January 2012

A Multifaceted Analysis of Early Stage Non-Small Cell Lung Cancer Data

Madhusmita Behera
University of South Florida, madhusmita19@gmail.com

A Multifaceted Analysis of Early Stage Non-Small Cell Lung Cancer Data

by

Madhusmita Behera

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Chemical and Biomedical Engineering
College of Engineering
University of South Florida

Major Professor: John J. Heine, Ph.D.
William E. Lee, Ph.D.
Maria Kallergi, Ph.D.
Steven Eschrich, Ph.D.
Srinivas Katkoori, Ph.D.

Date of Approval:
February 24, 2012

Keywords: Survival Analysis, Differential Evolution, TMA, DNA Repair, Statistical Learning

Acknowledgements

First and foremost, I would like to acknowledge my advisor and mentor, Dr. John Heine of the Moffitt Cancer Center & Research Institute. This work would not have been possible without his unfailing support and invaluable guidance all throughout. He never allowed me to give up. He has always been a constant source of inspiration and motivation during this journey. I really have no words to thank him enough. My heartfelt thanks to Dr. Maria Kallergi for believing in me and giving me my first break. She gave me the opportunity to work with her and to be part of the wonderful group she was leading. I am truly grateful to her for opening the door for me. I thank the committee members for their guidance and for reviewing my work. This support is truly appreciated. I would like to thank Ms. Erin Fowler of the Moffitt Cancer Center & Research Institute for her assistance in developing the programs and analytical methods for this work. Lastly, I would like to acknowledge and thank Dr. Suresh Ramalingam of Emory University for his support in many aspects of this work.

Table of Contents

List of Tables

List of Figures

Abstract

Some early stage NSCLC patients have a better survival prospects than others. In any event, the long-term prognosis for NSCLC patients is poor. Various measures were investigated to gain a better understanding of those patient characteristics that confer better survival or predict disease recurrence. A dataset comprised of stage 1 NSCLC patients (n=162) that underwent resection was investigated. Clinical variables (CVs) and tissue microarray (TMA) images with DNA repair protein and standard H&E expressions were investigated. Patients were dichotomized into two groups by survival characteristics and logistic regression (LR) modeling was used to predict favorable survival outcome. Various patient strata were investigated with Cox regression and Kaplan Meier survival analysis (i.e. accepted survival analysis methods). A statistical learning (SL) method comprised of a kernel mapping and Differential Evolution optimization was developed to integrate SL techniques with LR and accepted survival analysis methods by first combining various patient measures to form a hybrid variable. Younger age, female gender, and adenocarcinoma subtype confer better survival prospects, whereas recurrence confers poor survivability. The SL hybrid modeling produced greater favorable outcome associations and survival hazard relationships than the accepted approaches. Automated texture measures from the HE stained TMA images were significantly related to survival, tumor-type, and tumor-grade. DNA repair measures in isolation or in combination with CVs were not related to survival, favorable outcome or recurrence, and none of the CVs were related to recurrence.

A platform was established to incorporate automated TMA analysis and SL techniques into standard epidemiologic practice, and baseline predictive models were constructed. Future work will investigate novel biomarkers and larger datasets using this established framework to construct prognostic models for clinical applications for lung cancer patients in general and to better understand disease recurrence.

Chapter 1: Introduction

Lung cancer is the leading cause of cancer related mortality in the United States as well as globally [1-3]. Primarily, there are two types of lung cancer: non-small cell lung cancer (NSCLC), which accounts for about 80% of the cases [3], and small cell lung cancer (SCLC), which accounts for about 15% of the cancers [4]. Smoking is the leading risk factor for the development of lung cancer, and about 85% of lung cancer deaths are attributed to smoking [5]. Non-small cell lung cancer originates from the epithelial cells of the lung of the central bronchi to terminal alveoli. The common histological subtypes of NSCLC are (a) adenocarcinoma (AC), which represents about 40 % of the cases, (b) squamous cell carcinoma (SCC), which accounts for about 39% of the cases, and (c) large cell carcinoma, accounting for about 15% of the cases [6]. Although the NSCLC subtypes differ in cell size, shape, and chemical makeup, they are categorized as a monolithic group because they are treated similarly and have a similar prognosis.

The prognosis for lung cancer patients is generally poor. The five-year survival rate for NSCLC patients with stage IV disease is as low as 1% (see Table 1). Improvements in therapeutic modalities have resulted in modest improvements in outcome for patients in the past two decades. Related work by Behera et al and group shows the efficacy of certain treatments for patients with NSCLC in different settings [7-9]. In parallel, we have shown that SCLC patients that did not respond to frontline chemotherapy also responded poorly with second-line treatments [10]. Staging of the cancer is significant for determining suitable treatment, and patients with early stage cancer can benefit from surgical resection [3]. However, a cure remains elusive for patients with advanced stage disease and as well as for the majority of stage II and III patients [1, 6].

Lung cancer is often diagnosed at an advanced stage, largely due to the lack of effective modalities for early detection [2, 11]. Recent evidence from the National Lung Screening Trial shows that low-dose computed tomography (CT) scans can reduce lung cancer mortality in comparison with single-view chest radiography when screening high-risk patients [6]. Volumetric datasets and high resolution enable helical CT to better detect early stage cancers than chest radiography [12, 13]. Before this promising approach is incorporated into clinical practice, several important clinical issues must be addressed [6, 11].

For patients with early stage lung cancer, local therapy with surgical resection is associated with the best survival outcomes. This best case scenario is limited to those with NSCLC, which accounts for approximately 85% of all cases of lung cancer in the United States. Despite optimal surgical resection, recurrence of disease is noted in 30-75 percent of the patients with early stage disease. The development of prognostic models for predicting survival outcomes for patients with NSCLC after resection may have important healthcare implications [14]. Even in patients with early stage lung cancer, there is a critical need to improve cure rates, identify patients at higher risk for recurrence, and identify those patients that have better chances of survival. Previous work [2, 15] showed that early stage at diagnosis, younger age, and female gender are favorable prognostic indicators for NSCLC patients. It is also important to note that the incidence rates for the various forms of lung cancer appear to be shifting in time [16], and that there are both racial and regional differences throughout the United States [17]. Both serial and geographical variations in lung-cancer survival patterns indicate that survival rates require continual evaluation to ensure the knowledge-base is current.

The survival rates shown in Table 1 were obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database based on people who were diagnosed with NSCLC between 1998 and 2000 [18]. These clearly show that early stage patients have a survival advantage, but the 5 year survival probability cannot be considered favorable.

