PREDICTION OF SURVIVAL OF EARLY STAGES LUNG CANCER PATIENTS BASED ON ER BETA CELLULAR EXPRESSIONS AND EPIDEMIOLOGICAL DATA

by

EVGENY MARTINENKO
M.S. St. Petersburg Polytechnical Inst., 1990
M.S. of Physics, 2001

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematics
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2011

Major Professor: Marianna Pensky

ABSTRACT

We attempted a mathematical model for expected prognosis of lung cancer patients based on a multivariate analysis of the values of ER-interacting proteins (ERbeta) and a membrane bound, glycosylated phosphoprotein MUC1), and patients clinical data recorded at the time of initial surgery. We demonstrate that, even with the limited sample size available to use, combination of clinical and biochemical data (in particular, associated with ERbeta and MUC1) allows to predict survival of lung cancer patients with about 80% accuracy while prediction on the basis of clinical data only gives about 70% accuracy. The present work can be viewed as a pilot study on the subject: since results confirm that ER-interacting proteins indeed influence lung cancer patients' survival, more data is currently being collected.

To my parents.

ACKNOWLEDGMENTS

I would like to express my appreciation and gratitude to my professors, family, and friends for their support throughout my graduate studies. I am tremendously thankful for the patient supervision and guidance of Dr.Marianna Pensky. Her insightfulness and constructive comments have been instrumental in the completion of this thesis. I also would like to thank Dr.Tatiana Zhukov whose idea initiated current work and who supplied us with all biochemical data. I also want to say thanks to Mr. Raynald Levesque whose wonderful website helped me to safe a lot of time preparing SPSS code

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Lung cancer is the most common cause of mortality from malignancy throught the world [6]. Despite advances in surgical and chemotherapy treatments, the survival time of lung cancer patients in USA has not significantly improved in the past 25 years, remaining shorter than the corresponding survival times for colon cancer, breast cancer and prostate cancer.

The objective of current work is to predict survival of the lung cancer patients, based on variety of epidemiological and biochemical data. The biochemical data referred to expression of two biochemical markers: Estrogen Receptor protein (ER) and MUC1 (membrane-bound protein of the mucin family). Those two biomarkers have a long history of being associated with cancer. In particular, ER is believed to be involved in some aspects of carciogenesis [5] by either being activated by its ligand estrogen or by other pathways such as ligant-independent receptor activation via growth factor receptor. As for MUC1, several studies have suggested that 2 by a variety of mechanisms [7, 3, 10]. It may account for greater metastatic ability and also prevents formation of conjugates with lymphokine-activated killer cells and cytotoxic T-lymphocytes.

The motivation of our study of ER and MUC1 comes from the disparity between lung cancer survival rates of males and females. The lung cancer death rate in women has doubled over the past 25 years, while the male lung cancer death rate has continued to decline. Although several lines of evidence suggest that women may be more susceptible to develop tobacco-induced lung cancer than men, we lack definitive results related to gender disparity in lung cancer survival.

We want to test hypothesis that estrogen through interaction with estrogen receptors (ERs) may mediateactivity of ER-interacting proteins and affect their function in corresponding cell signaling pathways. Potentially, it may modulate the DNA damage/repair signaling network and lead to chromosomal instability and female lung cancer progression,

that might be the reason for gender differences in lung cancer survival. There exists [11] significant increases in expression of ERbeta and MUC1 in a subgroups of malignant tumors compared to normal adjacent lung. Mathematical modeling and correlation analysis of clinical data on differential lung cancer survival (after adjustment for patients gender, age, race, histology type, tumor stage at diagnosis, follow-up for recurrence and smoking history) with criteria of biomarkers cellular expression is presentewd. These data may open a new way to look on mechanisms of chromosomal instability under estrogen control in female lung cancer progression, and introduce an attractive, novel therapeutic targets. In particular, this approach may help to indicate novel targets for personalized chemotherapy following surgery to prevent lung cancer progression.

The objective of the study is to develop a mathematical model for expected prognosis of lung cancer patients based on a multivariate analysis of the values of ER-interacting proteins (ERbeta and MUC1), and patients clinical data recorded at the time of initial surgery. IN particular, the goal of this model is to predict survival of the lung cancer patients withing a 4 year period starting from the date of diagnosis. For this purpose we divide all the patients into two classes: long survival (more than 4 years) or short survival (less than 4 years) and formulate the problem as a classification problem where a patients needs to be assigned to one of these two classes on the basis of the clinical and biochemical record.

