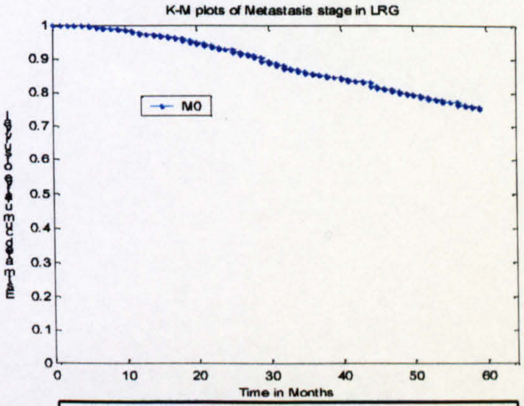
Pathological Size



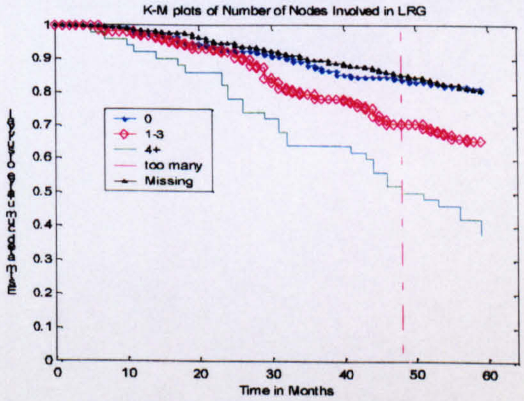
Tumour Stage



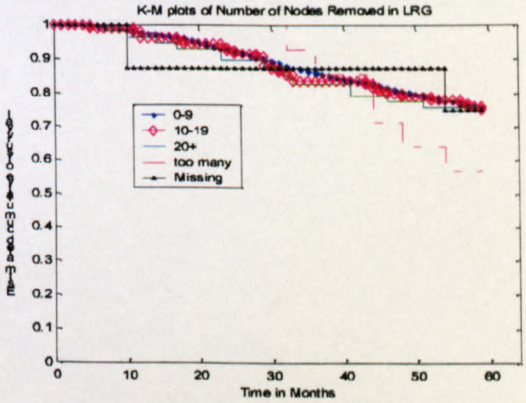
Node Stage



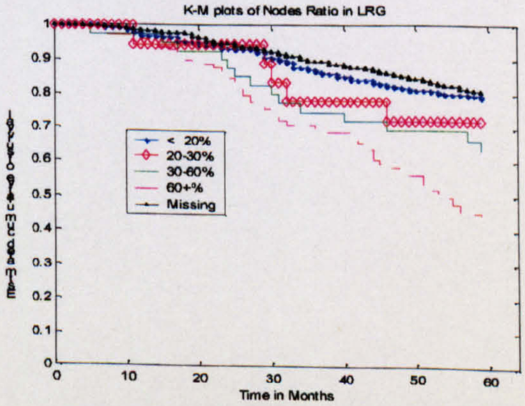
Metastasis Stage



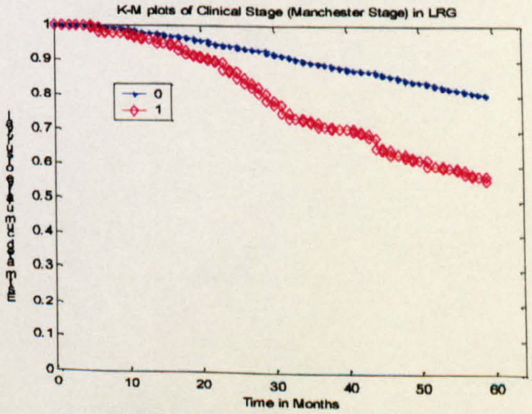
Number of Nodes Involved



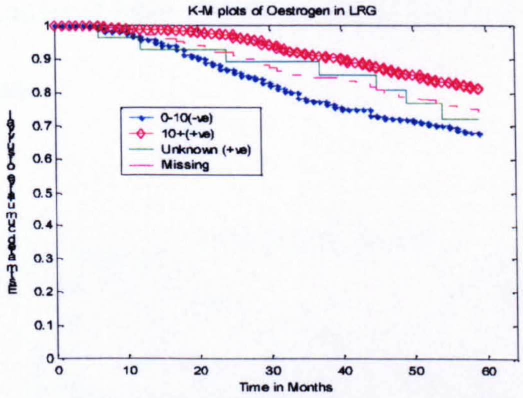
Number of Nodes Removed



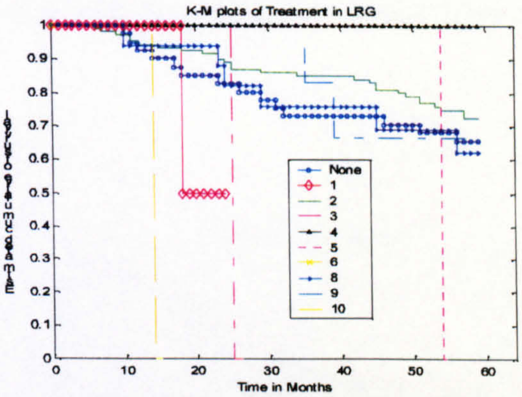
Nodes Ratio



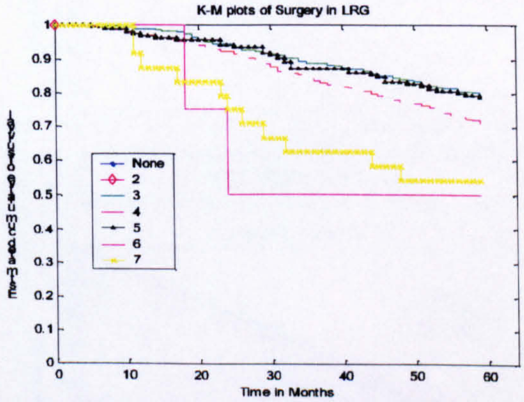
Clinical Stage (Manchester Stage)