Table 1: Five year survival rate across stages

Stage	**IA**	**IB**	**IIA**	**IIB**	**IIIA**	**IIIB**	**IV**
5- year survival rate	49%	45%	30%	31%	14%	5%	1%

Understanding the patient attributes that enhance longer-term survival probability or indicate recurrence is necessary to individualize treatment options. In this work, various aspects of stage I NSCLC were investigated. Clinical, pathological, and image measures, derived from tissue microarrays (TMAs), were analyzed. The dataset is comprised of stage 1 NSCLC patients (n=162) that underwent resection at the WellStar Kennestone Hospital, GA, from 2002-2008.   These patients were selected retrospectively and consecutively. This data was collected under an approved protocol by the Western Institutional Review Board (WCR20080401; approval # 20081986). This work is presented in four parts corresponding to chapters 2-5:

(i) Analyses of survival and recurrence using readily available clinical and pathological data.

(ii) Study of DNA-repair pathway expression variables.

(iii) Evaluation of statistical learning (SL) methods for survival analysis.

(iv) Automated measurements and analyses of multispectral tissue microarray (TMA) image data.

Part i was used to form baseline survival attributes for comparisons using readily available variables. The patient population and measures are described in Part i except for the tissue microarray (TMA) data and novel biomarker measures. This population sample was used for all of the subsequent investigations discussed in this dissertation. In Part ii, experimental biomarkers derived from novel protein stains were used to assess whether DNA repair can add to the findings in Part i [14]. Part iii represents an exploration to develop and evaluate a method for adapting statistical learning (SL) methods for accepted epidemiologic analyses. In this approach, SL methods were used as a pre-processing step to prepare the data for use in logistic regression and survival analysis (Cox regression and Kaplan Meier analysis) [19]. This essentially combines the strength of SL with these important epidemiologic analysis methods. In Part iv, a system was developed to analyze TMA images automatically. TMA images were stained with standard techniques and assessed with image processing methods to evaluate whether there is additional information that is not captured by human observation. Although we used low resolution data for this analysis due to technical difficulties, the methods are scalable to higher resolution.

Chapter 2: Survival Analysis of Stage I Non-Small Cell Lung Cancer

Readily available clinical variables and pathologic features from stage I NSCLC patients were analyzed [14]. Two forms of analysis were applied to evaluate the survival characteristics of this population. Logistic regression (LR) modeling was used to study two groups of patients dichotomized by their survival characteristics to form favorable and unfavorable survival outcome groups. Various models were studied to predict favorable outcome. Survival analysis (i.e. Cox regression and Kaplan-Meier analysis) was also used to study various patient strata. The full dataset as well as subsets of data dictated by full case ascertainment for the variables under consideration were used for this work. The NSCLC patient data described below was used exclusively for this work and will not be redefined in subsequent chapters. The evaluation described in this chapter was essentially excerpted from the work by Behera et al [14].

2.1 Study Population and Measures

The dataset is comprised of patients (n=162) with stage I NSCLC that underwent surgical resection at the WellStar Kennestone Hospital, GA, from 2002-2008.   These patients were selected retrospectively and consecutively. The selection criteria included all stage I patients that had complete case ascertainment for the variables under consideration. One hundred and one (n_1) of these patients were alive at last contact (censored), and 61 (n_2) patients died (incident) during the course of the contact interval. The clinical and pathological variables abstracted from the patient files included age (i.e. age of the patient at the time of surgery) with integer accuracy, gender (binary), smoking status (binary), four histological-subtypes [i.e. AC, SCC, LCC, and adenosquamous carcinoma (ASC)], tumor-grade, adjuvant treatment, disease recurrence, and tumor-location within the lung. Stage I was dichotomized as IA and IB subgroups (categorical binary variable) as ascertained from the pathology reports. Past or current smokers (i.e. smoking status) were categorized as either a smoker, past or present (yes), or as those that never smoked (no). Tumor was graded with a 1-3 integer scale (tumor-grade) describing the cancer cell differentiation (a measure of abnormality) derived from pathology reports (i.e. grade 1 implies well differentiated cells resembling normal cells and grade 3 implies poorly differentiated cells indicating abnormality). Adjuvant treatment was defined as systemic chemotherapy given to the patients after surgical removal of the tumor. Recurrence indicates the relapse of the disease after surgery. Tumor-location was defined primarily with four categories: lower lobe, middle lobe, upper lobe, and upper/lower lobe. This database (Winship Cancer Institute of Emory University Lung Cancer Database) was constructed and managed by the author [20] over the past three years (2009-2012) and is still under development. This database is a web-based archival management system, designed and developed to store clinical and pathological information of lung cancer cases [20]. Data from this system can be exported to SAS and MS-Excel. The statistical analysis was performed with the SAS software package (SAS Institute, NC) and PASW Statistics 18.0.0 (SPSS Inc).

Two forms of survival analysis were used below. First, the patients were dichotomized into two groups based on their survival outcomes. Logistic regression (LR) was used to investigate these two groups using the variables discussed above. Secondly, various patient strata were investigated with Cox-regression and Kaplan Meier survival analysis The rationale for using two forms of survival analysis is that they provide different endpoints (discussed below in the Modeling Strategies Section).

2.2 Analysis Methods

2.2.1 Favorable Outcome Analysis

Censored (n_1) and incident (n_2) patients were used to form favorable and unfavorable survival outcome groups, respectively. The LR modeling was referenced to the favorable outcome group (i.e. to predict the probability of a given patient experiencing a favorable survival outcome given a specific set of variables). Complete case-ascertainment for all the variables for the entire patient population was not available. The full dataset (full group) and various subgroups of this dataset were studied depending upon the case-ascertainment for the variables under investigation. The goal was to find those variables related to favorable survival outcome and characterize their association strengths. Another aim was to find the collection of variables that provided the greatest discrimination (i.e. predictive capability) between these groups. Odds ratios (ORs) were used to assess group associations and the area under the receiver operating characteristic curve (Az) was used to measure model predictive capability. The ORs are cited with 95% confidence intervals (CIs). In this portion of the analysis, Az was computed using standard SAS routine [i.e. assessing the range of (false positive, sensitivity) ordered pairs and performing integration with the trapezoid rule]. To avoid over fitting and user imposition, interaction terms within the LR model were not considered in the favorable outcome modeling. Alternatively, variable interactions were investigated as endpoints below. The justification for the favorable outcome modeling and the dichotomization strategy are discussed in more detail in subsequent sections and chapters. Briefly, this form of LR modeling has a different interpretation than the time-to-event methods, and the dichotomization technique was necessitated by the limited dataset.

2.2.2 Inter-Variable Association Analysis

Various combinations of variables were used to evaluate possible associations with the AC and SCC histology subtypes, gender, and disease recurrence. In this modeling, histology, gender, and recurrence were used as the dependent variables for LR. The LR models with the following independent (or input) variables were investigated: age (A), tumor-grade (Gr), and gender (G). The following relationships were investigated to predict the two class histology subtypes (i.e. predict SCC): LR(A, Gr), LR(A, G), and LR(A, Gr, G). A similar analysis was performed to predict male gender and disease recurrence. The independent variables used in the LR model to predict male gender and recurrence can be determined from the histology subtype analysis by replacing SCC with male gender and disease recurrence, respectively.