CHAPTER TWO: BACKGROUND

2.1 Specimen preparation and scanning

After approval by the Institutional Review Board, archived blocks of lung tumor resected from 33 individuals surgically treated for lung cancer, were provided by the Tissue Core of the Moffitt Cancer Center. We selected a set of lung cancer specimens from individuals with no prior chemotherapy. We obtained snap-frozen samples of the tumors; samples of formalin fixed tumors were paraffin embedded, and sections cut at 3 m from 33 lung cancer cases. We obtained correlative clinical information including gender, age, smoking history, tumor stage, grade and histologic type. Private identifiable patient information has been removed from these records in accordance with IRB and HIPAA regulations. Slides of paraffin sections stained with H&E, that correspond with primary tumor were reviewed prior to inclusion according to established morphological criteria [], and to assure presence of adjacent, uninvolved lung and pre-malignant lung lesions (peripheral AAH), as previously described [1, 1].

In this experiment, we used the following methodologies:

a) Clinical tissue specimens and patient information: Moffitt Tumor Tissue Bank provided surgically removed tumors from which we chose specimens of lung cancer patients without prior chemotherapy. We acquired samples of snap-frozen and formalin-fixed tumors in order to verify the pathologic diagnoses;

b) Immunohistochemistry (IHC)- we optimized this procedure on formalin fixed paraffin-embedded LC tissue sections (thickness:4 microns), the subsequent tissue sections were used for IHC for MUC1 and ERs followed by quantitative image analysis of biomarker expression.

Estrogen receptor immunohistologic staining: consecutive sections from the same tissue blocks described above were used for staining with antibodies to ER-alpha and ER-. We op-

timized procedure for immunostaining of ER- with mouse monoclonal antibody to Oestrogen R-beta 1, MCA1974 (Serotec, UK). Deparaffinized sections underwent antigen retrieval in citrate buffer under microwaves (700 W), blocking with avidin/biotin and then incubation with primary antibody at 1:400 dilution, applied overnight at 4C in a humid chamber. IHC was completed on the DAKO autostainer

PX Mouse detection and DAB chromogen. ER- immunohistologic staining on tissue sections was performed after antigen retrieval described above, and with application of the standard Ventana test (anti-ER-, clone 6F11). The positive control for ER-alpha and ER- were breast cancer and uterus tissues; in the negative control primary antibodies were replaced with PBS, following secondary antibody detection technique.

Histology slides were scanned using the Aperio (Vista, CA) ScanScope XT with a 200x/0.8NA objective lens at a rate of 2 minutes per slide via Basler tri-linear-array. Image analysis was performed using an Aperio Positive Pixel Count v9.1 algorithm with the following customized thresholds [Hue Value =.2; Hue Width =.6; Color Saturation Threshold =0.05; IWP(High) = 210; Iwp(Low)=Ip(High) = 160; Ip(low) =Isp(High) =80 Isp(Low) =0] . The algorithm was applied to the entire scanned slide image to detect regions of increased ER and MUC1 expression by detecting pixels that satisfy the color and intensity specification defined above.

2.2 Description of the data

Our analysis was based on the data for 33 patients: 13 females and 20 males. For each of the patients, we had clinical and biochemical data. Epidemiological data was derived from the Florida Cancer Data System (FCDS) while the biochemical data was obtained as described above. The epidemiological data are listed in the Table 1. The patients in the group were ages 57 to 84 with average age at diagnostic for females being 68 years, and for males 71.5 years. Out of 24 patients, at the moment of recording, 11 patients were alive and 22 were dead. In terms of smoking history all patients were divided in the following

Table 2.1: Epidemiological variables

Variable Name	Meaning	Type	Range
Age	Age at diagnosis	Numerical	57-84
Gender	Gender	Categorical	m or f
Status	Vital Status	Categorical	dead or alive
Time	Survival time in months	Numeric	1 -177
Site	Primary location of tumor	Categorical	5 sites
Hist	Hitological type of cancer	Categorical	7 types
Seq	Sequence Number	Categorical	5 numbers
TNM	TNM Mixed Stage	Categorical	7 stages
Smoking	History of tobacco use	Categorical	4 categories
Treat	methods of treatment	Categorical	5 methods

groups: non-smokers, history of smoking, light, moderate or heavy smoking, non-specified or unknown.

The biochemical data were extracted from cancer and adjacent normal cells and referred to two types of biomarkers: ER and MUC1. While ER is activated in the nucleus of the cells, the MUC1 marker is located in the membranes.

The effects of ER and MUC1 are represented by the grey level of the image (of the nuclei for ER and of the membranes for the MUC1) and is subsequently divided into four categories (3+, 2+, 1+ and 0+). In particular, the ER data is represented by the following characteristics: percentages and numbers of cells in each of the four the categories and intensity score (the most represented intensity). The MUC1 data is represented by the following characteristics: percentages and numbers of cells in each of the four the categories, percentage of the membrane stained, the number and the percentage of the total number of

cells and the complete cells.