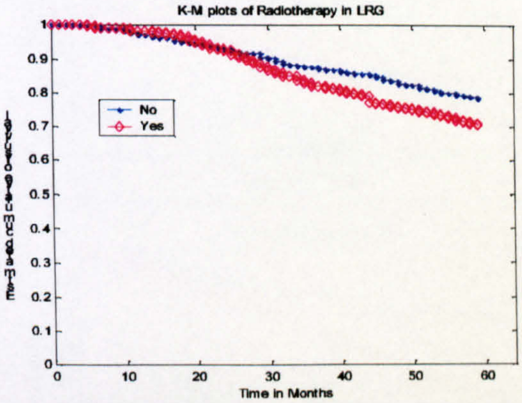
Oestrogen



Treatment

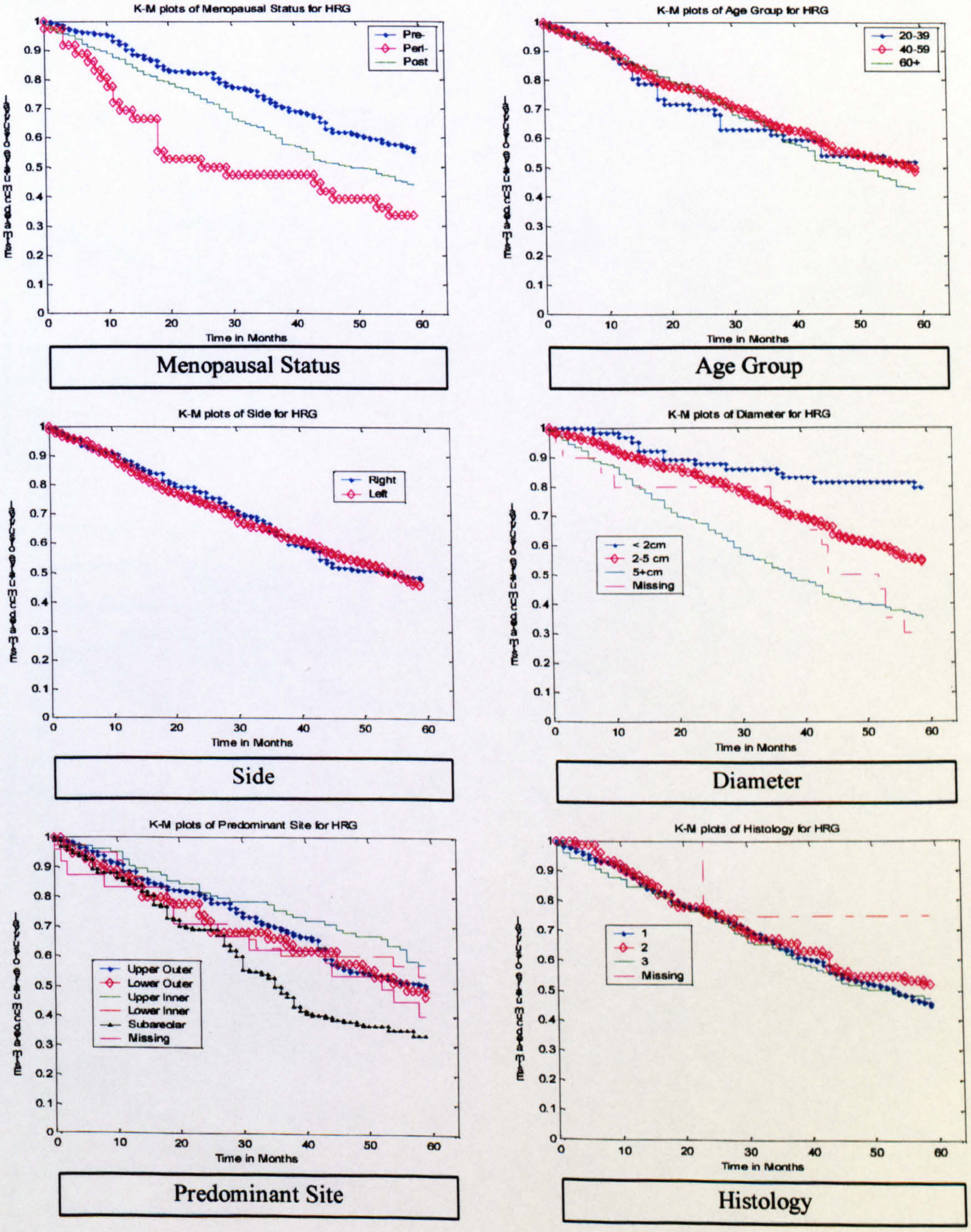


Surgery



Radiotherapy

Kaplan-Meier estimate of the survival functions for the high-risk cohort of design data.