2.2.3 Survival Analysis

Kaplan-Meier survival probability analysis was applied to evaluate survival differences between various patient strata. Hazard ratios (HRs) estimated with Cox regression were used to assess group survival characteristics with 95% CIs. To study age-related survival, the patients were dichotomized using the population median age (i.e. 67 years) as the cut-point. The below median age group was used as the reference. The patient population was also dichotomized by two-group histology subtype (i.e. SCC and AC), recurrence, adjuvant treatment, and gender; the respective references were AC, no-recurrence, no-adjuvant treatment, and female gender. Additionally, stage I subgroups were investigated in various strata (i) IA and IB using IA as the reference using the full group dataset, (ii) lower age-group patients with stage IA histology as the reference compared to the remaining patients in the full group dataset, (iii) all patients with stage IA with AC as the reference compared to those remaining patients in the full group, and (iv) lower-age group patients with both stage IA and AC as the reference compared to remaining patients in the full group.

2.2.4 Modeling Strategies

These two forms of modeling (i.e. time-to-event and LR favorable outcome analysis) convey different information to both the patient and the clinician. Cox regression is not typically used to make point estimates at the patient level [21] but can provide an instantaneous relative risk given a set of covariates for a given patient. Developing methods derived from Cox regression for point estimates is an active field of research [21]. Kaplan-Meier analysis is non-parametric and not useful for point estimates. Reducing the data resolution to a binary outcome makes the dataset amenable to both LR modeling for point estimates specifically and more generally to all forms of binary classification applications. The LR model provides the probability of a pre-defined endpoint given a set of specific covariates and thus gives an output that is easily interpretable at the patient level. More generally, converting survival data to a binary outcome (i.e. for classification purpose) is an accepted approach in survival prognosis predication [22]. Our approach [14, 19] to survival analysis has a similar prognostic aim and it also serves as a simplifying mechanism for accepted time-to-event analysis, as demonstrated below. Thus, the dichotomized analysis and the accepted survival analysis methods (i.e. Cox regression and Kaplan-Meier analysis) are complementary. We also developed a non-standard dichotomization method for the binary modeling. Often such dichotomization is based on a pre-defined survival time cut-point [22], not incident-censored group status and is discussed in detail below.

2.3 Results

2.3.1 Patient Characteristics

Table 2 shows the patient characteristics summarized by combined population, incident group, and censored group. The median age of the patients included in this analysis was 67 years. There was a near equal representation of male and females in total. The censored patients were younger, more likely to have tumor-grade 1 than those in the incident group, whereas the other grades were similar across the groups. The censored patients were more likely female and to have AC rather than SCC histology. Although only a small number of patients had either ASC or LCC histology-subtypes, those in the censored group were more likely to have LCC than those in the incident group. Smoking status was similar across the groups, and approximately 20% of the patients were non-smokers. The incident patients were more likely to have experienced disease recurrence and received adjuvant treatment. Tumor-location (upper lobe, lower lobe, middle lobe, upper/lower, upper/middle, lower/middle, chest wall, main stem bronchus) was similar across the groups. The censored patients were more likely to have stage IA disease in comparison with the incident patients. For the favorable group, the censored time distribution mean and standard deviation (SD) were 3.94 and 1.3 years, respectively. For the unfavorable group, the overall survival time distribution mean and SD were 2.19 and 1.8 years, respectively (not shown in the table). The separation between the group means provides the justification for the stratification method. The incident group patients are more likely positive for recurrence, but roughly 64% of these patients are negative for recurrence. It also follows that the recurrence status for most of the censored group patients is unknown. This suggests that the recurrence variable in this work could be qualified more accurately as *early recurrence*.

Table 2: Patient characteristics. This table provides the patient characteristics (Char) for the incident group (I), censored group (C), and combined total (Tot). The number of samples (n), mean values (Mean), standard deviation (SD) and percentages (%) are provided for each characteristic where applicable. The C and I labels correspond to the favorable and unfavorable outcome groups, respectively.

Char	**I**		**C**		**Tot**	
	N	**Mean / SD or %**	**N**	**Mean / SD or %**	**n**	**Mean / SD or %**
Age	61	69.6 / 7.66	101	65.7 / 8.6	162	67.2 / 8.4
Tumor-Grade	61	2.23 / 0.62	101	2.10 / 0.69	162	2.15 / 0.66
One	6	9.84%	19	18.81%	25	15.43%
Two	35	57.38%	53	52.48%	88	54.32%
Three	20	32.79%	29	28.71%	49	30.25%
Gender						
Male	39	63.93%	39	38.61%	78	48.15%
Female	22	36.07%	62	61.39%	84	51.85%

Table 2 (Continued)

Tumor-type						
Adenocarcinoma	30	49.18%	63	62.38%	93	57.41%
Adenosquamous	2	3.28%	2	1.98%	4	2.47%
Large Cell	2	3.28%	11	10.89%	13	8.02%
Squamous	26	42.62%	22	21.78%	48	29.63%
Unknown	1	1.64%	3	2.97%	4	2.47%
Smoking						
Non-Smoker	12	19.67%	19	18.81%	31	19.14%
Smoker	47	77.05%	74	73.27%	121	74.69%
Unknown	2	3.28%	8	7.92%	10	6.17%
Recurrence						
Yes	20	32.79%	6	5.94%	26	16.05%
No	39	63.93%	93	92.08%	132	81.48%
Unknown	2	3.28%	2	1.98%	4	2.47%
Tumor-Location						
Lower Lobe	18	29.51%	26	25.74%	44	27.16%
Middle Lobe	5	8.20%	5	4.95%	10	6.17%
Upper Lobe	36	59.02%	63	62.38%	99	61.11%
Upper/Lower Lobes	0	0.00%	2	1.98%	2	1.23%
Upper/Middle Lobes	0	0.00%	1	0.99%	1	0.62%
Chest Wall	0	0.00%	1	0.99%	1	0.62%
Main Stem Bronchus	0	0.00%	1	0.99%	1	0.62%
Unknown	2	3.28%	2	1.98%	4	2.47%

Table 2 (Continued)

Treatment						
Yes	10	16.39%	9	8.91%	19	11.73%
No	51	83.61%	92	91.09%	143	88.27%
Stage I						
A	37	60.66%	73	72.28%	110	67.90%
B	24	39.34%	28	27.72%	52	32.10%

2.3.2 Favorable Outcome Analysis

For the full group dataset (n=162, with n_1=101 and n_2 = 61), complete case ascertainment for age, gender, adjuvant treatment, tumor-grade, stage I subgroup was available. The forward stepwise selection procedure used with LR resulted in a bivariate model. As shown in Table 3, the ORs for age and gender were significant (i.e. the CIs do not include unity). When adjusting for gender, the age association [OR = 0.64 per standard deviation (SD) increase] and gender association [OR = 0.39] show increasing age and male gender confer an unfavorable survival outcome (i.e. females are 2.6 times more likely to be in the favorable group and younger age patients are 1.5 times more likely to be in the favorable group). In this age-gender model, Az = 0.683. Adjuvant treatment, tumor-grade, and stage I subgroup measures were not significant independent predictive factors (i.e. Az < 0.600) and their OR associations were not significant. We estimated the standard error (SE) in Az was 0.035.