The total number of epidemiological characteristics for each patient is 11 while the total number of biochemical characteristics is 26. Therefore, the total number of features, 37, exceeds the sample size (even if the male and the female patients are bundled together for classification purposes). In addition, many of the characteristics are redundant (e.g., the percentage and the number of cells with certain intensity). This fact makes it necessary to reduce dimensionality of the data via the model selection process, i.e. by selecting the features which are useful for classification purposes and discarding the rest of them.

2.3 Description of the methods

2.3.1 Multiple linear regression with stepwise choice of predictors

Multiple liner regression attempts to model the relationship between several explanatory variables and a response variable by fitting a linear equation to observed data. A general form of a multiple linear regression model (see for example [2]) is given by :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

or in matrix form :

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$$

where:

$\vec{y}^T = (y_1, y_2, \ldots, y_N)$ is a response variable,

$\vec{\beta}^T = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ are the regression coefficients,and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} \ldots & x_{1,p} \\ . & .... & . \\ . & .... & . \\ . & .... & . \\ 1 & x_{N,1} \ldots & x_{N,p} \end{pmatrix}$$

is the matrix of explanatory (independent) variables, and finally $\vec{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_N)$ is the vector of random errors which are assumed to have zero mean and are independently sampled from the same distribution which has a finite variance σ^2.

The regression coefficients are estimated using the least squares principle. If the matrix $\mathbf{X}$ is of full column rank, the ordinary least square estimator for$\vec{\beta}$ exists and given by

$$\vec{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y}.$$

Note that in our case the assumption of the full column rank is violated since $N < p$. In this situation, it is desirable to select a subset of variables as predictors (explanatory) variables. A linear regression model with more variables may not always perform better than the regression model with fewer variables since inclusion of extra variables leads to the increase in the variance of the total prediction (over fitting). The method of testing all possible subsets of variables is infeasible when the number of possible predictors p is large, since it requires testing $2^p - 1$ possible subsets. A common alternative in this case is to apply a stepwise algorithm. There are three types of stepwise procedures available: backward elimination, forward addition, and stepwise search. It all of these approaches, variables are added into or deleted from the model in an iterative manner, one at a time. Below we will give a short description of the each of the methods.

Backward Elimination Starting with fitting the whole model than includes all N predictors. For each predictor X the F test statistic is computed that compares the whole model with the reduced model that excludes X. Using preassigned threshold significance level, the least significant predictor being removed. The remaining model contains N-1 predictors and again the least significant predictor is identified and may be removed be examining the F test statistics and their p-values. The procedure is repeated till all p-values in the model are less than preassigned value. The resultant model is then claimed as the final model.The major problem with backward elimination is that a dropped variable has no more chance

to re-enter the model. However, a variable that has been excluded in an earlier stage may become significant after dropping other predictors. Due to specifics of our case backward elimination is not useful since the model will be instantly over fitted as soon there are the same number of linearly independent variables as there are observation.

Forward addition Forward addition works in reverse fashion with respect to backward elimination (see [9]). Starting with the simple regression model that has the only one predictor which has the biggest sample correlation in absolute value with the response variable Y, we add to the model the predictor which meets three equivalent criteria:

1. it has the highest sample partial correlation in absolute value with the response, adjusting for the the predictors in the equation already;
2. adding the variable will increase R^2 more than any other single variable;
3. the variable added would have the largest $t-$ or F-statistics of any of the variables that are not already in the model

Thus starting with a subset of size 1, and, at each step we add another variable to the model. This procedure is repeated until a stopping rule is met. The possible stopping rules are:

- Stop with a subset with predetermined size
- Stop if the F-test for each of the variables not yet entered would be less than some predetermined number (F in)
- Stop when the addition of the next predictor will make the set of predictors too close to collinear. This is called a tolerance check and is usually related to the square of the multiple correlation between the next predictor to be added and the predictor already included in the equation

The problem associated with forward addition is that once added, a variable would stay in the final model, even if it will become insignificant after including other predictors.

Stepwise Search The stepwise search method is intended to avoid the problems with both backward elimination and forward addition so that variables already in the model may be removed due to insignificance and variables excluded may be added later on when it becomes significant. The procedure itself is more similar to the forward addition algorithm. As in forward addition, the most significant variable is added to the model at each step, it its corresponding F test is significant at the level of entry. Before the next variable is added, however, the stepwise search method takes an additional look-back step to check all variables included in the current model and deletes any variable that has a p-value greater than the level of stay. Only after the necessary deletions are accomplished can the procedure move to the next step of adding another variable into the model.