K-M plots of Menopausal Status for HRG

Menopausal Status

K-M plots of Age Group for HRG

Age Group

K-M plots of Side for HRG

Side

K-M plots of Diameter for HRG

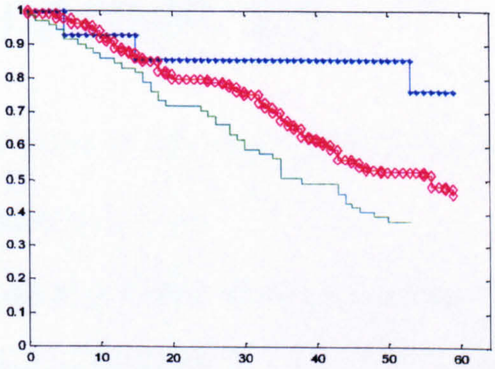
Diameter

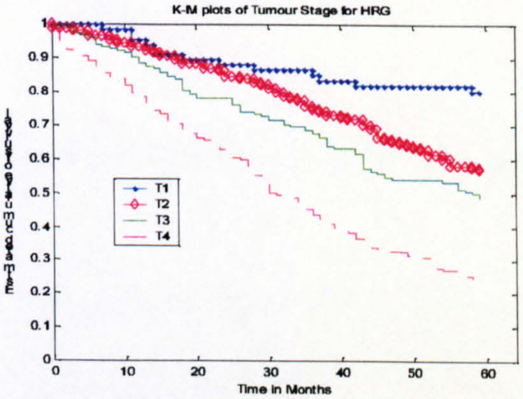
K-M plots of Predominant Site for HRG

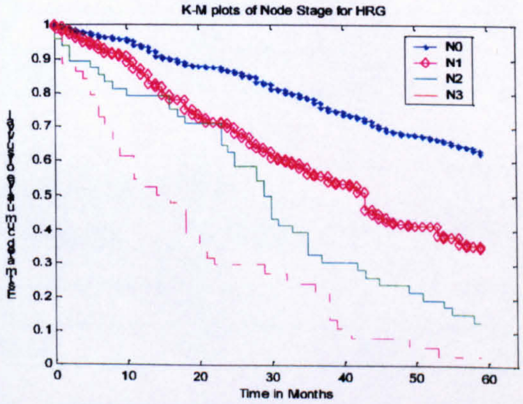
Predominant Site

K-M plots of Histology for HRG

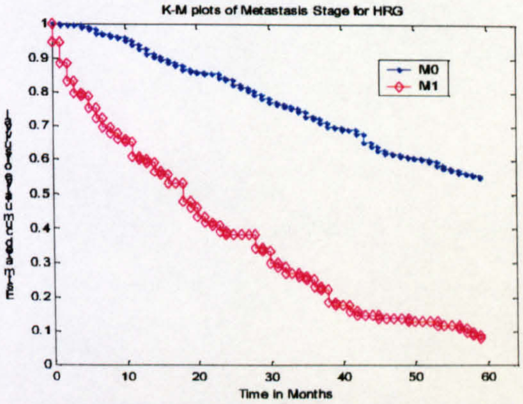
Histology



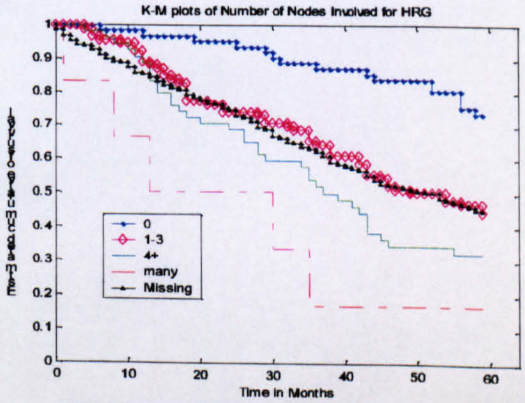




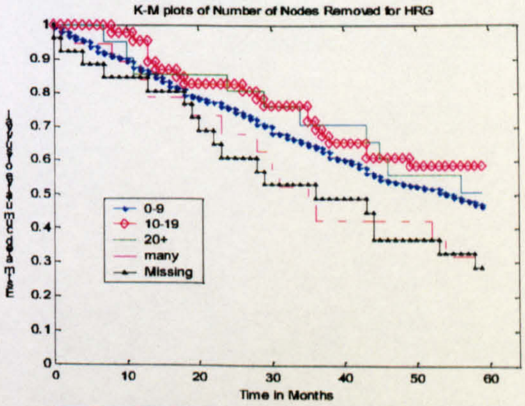
Node Stage



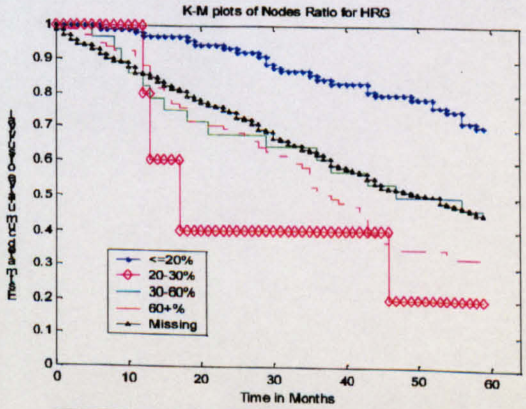
Metastasis Stage



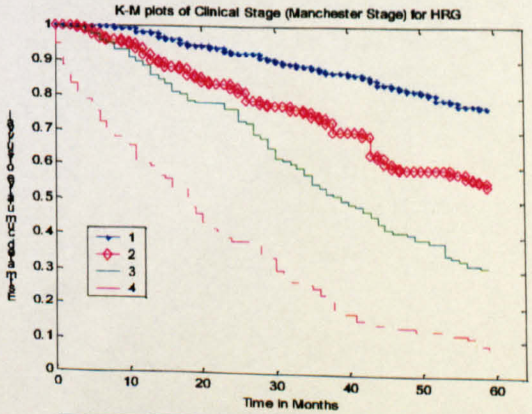
Number of Nodes Involved



Number of nodes Removed



Nodes Ratio



Clinical Stage (Manchester Stage)_

Appendix (II)

The variation of sample distribution for the selected variables of the design and the validation data set.

Low-Risk Cohort of the Design Data

Clinical Node Stage		
Category	Frequency	Percentage
0	734	80
1	183	20
total	917	100

Pathological Size		
Category	Frequency	Percentage
1	383	41.8
2	534	58.2
total	917	100

Histology		
Category	Frequency	Percentage
1	724	79
2	95	10.4
3	98	10.7
total	917	100

Nodes Ratio		
Category	Frequency	Percentage
<20%	256	27.9
20-30%	18	2.0
30-60%	40	4.4
60+%	98	10.7
Missing	505	55.1
total	917	100

Low-Risk Cohort of the Validation Data

Clinical Node Stage		
Category	Frequency	Percentage
0	800	85.9
1	131	14.4
total	931	100