Table 3: Favorable outcome predictions and associations. The full group modeling used to predict favorable outcome included: age (A), gender (G), adjuvant treatment (T) noted as ‘Treat’ in the covariates (Cov) column, tumor-grade (Gr), stage I subgroup (S1). The step-forward selection was used to build a model to predict the censored group (C). Only age and gender were significant. The odds ratios (ORs) with 95% confidence intervals are provided for each covariate, and the area under the receiver operating characteristic curve (Az) is provided for the various arrangements. Non-applicable (NA) entries are labeled. The functional notation LR(x,y) was used to indicate the variable(s) within the logistic regression (LR) model used to predict the censored group.

Model	**Cov**	**Unit / Ref**	**Versus**	**Cov OR**	**Az**
LR(A) = censored group	Age	8.4	NA	0.61 (0.43, 0.87)	0.630
LR(G) = censored group	Gender	Female	Male	0.36 (0.18, 0.69)	0.627
LR(T) = censored group	Treat	No	Yes	0.50 (0.19, 1.31)	0.537
LR(Gr)= censored group	Grade	1.0	NA	0.74 (0.45, 1.20)	0.549
LR(S1) = censored group	Stage	IA	IB	0.59 (0.30, 1.16)	0.558
LR(A,G)= censored group	Age Gender	8.4 Female	NA Male	0.64 (0.44, 0.91) 0.39 (0.20, 0.75)	0.683

In the subgroup-1 dataset (n =149, with n_1 =91 and n_2= 58), patients that had complete ascertainment for age, gender, tumor-grade, tumor-location, and histology-subtype were included to predict favorable outcome. This modeling considered all of the measured tumor characteristics in combination with age. The univariate analysis found significant associations for age [OR= 0.65 (CI: 0.45, 0.92)] per standard deviation (SD) increase and gender [OR=0.32 (CI: 0.16, 0.64)] per unit increase, indicating increasing age and male gender confer an unfavorable outcome. The forward selection process resulted in a bivariate model with age and gender, which had similar ORs as the univariate models and Az = 0.680. None of the tumor characteristics, including stage I subgroup, demonstrated significance (findings not shown).

The subgroup-2 dataset (n= 134 with n_1= 80 and n_2= 54) was investigated to predict favorable outcome by restricting the analysis to patients that had full case ascertainment for the SCC and AC histology-subtypes in conjunction with age, gender, tumor-location, adjuvant treatment, tumor-grade, and stage I subgroup. The univariate analysis found significant findings for gender [OR = 0.34 (CI: 0.17, 0.70), Az = 0.630] and histology subtype [OR= 0.44 (CI: 0.21, 0.91) and Az=0.594]. The forward stepwise procedure found the corresponding bivariate model with gender and histology subtype. In this model (i) the ORs were similar to that of the respective univariate models, (ii) the combined Az increased to 0.668, and (iii) the OR for histology was not significant. The findings show that both AC and female gender confer a favorable survival outcome, and these two variables in combination provided an increased Az in comparison with either in isolation. The other variables were not significant (data not shown). It is important to note that age was not included in the selection process. The reasons for this were investigated below in the interaction analysis.

Different combinations of variables were investigated to determine models with increased predictive capability for favorable outcome using the forward stepwise procedure. This subgroup is referred to as the best model dataset (n= 123 with n_1=71 and n_2=52). This evaluation included patients with complete case ascertainment for age, gender, adjuvant treatment, tumor-grade, stage subgroup, SCC and AC histology-subtypes, tumor-location, smoking status, and recurrence. The model with the greatest predictive capability included gender, histology, and recurrence, which resulted in a combined Az of 0.788. In this model the associations for gender [OR = 0.32 (CI: 0.14, 0.76)], histology subtype [OR = 0.41 (CI: 0.17, 0.98)], and recurrence [OR = 0.04 (CI: 0.01, 0.20)] were significant and all conferred an unfavorable outcome. The respective ORs do not vary much from their respective univariate values (see Table 3) indicating they provide *independent* contributions. Age was then forced into the selected model. Although the OR association for age was not significant, its contribution increased the model's predictive capability giving Az = 0.796. Although recurrence was a strong indicator of limited survival, the variable has limited application in general predictive modeling in that over 63% of patients in the unfavorable group were negative for recurrence as was over 94% of those patients in the censored group (i.e. not known). However, understanding the variables related to recurrence is important because given recurrence, poor survival is likely.

2.3.3 Interaction Analysis

To understand possible interactions between age and other variables, inter-measurement analyses were performed using the subgroup 2 patients. This analysis was portioned into three outcomes by considering those variables that could predict (i) the two-subgroup histology [i.e. AC or SCC], (ii) gender, and (iii) recurrence. To limit the presentation, only models that provided an Az equal to or above 0.600 are shown when the corresponding ORs were not significant. The Az quantities were calculated with SAS as described in Section 2.2. The histology relationships are shown in Table 4. Although there are many models that provided some predictive capability, the majority of the OR associations were not significant. In all models that contained age, increasing age was associated with SCC. The association for age in isolation was [OR = 1.70 (CI: 1.14, 2.52) per SD increase] with Az = 0.637. The gender associations are shown in Table 5. Although tumor-grade in isolation provided weak predictive capability (Az = 0.596), its association [OR = 1.84 per unit increase] indicates increasing grade is significantly related to male gender; in models that included grade, similar associations were found (i.e. 1.46 - 2.01 range of ORs). To understand possible interactions with disease recurrence, several LR models were investigated to predict recurrence. No significant OR relationships or predictors of recurrence were found (data not shown).

Table 4: Interaction analysis to predict histology. In this analysis we use the subgroup two dataset to predict two class histology: adenocarcinoma (AC) and squamous cell carcinoma (SCC). This model included age (A), gender (G), and tumor-grade (Gr). We used the functional notation LR(x,y) to indicate the variable(s) within the logistic regression model to predict SCC. Odds ratios (ORs) are provided with 95% confidence intervals parenthetically. The area under the receiver operating characteristic curve (Az) is provided for each model. The unit and reference (ref) and covariate (Cov) ORs are also provided for each model. Non-applicable (NA) entries are labeled.