2.3.2 Principal component analysis

Principal component analysis (PCA) is the method of dimension reduction [8] that works as follows:

- for the data set covariance matrix is computed
- for the covariance matrix the eigenvectors and eigenvalues are being found
- by ordering the eigenvectors according the eigenvalues starting with the larger one, the orthogonal basis is formed
- the matrix composed of eigenvectors is used to transform the original data vector , the components of the transformed vector in the orthogonal basis are called the principal components

It is usually happens that only first few components will accumulate almost all variation of the data set. Subsequently, one can either carry out the change of variables (setting new variables to be the principal components), or choose predictors to be the variables which appear with the largest coefficients in the first few principal compponents.

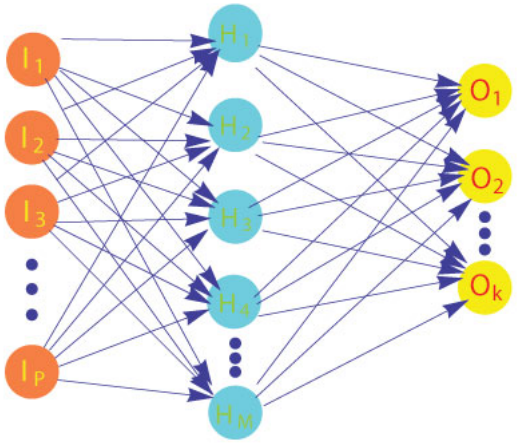

Figure 2.1: General 2 layers ANN

2.3.3 Artificial Neural Network classification

Artificial neural networks (ANN) provide a robust approach to approximating functions. The study of ANNs has been inspired in part by the observation that biological learning systems are built of very complex webs of interconnected neurons. ANNs consists of a pool of simple processing units with communicate by sending signals to each other over a large number of weighted connections (ref here) Each unit performs a simple job: receive input from neighbours or external sources and use this to compute an output signal which is propagated to other units. From the multitude of network designs we will consider the simplest one , sometimes called the single hidden layer back-propagation network. Within the neural system there are three types of units : input units which receive data from outside the neural network, output units which send data out of the neural network, and hidden units whose input and output signals remain within the neural network.

The neural network is represented by a diagram in Figure 1. There is a network for K-class classification, there are K output units[8]. Derived features H_m are created from linear combinations of the inputs I_p and the outputs are modelled as a function of linear

combinations of the H_m.

$$H_m = \alpha_{0m} + \alpha_m^T I,\ m = 1, ..., M$$

$$O_k = \sigma(\beta_{0k} + \beta_k^T H),\ k = 1, \ldots, K,$$

where $H = (H_1, H_2, ..., H_M)$, and $O = (O_1, O_2, ..., O_K)$. The output function σ is typically chosen to be *sigmoid* function $\sigma(x) = 1/(1 + e^{-x})$, identity or hyperbolic tangent function. ANN training is finding so called *weights* which will fit data the best. The complete set of weights for the network on Fig **??**. consists of:

$$\{\alpha_{pm}; \qquad p = 0, 1, \ldots, P \qquad m = 1, 2, \ldots, M\} \qquad M(P+1) \qquad \text{weights},$$

where α_{pm} is the weight for the mth input to hidden neuron p

$$\{\beta_{mk}; \qquad m = 0, 1, \ldots, M \qquad k = 1, 2, \ldots, K\} \qquad K(M+1) \qquad \text{weights}$$

where β_{mk} is the weight for the kth input to output neuron m.
We use use sum-of-squared errors (SSE) as a measure of fit:

$$R = \frac{1}{2} \sum_{k=1}^{K} (O_k - T_k)^2$$

where $\vec{T}$ is the vector of target values

The generic approach to minimizing R) is by the version of the gradient descent algorithm called *back-propagation*. The gradient can be easily derived by using simple differentiation. Upon taking derivatives, a gradient descent update at the $(r+1)$-st iteration is carried out as follows (see Fig. **??**):

- Rule for output weights:

$$\beta_{mk}^{(r+1)} = \beta_{mk}^{(r)} - \gamma_r \frac{dR}{d\beta_{mk}^{(r)}}$$

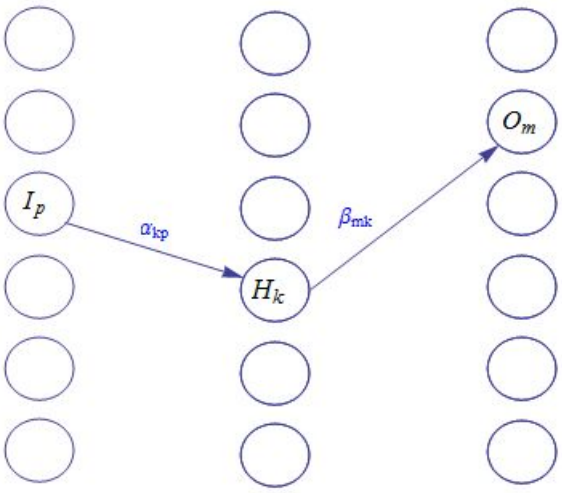

Figure 2.2: Derivation of back -propagation rules

where

$$\frac{dR}{d\beta_{mk}} = (O_k - T_k)\sigma'(\beta_{0k} + \beta_k^T H)H_m$$

- Rule for input weights:

$$\alpha_{kp}^{(r+1)} = \alpha_{kp}^{(r)} - \gamma_r \frac{dR}{d\alpha_{kp}^{(r)}}$$

where

$$\frac{dR}{d\alpha_{kp}} = \sum_{k=1}^{K}(O_k - T_k)\sigma'(\beta_{0k} + \beta_k^T H)I_p$$

Here, γ_r is the so-called *learning rate* which is usually set to be a constant, or can be optimized to minimize the error function at each update.

Using updates for the weights, the back propagation algorithm searches the space of possible hypotheses to iteratively reduce the error of the fit to the training examples in the network. Gradient descent algorithm converges to a local minimum of the training error with respect to the network weights. The advantage of the back-propagation algorithm is its simple, local nature. In the back propagation algorithm, each hidden unit passes and receives information

only to and from units that share a connection. The one of the intriguing properties of the back-propagation is its ability to invent new features that are not explicit in the input of the network. For example, the hidden layer learns to represent intermediate features that are useful for learning the target function and that are only implicit in the network inputs.

2.3.4 Validation and sensitivity analysis

In order to evaluate precision of the variable selection algorithms and subsequent classification we use Monte Carlo simulation procedures.

In particular, we randomly partition our data into two parts: the training set and the evaluation set. Data in the training set is used for variable selection and construction of a classification rule. After that, this classification rule is tested on the evaluation set and the percentage of mis-classified cases is recorded. The process is repeated many times and results are averaged. Results of this calculations allows to predict how classification algorithms will work on a new data.

In addition, we evaluate how much classification precision will be lost if biochemical characterisitcs are not available and classification is carried out entirely on the basis of epidemiological data. The latter allows to calibrate how much advantage is received by employing biomarkers for prediction of the patient's survival time.

2.4 Application to the data set

Our objective is to classify patients according to survival times, in particular, to reveal relationship between biomarkers (ER and MUC1) characteristics and survival time.

We partitioned the patients into two classes: a class containing patients whose survival was less than 4 years and another one, containing patients with more than 4 years survival time. Our analysis was based on the data for 33 patients, 13 females and 20 males, with about 60 variables per patient. The high number of covariates compared to the sample size

Table 2.2: Coding the histological type of cancer (categorical variable).

Dummy variables / **Histological type of cancer**	h1	h2	h3	h4	h5	h6
Squamous cell carcinoma nos	0	0	0	0	0	0
Adenocarcinoma nos	1	0	0	0	0	0
Large cell carcinoma nos	0	1	0	0	0	0
Carcinoma nos	0	0	1	0	0	0
Squamous cell carcinoma keratinizing nos	0	0	0	1	0	0
Bronchiolo-alveolar adenocarcinoma nos	0	0	0	0	1	0
Mucin-producing adenocarcinoma	0	0	0	0	0	1

(the total number of patients) made variable selection (dimension reduction) a matter of uttermost importance.

2.4.1 Linear regression approach

In order to successfully deal with variable selection in the presence of categorical variables, we need to record those variables as a collection of dummy variables as it is presented in the following tables. Other categorical variables were coded similarly. Coding of ordinal variables is displayed in Table 2.5. For computations, we were using software packet PASW 18 (former SPSS). Monte Carlo simulations confirm that classification results are more accurate if the data is split by gender before variable selection is carried out. Simultaneous variable selction for both genders immediately produces over-fitted solution which is of no use for prediction purposes.

Table 2.3: Coding

Dummy variables / **TNMCS/Mixed Stage**	m1	m2	m3	m4	m5	m6
1B	0	0	0	0	0	0
1A	1	0	0	0	0	0
2B	0	1	0	0	0	0
1	0	0	1	0	0	0
2A	0	0	0	1	0	0
3A	0	0	0	0	1	0
3B	0	0	0	0	0	1

Table 2.4: Coding the tumor location (categorical variable).