Pathological Size		
Category	Frequency	Percentage
1	541	58.1
2	390	41.9
total	931	100

Histology		
Category	Frequency	Percentage
1	633	68.0
2	111	11.6
3	187	20.1
total	931	100

Nodes Ratio		
Category	Frequency	Percentage
<20%	624	67.0
20-30%	28	3.0
30-60%	51	5.5
60+%	43	4.6
Missing	185	19.9
total	931	100

High Risk Cohort of the Design data

Menopausal Status		
Category	Frequency	Percentage
1	177	28.0
2	36	5.7
3	420	66.4
Total	633	100

Clinical Node Stage		
Category	Frequency	Percentage
0	355	56.1
1	186	29.4
2	48	7.6
3	44	7.0
Total	633	100

Clinical Stage		
Category	Frequency	Percentage
1	184	29.1
2	171	27.0
3	165	26.1
4	113	17.9
Total	633	100

Predominant Site		
Category	Frequency	Percentage
1	256	40.4
2	54	8.5
3	103	16.3
4	38	6.0
5	158	25.0
Unknown	24	3.8
Total	633	100

Tumour Stage		
Category	Frequency	Percentage
1	68	10.7
2	186	29.4
3	156	24.6
4	223	35.2
Total	633	100

High-Risk Cohort of the Validation data

Menopausal Status		
Category	Frequency	Percentage
1	107	32.0
2	16	4.8
3	211	63.2
total	334	100

Clinical Node Stage		
Category	Frequency	Percentage
0	199	59.6
1	100	29.9
2	25	7.5
3	10	3.0
total	334	100

Clinical Stage		
Category	Frequency	Percentage
1	57	17.1
2	109	32.6
3	79	23.7
4	89	26.6
total	334	100

Predominant Site		
Category	Frequency	Percentage
1	143	42.8
2	17	5.1
3	40	12.0
4	14	4.2
5	59	17.7
Unknown	61	18.3
Total	334	100

Tumour Stage		
Category	Frequency	Percentage
1	57	17.1
2	109	32.6
3	79	23.7
4	89	26.6
Total	334	100

High Risk Cohort of the Design data

Pathological Size		
Category	Frequency	Percentage
<2cm	14	2.2
2-5cm	134	21.2
5+cm	71	11.2
Missing	414	65.4
Total	633	100

Histology		
Category	Frequency	Percentage
1	362	57.2
2	78	12.3
3	189	29.9
Missing	4	0.6
total	633	100

Nodes Ratio		
Category	Frequency	Percentage
<20%	91	14.4
20-30%	5	0.8
30-60%	28	4.4
60+%	92	14.5
Missing	417	65.9
total	633	100

High-Risk Cohort of the Validation data

Pathological Size		
Category	Frequency	Percentage
<2cm	12	3.6
2-5cm	51	15.3
5+cm	51	15.3
Missing	220	65.9
Total	633	100

Histology		
Category	Frequency	Percentage
1	164	49.1
2	33	9.9
3	97	29.0
Missing	40	12.0
total	334	100

Nodes Ratio		
Category	Frequency	Percentage
<20%	102	30.5
20-30%	17	5.1
30-60%	22	6.6
60+%	45	13.5
Missing	148	44.3
total	334	100

Appendix (III)

The number of patients in each prognostic group of design data for Cox regression and neural network models.

Low-risk cohort: D – Design data (917 cases), T - Test data (931 cases)

	Prognostic group 1		**Prognostic group 2**		**Prognostic group 3**		**Prognostic group 4**	
	D	**T**	**D**	**T**	**D**	**T**	**D**	**T**
Cox Regression	127	229	189	355	487	237	114	110
Neural Networks using Cox Model	56	126	359	461	460	328	42	16
Cox Regression Involving Interaction Term (*Histology*Node Stage*)	61	0	207	610	579	321	68	0
Cox Regression Involving Interaction Term (*Nodes ratio*Node Stage*)	116	237	331	375	427	303	43	16

High-risk cohort: D – Design data (613 cases), T - Test data (334 cases)

	Prognostic group 1		**Prognostic group 2**		**Prognostic group 3**	
	D	**T**	**D**	**T**	**D**	**T**
Cox Regression	171	126	275	129	187	78
Neural Networks using Cox Model	248	125	174	69	211	140
Neural Networks ARD model	244	99	171	87	218	148
Cox Regression Involving Interaction Term (***nodes ratio*tumour stage***)	214	130	278	138	139	66