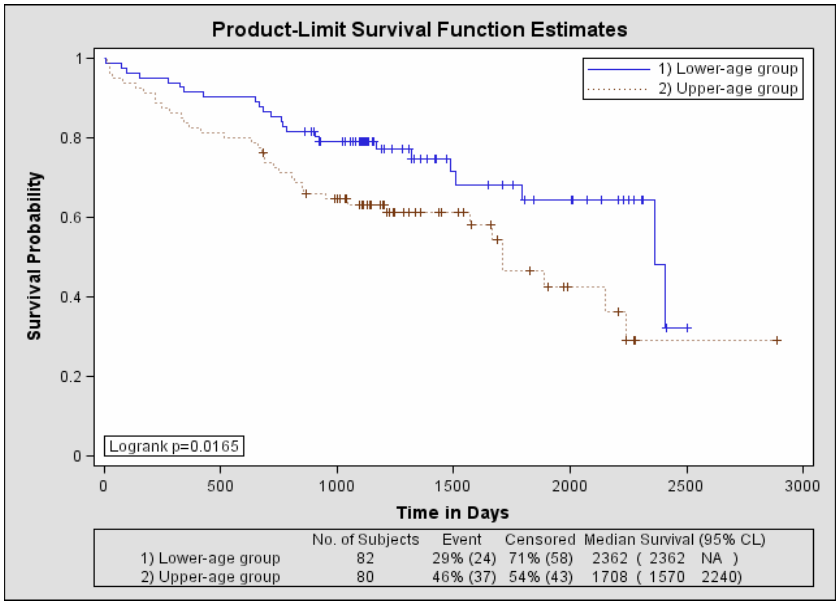
Model	Cov	Unit/Ref	Versus	Cov OR	Az
LR(A)=SCC	Age	8.0	NA	1.70 (1.14, 2.52)	0.637
LR(A, Gr)=SCC	Age	8.0	NA	1.76 (1.18, 2.65)	0.651
	Grade	1	NA	1.43 (0.79, 2.58)	
LR(A, G)=SCC	Age	8.0	NA	1.65 (1.11, 2.46)	0.660
	Gender	Female	Male	1.71 (0.82, 3.58)	
LR(A, Gr, G)=SCC	Age	8.0	NA	1.70 (1.13, 2.56)	0.665
	Grade	1	NA	1.32 (0.72, 2.42)	
	Gender	Female	Male	1.58 (0.74, 3.39)	

Table 5: Interaction analysis to predict male gender. In this analysis we use the subgroup-2 dataset to predict gender. This model included age (A), tumor-grade (Gr), and histology (H), including adenocarcinoma (AC) and squamous cell carcinoma (SCC). We use the functional notation LR(x,y) to indicate the variable(s) within the logistic regression model used to predict male gender. Odds ratios (ORs) are provided with 95% confidence intervals parenthetically. The area under the receiver operating characteristic curve (Az) is provided for each model. The unit, reference (ref) and covariate (Cov) ORs are also provided for each model. Non-applicable (NA) entries are labeled.

Model	Cov	Unit/Ref	Vs	Cov OR	Az
LR(A)=SCC	Age	8.0	NA	1.70 (1.14, 2.52)	0.637
LR(A, Gr)=SCC	Age	8.0	NA	1.76 (1.18, 2.65)	0.651
	Grade	1	NA	1.43 (0.79, 2.58)	
LR(A, G)=SCC	Age	8.0	NA	1.65 (1.11, 2.46)	0.660
	Gender	Female	Male	1.71 (0.82, 3.58)	
LR(A, Gr,G)= SCC	Age	8.0	NA	1.70 (1.13, 2.56)	0.665
	Grade	1	NA	1.32 (0.72, 2.42)	
	Gender	Female	Male	1.58 (0.74, 3.39)	

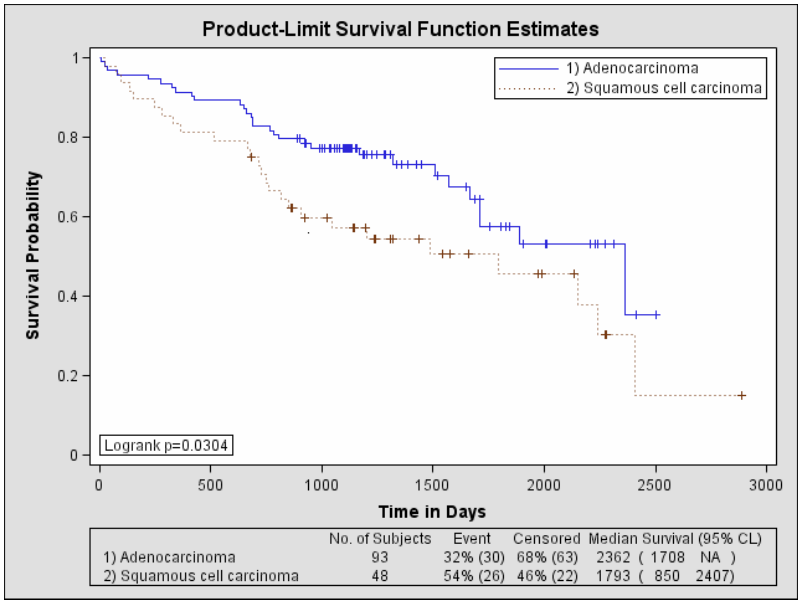
2.3.4 Survival Analysis

The survival analysis statistical test results, HRs, and survival probabilities are provided in Table 6. Figure 1 shows the dichotomous age grouping survival curves. The upper-age group is at significantly greater hazard compared with the lower-age group [HR=1.86, (CI: 1.11, 3.12)]. This table also provides proportional estimates for those surviving past 3, 5, and 7 years. This shows that 64% of the lower-age group survived past five years, whereas 47% of the upper-age group survived past this time. Controlling for grade in the age hazard model was not significant [HR=1.93, (CI: 1.11, 3.12)]. Figure 2 shows the survival curves for patients with SCC compared to patients with AC histology subtypes. Patients with SCC are at a significantly increased hazard [HR = 1.78, (CI: 1.05, 3.01)]. Over 35% of the patients with AC survived past 7 years, whereas only 15% of the patients with SCC survived past this time. Controlling for grade with histology subtype somewhat confounded the hazard relationship [HR= 1.68, (CI: 0.99, 2. 28)], but the change was not significant. The elevated hazard for disease recurrence [HR = 4.16, (CI: 2.37,7.31)] significantly limits survival (Figure 3) but note the limited number of positive recurrence patients. As shown in Table 6, approximately 38% of the non-recurrence patients survived past 7 years whereas none of the recurrence patients survived past this time. Although the curves indicate adjuvant treatment limits survival (shown in Figure 4), the findings [HR = 1.82, (CI: 0.92, 3.61)] were not significant (i.e. considered as a trend). Figure 5 shows the gender stratified curves indicating males have an elevated hazard [HR = 2.03, (CI: 1.20,3.43)] relative to females. However, the favorable survival characteristic for females is present mainly for the short and mid-term. Past seven years, survival appears similar for both genders (i.e. about 26% males and 26% of females survived).