Dummy variables / **Tumor Location**	s1	s2	s3	s4
Lung upper lobe	0	0	0	0
Lung lower lobe	1	0	0	0
Lung nos	0	1	0	0
Lung overlapping lesion	0	0	1	0
Lung middle lobe	0	0	0	1

Table 2.5: Coding of ordinal variables

Coding	Variable	Description
N2PN	Norm2PercentNuclei	% of nuclei with intensity score 2+ in normal region (ER)
T2PN	Tum2PercentNuclei	% of nuclei with intensity score 2+ in tumor region (ER)
T3PN	Tum3PercentNuclei	% of nuclei with intensity score 3+ in tumor region (ER)
TAPI	TumAveragePositiveInt	average positive stain intensity in tumor cell nuclei (ER)
TIS	TumIntensityScore	prevalent intensity score in tumor cells (ER)
NPCO	NormPercentComplete	% of cells with membrane affected in normal region (MUC1)
N2PC	Norm2PercentCells	% of membranes with intensity score 2+ in normal cells (MUC1)
N3PC	Norm3PercentCells	% of membranes with intensity score 3+ in normal cells (MUC1)

In order to efficiently classify the patients according to their survival times and to reduce the number of variables, we used the stepwise variable selection for linear regression and the PCA approaches described in Sections 2.3.1 and 2.3.2, respectively. In regression model selection, we employed forward addition algorithm. Subsequently, we applied linear regression and the neural network-based machine learning described in Sections 2.3.1 and 2.3.3 to predict survival times.

In order to test the accuracy of variable selection and the precision of the subsequent classification, we carried out Monte Carlo simulation algorithm. In particular, we randomly divided all patients into a training set containing 75% and an evaluation set containing 25% of total data. proportion 75%/25%. At each run of the simulations (i.e., for each of the splits of the data) we evaluated percentage of the incorrectly classified patients and then averaged those percentages over all runs.

We repeated this process 100 times with female patients and 100 times with male patients and recorded the percentage of models in which each of the variables appeared (separately, for males and females). Subsequently, all variables were ranked and variables with the low rankings were eliminated from the final model. Tables 3.1 and 3.2 show the higher ranked variables for female and male patients, respectively. Table 3.4 displays the means and the variances of the regression coefficients for the variables in the models chosen over 100 runs. Those tables confirm that the coefficients are indeed have relatively low variability and, therefore, the model selection is reliable.

CHAPTER THREE: RESULTS

We have arranged this section in the same way as we did for the previous one. First, we describe the models obtained separately for male and female patients using linear regression and, after that, results obtained using the PCA and the ANN classification.

3.1 Results Obtained by Liner Regression Approach

There were 13 records for female and 20 records for male patients. We arranged training set sizes to be 10 and 15 records and evaluation sets - 3 and 5 records for female and male patients, respectively.

In order to choose variables which should appear in classification rule, variable were ranked according to the percentage of times of they appear in the suitable regression models. Selection of suitable regression models was based on the number of classification errors appearing in the training set. In particular, if the number of classification errors in the training set exceeded four, the model was discarded. Otherwise, the variables appearing in the model were recorded. Then, the probability of appearing in the model was averaged over 50 suitable models Tables 3.1 and 3.2 present the list of the variables which appear in the majority of regression models for female and male patients, respectively.

Here, $m1$, $m2$ and $s2$ are epidemiological categorical variables appearing in Tables 2.2 and 2.3, respectively, AgeDiag is the age of the patient at the moment of diagnostics, Variables N2PN, NPPN, TIC, T2PN and TAPI are biochemical characteristics associated with ER biomarker (see Table 2.5 for coding). The following table presents the set of variables which appeared in the highest percentage of models over the 50 runs.

Table 3.1: Variables appearing in the highest percentage of models for females.

Variable	Description	Percentage of models
m2	Categorical, TNMCS- Large cell carcinoma nos	44%
s2	Categorical, Tumor location - Lung Nos	28%
N2PN	ER data for normal cells	28%
NPPN	ER data for normal cells	17%
TIS	ER data for cancer cells	13%

Table 3.2: Variables appearing in the highest percentage of models for males.

Variable	Description	Percentage of models
AgeDiag	Age at diagnostic	67%
T2PN	ER data for cancer cells	56%
m1	Categorical, for Ademocarcinoma nos	30%
TAPI	ER data for cancer cells	25%
s2	Categorical,Lung nos	16%

Table 3.3: Average percentage of classification errors in the training and evaluation sets with and without biomarker ER data (over 100 runs).

	Females		Males	
	training set	evaluation set	training set	evaluation set
With ER data	5%	19%	10%	20%
Without ER data	6.3%	28%	11%	31%

Using the top three of those variables for females and five for males, we carried out 100 runs with the randomly partitioned sample, keeping 10 and 15 observations as training samples and 3 and 5 as evaluation sets for females and males, respectively, and recorded percentage of classification errors and regression coefficients. In order to assess the advantage we receive by using biochemical variable, we also constructed classification rules in the absence of ER data and evaluated average percentage of misclassified cases. Table 3.3 below presents the average percentage of classification errors in the training and evaluation sets, with and without biomarker ER data (over 100 runs). Table 3.4 reports the means and the standard deviations of the regression coefficients confirming validity of the models for female and male patients.

We have also computed the regression models based on the the whole data sets. The resulting equations for females and males, respectively, are

Table 3.4: The means and the standard deviations of the regression coefficients (over 100 runs).

Females			Males		
Variable	mean	st. deviation	Variable	mean	st. deviation
Constant	1.12	0.22	Constant	8.4	2.07
m2	-0.61	0.22	AgeDiag	-0.057	0.006
s2	-0.90	0.08	T2PN	-0.0217	0.0062
N2PN	-0.007	0.004	m1	0.547	0.108
			TAPI	-0.02	0.01
			s2	-1.01	0.18

$$\textbf{Females} : \textbf{SC} = 1.099 - 0.601\textbf{m2} - 0.919\textbf{s2} - 0.005\textbf{N2PN},$$

$$\textbf{Males} : \ \textbf{SC} = 0.812 - 0.057\textbf{AgeDiag} - 1.008\textbf{h2} + 0.551$$
$$- 0.022\textbf{T2PN} - 0.016\textbf{TAPI}, \quad (3.1)$$

Model (3.1) has 7% prediction error rate for males and 15% prediction error rate for females.

3.2 ANN classification Results

AS an alternative to linear regression technique, we used ANN as a classification technique in conjuction with PCA variable selection as a model selection methodsince PCA is known to work well together with ANN. We use PCA separately to choose ER and MUC1-related variables. Results are presented in Table 3.5 with coding of the ordinal variables displayed in Table 2.5.

Table 3.5: Most influential MUC1 and ER-related variables.

Gender	**MUC1**	**ER**
Males	N3PC; NPCO	T3PN; TIS
Females	TMA ; N2PC	TAPI ; T2PN

ANN classification has been implemented using MatLab 2007. For each gender a simple "feed forward" ANN has been built with 15 hidden layers of neurons and the number of inputs which is dependent on the number of variables. We included 22 variables for males and 17 variables for females which were found by the PCA algotithm. After that, we randomly split cases in proportion 75%-25% and carried out Monte-Carlo simulations. In order to evaluate the advantage provided by using bio-chemical variables, we also repeated ANN classification in the absense of bio-chemical data. Table 3.6 displays the means and the standard deviations of the correctly classified cases in the evaluation and training sets over 100 runs. Table 3.6 shows that the prediction accuracy increases when biochemical variables are used.

Table 3.6: Percentage of cases correctly classified by ANN averaged over 100 runs. Standard deviations are presented in paretheses.

Gender	**Variables**	**evaluation set**	**training set**
Males	with bio	78% (5%)	79% (5%)
	without bio	73% (21%)	74% (6%)
Females	with bio	83% (18%)	89% (5%)
	without bio	73% (23%)	78% (7%)

CHAPTER FOUR: DISCUSSION

The results obtained shows that however biochemical data definitely increase the accuracy of the prediction of the survival of the cancer patients but there are not enough evidence to support the hypothesis that estrogen leads to chromosomal instability and female lung cancer progression, that might be the reason for gender differences in lung cancer survival. On contrary we have more evidence supporting the point of view that estrogen affects male patients more than females - this can be concluded comparing the size of "biochemical" coefficients in regression equations (see page 13). Still, the data support the conclusion that the expression of ER receptor leads higher mortality. The more representative number of patient records needed for more precise conclusion. As for the methods ,the prediction accuracy of ANN yields to that of linear regression, but still is accurate for the number of cases we had.

APPENDIX: PASW CODE

A.1 Forward regression

This program performs random selection from the loaded dataset predefined number of records - 15 in this example and then performs forward regression for the SC variable (Survival class) over the all descriptors

```
compute scramble=uniform(1).\** the variable for permutation
sort cases by scramble.
COMPUTE temp=\$casenum.
compute selectvar = temp LE 15.
REGRESSION
  /SELECT = selectvar EQ 1
  /MISSING LISTWISE
  /STATISTICS COEFF  R
/CRITERIA=PIN(.06) POUT(.10)
   /NOORIGIN
  /DEPENDENT SC
 /METHOD=FORWARD SequenceNumber AgeatDiagnosis etc.
/SAVE pred (mypred).
COMPUTE newpred=0.
if (mypred GE .5) newpred =1.
COMPUTE newvar = ABS(newpred-SC).

compute err= 100*newvar/14.
list newpred  SC newvar err .

DElETE VARIABLES mypred newpred newvar err selectvar scramble temp.
```