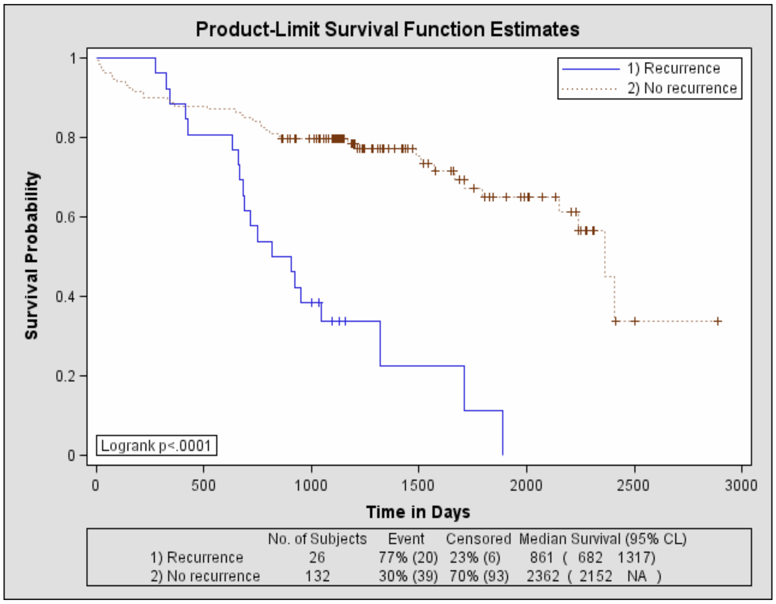
	No. of Subjects	Event	Censored	Median Survival (95% CL)
1) Lower-age group	82	29% (24)	71% (58)	2362 (2362 NA)
2) Upper-age group	80	46% (37)	54% (43)	1708 (1570 2240)

Figure 1: Age survival. This shows the survival probability curves for the full (full group) dataset. The lower-age group (upper blue curve) and upper-age group (lower brown curve) were formed by using the median age as the cut-point. The lower-age group has a survival advantage [HR = 1.86].



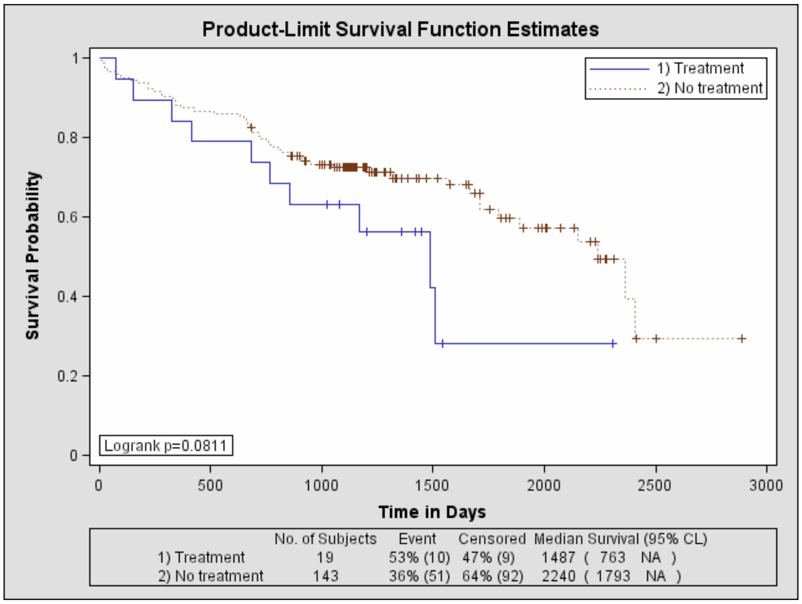
	No. of Subjects	Event	Censored	Median Survival (95% CL)
1) Adenocarcinoma	93	32% (30)	68% (63)	2362 (1708 NA)
2) Squamous cell carcinoma	48	54% (26)	46% (22)	1793 (850 2407)

Figure 2: Histology subtype survival. This shows the survival probability curves for patients with either adenocarcinoma (upper blue curve) or squamous cell carcinoma (lower brown curve) histology-types (subgroup-2). The patients with adenocarcinoma clearly have a survival advantage [HR=1.7].



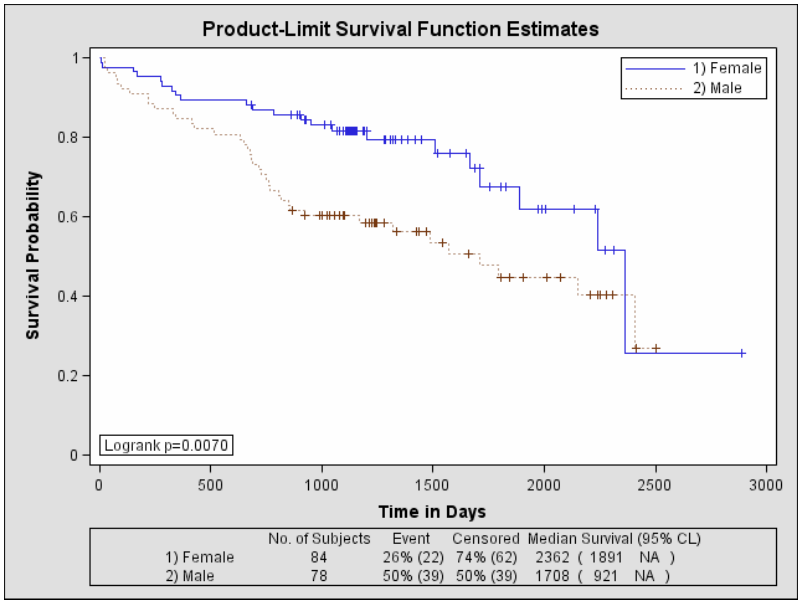
	No. of Subjects	Event	Censored	Median Survival (95% CL)
1) Recurrence	26	77% (20)	23% (6)	861 (682 1317)
2) No recurrence	132	30% (39)	70% (93)	2362 (2152 NA)

Figure 3: Disease recurrence survival. This shows survival probability curves for patients with disease recurrence (lower blue curve) and without recurrence (upper brown curve) for the full group. Disease recurrence limits survival [HR= 4.16].



	No. of Subjects	Event	Censored	Median Survival (95% CL)
1) Treatment	19	53% (10)	47% (9)	1487 (763 NA)
2) No treatment	143	36% (51)	64% (92)	2240 (1793 NA)

Figure 4: Adjuvant treatment survival. This shows the survival probability curves stratified for patients with treatment (lower blue curve) and without treatment (upper brown curve) for the full group. The apparent poorer survival trend for those with treatment [HR=1.82] was not significant.



	No. of Subjects	Event	Censored	Median Survival (95% CL)
1) Female	84	26% (22)	74% (62)	2362 (1891 NA)
2) Male	78	50% (39)	50% (39)	1708 (921 NA)

Figure 5: Gender survival. This shows the survival curves for males (lower brown curve) and females (upper blue curve) for the full group. Males have a significantly elevated hazard [HR = 2.03] compared to the females. The survival advantage for females appears limited to the short-mid terms.