A.2 Monte Carlo

This code idea borrowed from [4].

```
DEFINE !doit(nbvar=!TOKENS(1)) \* MACRO starts
* Save the regression parameters of each case in a separate file.
!DO !cnt=1 !TO !nbvar
compute scramble=uniform(1).
sort cases by scramble.
COMPUTE temp=\$casenum.
compute selectvar = temp LE 15.
REGRESSION
  /SELECT= selectvar EQ 1
  /MISSING LISTWISE\\
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.07) POUT(.10)
  /NOORIGIN
  /DEPENDENT SC
  /METHOD=ENTER  AgeatDiagnosis  s2   m1 Tum@2PercentNuclei  TumAveragePositiveIntensity
 /OUTFILE=COVB(!QUOTE(!CONCAT('C:/Users/param',!cnt,'.sav')))
/SAVE pred (mypred).
COMPUTE newpred=0.
if (mypred GE .5) newpred =1.
COMPUTE newvar = ABS(newpred-SC).
compute err= 100*newvar/20.
list newpred  SC newvar err .

DElETE VARIABLES mypred newpred newvar err selectvar scramble temp.
```

```
!DOEND \* end of MACRO

* Get all parameters in the same file; keep only the parameters estimates.
GET FILE= 'C:/Users/param1.sav'.
!DO !cnt=2 !TO !nbvar
ADD FILES FILE=* /FILE=!QUOTE(!CONCAT('C:/Users/param',!cnt,'.sav')).
!DOEND

SELECT IF RTRIM $(rowtype_)$= "EST".

* then add them to the original data file.
MATCH FILES /FILE=*
 /RENAME $(depvar_ rowtype_ varname_ = d0 d1 d2)$
 /FILE='C:/Users/mydata.sav'
 /DROP= d0 d1 d2.
EXECUTE.
!ENDDEFINE.

!doit nbvar=100.
```

APPENDIX B: MATLAB CODE

Function **trainset** used to prepare random training and testing sets of given size from the given data (**mydata**)

```
function [traindata,simdata,train_truth,sim_truth] = trainset (mydata,percent)
% define what is true value
[ vsego, compon]=size$(mydata);
skolko=round(percent*vsego);
vibor=randperm$(vsego)$;
truth=mydata(:,compon);
mydata=mydata(:,1:compon-1);
for i = 1:vsego
    if(i<=skolko)
        traindata(i,:)=mydata(vibor(i),:);
        train-truth$(i)$=truth$(vibor(i))$;
    else\\
        simdata$(i-skolko,:)$=mydata$(vibor(i),:)$;
        sim-truth$(i-skolko)$=truth$(vibor(i))$;
    end;
end;
```

Function **start** used to train and simulate network for gives training set of data and report the results

```
function [acc_tr,acc_test] =  start$(network,data, n_attemp,percent)
for i=1:n-attemp
    [st,ss,tt,ts]$=trainset$(data,percent);
  train(network,st',tt);
  a=sim(network,ss');
b=sim(network,st');
```

```
  [attemp,acc_test(i)]= pred(a,ts);
  [attemp1,acc_tr(i)]= pred(b,ts);
end;

[acc1,acc2]$=start$( network1,fornet, 100,.8);
```

LIST OF REFERENCES

[1] D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M. Taylor, M.D. Iannettoni, M.B. Orringer, and S. Hanashand. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, 8:816–824, 2002.

[2] Jürgen Groß. *Linear Regression*. Springer, 2003.

[3] Nikolai N. Khodarev1, Sean P. Pitroda1, Michael A. Beckett1, Dhara M. MacDermed1, Lei Huang2, Donald W. Kufe2, and Ralph R. Weichselbaum1. Muc1-induced transcriptional programs associated with tumorigenesis predict outcome in breast and lung cancer. *Cancer Research*, 69, 2009.

[4] Raynald Levesque. Raynald's spss tools. http://pages.infinit.net/rlevesqu/SampleSyntax.htm#MultipleResp, October 2004.

[5] D. C. Mrquez-Garbn and R. J. Pietras. Estrogen-signaling pathways in lung cancer. *Advances in Experimental Medicine and Biology*, 617:281–289, 2008.

[6] H. I. Pass, D. P. Carbone, D.H. Johnson, J. D. Minna, G. V. Scagliotti, and A. T. Turrisi. *Principles and Practice of Lung Cancer: The Official Reference Text of the International Association for the Study of Lung Cancer (IASLC)*. Springer, 2010.

[7] Eva Szabo. Muc1 expression in lung cancer. *Methods in Molecular Medicine*, 74(3):251–258, 2003.

[8] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* , 2001.

[9] Sanford Weisberg. *Applied Linear Regression*. Wiley, New York, 2 edition, 1985.

[10] Yongchun Zhou, Hasan Rajabi, and Donald Kufe. Mucin 1 c-terminal subunit oncoprotein is a target for small-molecule inhibitors. *Molecular Pharmacology*, 79(5):886–893, 2011.

[11] T. Zhukov. Personal communication. .