Table 6: Survival analysis associations. This table provides the hazard ratios (HRs) with 95% confidence intervals, the Wilcoxon (Wilcox), chi-square (chi-sq), and Log-rank (LgR) test p-values (p-val) and the percentage of patients surviving (Sur) past 3, 5, and 7 years for the various groups. The number of patients in each stratification belonging to the censored group (n_c), incident group (n_I), and totals (n) for each experiment are also provided. We show the survival statistics for age, histology subtype restricted to adenocarcinoma (AC) and squamous cell carcinoma (SCC), disease recurrence (Rec), adjuvant treatment (treatment), and gender. The reference (ref) groups are designated below. Recurrence and adjuvant treatment are denoted by Rec and treatment respectively.

Model / Group	**N** **n_I, n_C**	**Wilcox Ch-sq**	**LgR Ch-sq**	**HR (95% CI)**	**3 Year % Sur**	**5 Year % Sur**	**7 Year % Sur**
Survival Age	162 61,101	5.01 (0.025)	5.75 (0.016)	1.86 (1.1,3.1)			
Lower-age group (ref)	82 24, 58				79.1%	64.4%	32.2%
Upper-age group	80 37, 43				63.3%	46.7%	29.1%
Survival Histology	141 56, 85	5.08 (0.024)	4.68 (0.030)	1.78 (1.1,3.0)			
AC (ref)	93 30, 63				77.3%	57.6%	35.4%
SCC	48 26, 22				57.3%	45.5%	15.2%
Survival Rec	158 59, 99	18.28 (0.000)	28.79 (0.000)	4.16 (2.4,7.3)			
No Rec (ref)	132 39, 93				79.6%	65.0%	33.9%
Rec	26 20, 6				33.7%	11.2%	0.00%
Survival Treatment	162 61,101	1.62 (0.203)	3.04 (0.081)	1.82 (0.92,3.6)			
No Treatment (ref)	143 51, 92				72.4%	59.8%	29.6%

Table 6 (Continued)

Treatment	19 10, 8				63.2%	28.1%	28.1%
Survival Gender	162 61,101	9.39 (0.002)	7.27 (0.007)	2.03 (1.20,3.4)			
Female (ref)	84 22, 62				81.7%	67.6%	25.80 %
Male	78 39, 39				60.2%	44.7%	26.9%

Additional analysis was applied to evaluate various strata based on the stage I subgroups (graphs and tables not shown). Stratification by stage IA and IB subgroups did not result in a significantly different hazard. However, stratification by considering all patients in the lower-age group with stage IA compared to the remaining population was associated with a significant hazard [HR = 2.44, (CI: 1.30, 4.59)], indicating that older age in combination with stage IB confers poor survival relative to the remaining stage I population. Similarly, stratification by considering patients with both AC and Stage IA compared to the remaining population resulted in a significant hazard [HR = 2.18, (CI: 1.22, 3.88)]. Stratification by considering patients with AC, stage IA and in the lower-age group compared with the remaining population was associated with a significant hazard [HR= 2.65, (CI: 1.20, 5.84)]. Thus, stage subgroup was related to survival when considering those patients with specific clinical factors.

As previously demonstrated [19], the work also shows that the LR analysis based on the incident and censored group for favorable outcome predictions provided a means for determining those variables related to survival as evaluated with accepted analysis methods (i.e. Cox regression and Kaplan Meier analysis).

The analysis methods presented in this chapter were repeated using measures of DNA repair protein expression in the tumor tissues of the patients from this dataset presented in the next chapter.

Chapter 3: Analysis of DNA Repair Pathway Expression in Stage I NSCLC

Both clinical and molecular characteristics of tumors can be used to determine prognosis. In particular, the expression of excision repair cross complementing gene 1 (ERCC1) was shown as a prognostic factor in patients with early stage NSCLC [23]. ERCC1 plays an important role in the nucleotide excision repair pathway. Given that DNA repair is mediated by a number of other important pathways as well, the impact of PARP and Ku86 expression was investigated along with clinical factors [2, 15] to identify the variables that either limit survival or confer survivability [24, 25]. Ku86 and PARP are involved in the non-homologous end joining and base excision repair pathways respectively. It was hypothesized that the expression of each of these proteins could influence prognosis and treatment selection for patients with NSCLC. We also hypothesized that these expression measures may be related to disease recurrence.

3.1 Experimental Protein Expression Biomarker Measures

DNA repair protein expression for Ku86 and PARP were evaluated as disease biomarkers from digitally scanned, immunohistochemically stained NSCLC tissue microarrays (TMA). A board certified pathologist reviewed selected hematoxlyin and eosin (H&E) stained slides to confirm the diagnosis and grade of the tumors. Subsequently, areas of interest were identified on the H&E stained slides and tissue cores were obtained from the corresponding areas of the originating formalin fixed paraffin embedded tissue block using a semi-automated tissue microarrayer (Pathology Devices, Westminster, MD). Due to the inherent histological heterogeneity of non small cell lung carcinomas, especially adenocarcinomas, triplicate tumor tissue cores were obtained to account for tumor heterogeneity. Immunohistochemical staining of the 4 μm thick sections from the TMA blocks was performed using monoclonal antibodies to Ku86(SC- 56136 Santa Cruz) and PARP (Cat # 630210; Clontech) according to the manufacturer's instructions for immunhistochemical staining using an automated stainer (Dako, Carpinteria, CA). The immunohistochemically stained TMA slides were digitized using a Nanozoomer© whole slide scanner (Olympus, Center Valley, PA). Protein expression was evaluated from digitized, stained TMA pathology images using a modified scoring methodology. The stain intensity was assessed for each tissue core manually by the pathologist. A final intensity score for each patient was determined by averaging the scores of the 3 tumor cores of a sample. This method of scoring was repeated for every patient. The three measures were derived for each DNA repair protein expression that measured intensity (I), proportion (P) and total score (S) [i.e. intensity × proportion]. For Ku86, we refer to these as K_I, K_P, and K_S, respectively. Similarly, the corresponding PARP measures were referred to as P_I, P_P, and P_S respectively.

Due to the proprietary nature of the novel methodology described above, these TMA images were not available or accessible for automated analysis. Only the manual scoring data derived from the pathologist was available for the work presented in this chapter.

3.2 Study Population and Methods of Analysis

This data corresponds to the NSCCLC patients and data described in the previous chapter. Please refer to Chapter 2, section 2.1 for the description of study population and section 2.2 for analysis methods. Favorable outcomes were modeled in a similar fashion. Variable interactions were also investigated. In this modeling, histology (either AC or SCC), gender, and recurrence were used as the dependent variables for the LR model (as demonstrated in Chapter 2). The LR models included the following variables: age (A), tumor-grade (Gr), gender (G) and the protein expression measures. Each of these clinical and pathological variables are generically referred to as CL1, and each of the six DNA repair protein expression variables are generically referred to as PR in this description. The following relationships were investigated to predict the two class histology subtypes (i.e. predict SCC): LR(CL1, PR), LR(A, Gr, PR), LR(A, G, PR), and LR (A, Gr, G, PR). This evaluation included 56 different LR models. We performed a similar analysis to predict male gender. The independent variables used in the LR model to predict male gender can be determined from the histology (H) subtype analysis by replacing G with H (i.e. 56 different LR models). Various relationships were used to predict recurrence. We refer to A, Gr, H, and G individually as the CL2 variables in this description. We investigated, LR(CL2), LR(PR), LR(CL2, PR), LR(A, Gr), LR(A, H), LR(A, G), LR(CL2, PR), LR(A, Gr, H), LR(A, Gr, G), LR(A, H, G), LR(A, Gr, PR), LR(A, H, PR), LR(A,G, PR). LR(Gr, H, PR), LR(Gr, G, PR), LR(H, G, PR), LR(A, Gr, H, PR), LR(A, H, G, PR), LR(Gr, H, G, PR), LR(A, Gr, H, G, PR). This evaluation included 110 LR models. The models were formed manually (without automated selection). For the survival analysis using DNA repair expression measures, similar methods were followed, as described in Chapter 2, section 2.2.

3.3 Results

3.3.1 Favorable Outcome Analysis

The additional patient characteristics for biomarker groups are summarized in Table 7 by incident and censored group. The Ku86 and PARP expression for each of the three measures were similar across the two groups. In the modeling and survival analyses below, K_I findings were excluded because all males had K_I =3, as did most females (majority of patients had the same value).

Table 7: Patient characteristics for biomarker groups. This table provides the patient characteristics for the incident group (I), censored group (C), and total (Tot) for the biomarkers. The number of samples (n), mean values (Mean), standard deviation (SD) and percentages (%) are provided for each characteristic where applicable. The C and I labels correspond to the favorable and unfavorable outcome groups, respectively.

Biomarker	**I**		**C**		**Tot**	
	n	**Mean / SD or %**	**n**	**Mean / SD or %**	**n**	**Mean / SD or %**
K_P	61	96.9 / 6.6	101	97.1 / 4.0	162	97.1 / 5.1
K_S	61	289.5 /24.8	101	288./ 21.5	162	288.7 / 22.7
P_I	61	2.58 / 0.60	101	2.51 / 0.62	162	2.54 / 0.61
P_P	61	83.6 / 17.5	101	84.9 / 18.2	162	84.4 /17.9
P_S	61	223.4/ 76.9	101	221.6 /77.4	162	222.8 / 76.7

For the full group dataset (n=162, with n_1=101 and n_2 = 61), the DNA repair protein expression measures were not significant independent predictive factors for favorable outcome (i.e. Az < 0.600) and their OR associations were not significant as shown in Table 8.

Table 8: DNA repair associations for the full group dataset. The odds ratios (ORs) with 95% confidence intervals, and area under the receiver operating characteristic curve (Az) are provided for the various arrangements. We use the functional notation LR(x,y) to indicate the variable(s) within the logistic regression (LR) model used to predict the censored group. The unit and reference (ref) and covariate (Cov) ORs are also provided for each model.

Model	Cov	Unit/Ref	Cov OR	Az
LR(K_P)=censored group	K_P	5.121	1.03 (0.75, 1.41)	0.479
LR(K_S)=censored group	K_S	22.753	0.95 (0.68, 1.32)	0.530
LR(P_I)=censored group	P_I	0.612	0.90 (0.65, 1.25)	0.527
LR(P_P)=censored group	P_P	17.896	1.08 (0.79, 1.47)	0.540
LR(P_S)=censored group	P_S	76.711	0.98 (0.71, 1.34)	0.498

In the subgroup-1 dataset (n =149, with n_1 =91 and n_2= 58), patients that had complete ascertainment for age, gender, tumor-grade, tumor-location, histology-subtype, and DNA repair protein expression measures were included. None of DNA repair protein expression measures demonstrated significant association with favorable outcome (data not shown).

The subgroup-2 dataset (n= 134 with n_1= 80 and n_2= 54) was investigated by restricting the analysis to patients that had full case ascertainment for the SCC and AC histology-subtypes in conjunction with age, gender, tumor-location, adjuvant treatment, tumor-grade, stage I subgroup, and DNA repair protein expression measures. None of the DNA expression measures demonstrated significant association with favorable outcome or significantly altered the associations found previously in Chapter 2 when modeled simultaneously with the respective covariates (associations not shown).

3.3.2 Interaction Analysis

To understand possible interactions between age and the DNA repair protein expression measures, inter-measurement analyses were performed using the subgroup-2 patients. This analysis was portioned into three outcomes by considering those variables that could predict (i) the two-subgroup histology [i.e. AC or SCC], (ii) gender, and (iii) recurrence. To limit the presentation, only models that were associated with Az equal to or above 0.600 are shown when the corresponding ORs were not significant.

In the LR histology association analysis, there were many models that provided some predictive capability, whereas, the majority of the OR associations were not significant as show in Table 9. The associations for the PARP and Ku86 expression measures were not significant, although the gender association [OR = 2.17] gained significance in the LR(Gr, G, K_S) model indicating males are more likely to have the SCC histology subtype than females.

In the LR models for gender interaction analyses (Table 10), the majority of the associations were not significant except for those provided by tumor-grade and Ku86 expression. Although grade in isolation provided weak predictive capability (Az=0.596), its association [OR = 1.84 per unit increase] indicates increasing grade is significantly related to male gender; in models that included grade, similar associations were found (i.e. 1.46 - 2.01 range of ORs). Similarly, K_S in isolation provided a significant association [OR = 2.03 per SD increase] with Az = 0.635, indicating a relationship with gender (i.e. increasing K_S is related to male gender). In most models that included K_S, it provided significant associations (i.e. 1.84 - 2.14 range of ORs). Histology gained significance when including K_S. In this bivariate model the associations for histology subtype, [OR=2.14] and K_S [OR = 2.05 per SD increase] with the Az = 0.681, show males are more likely to have SCC and an increased K_S measure.

To understand possible interactions with disease recurrence, several (i.e. 110) LR models were investigated. We found only weak predictors of recurrence with no significant OR relationships. For example in summary, no univariate model was associated with an Az greater than 0.582 (provided by tumor-grade), no bivariate model had an Az greater than 0.619 (grade and P_I), no trivariate model had an Az greater than 0.625 (grade, gender, P_I), and no models with four variates resulted in an Az as large as the best trivariate model (data not shown).

Table 9: Histology association analysis. In this analysis we use the subgroup-2 dataset to predict two class histology: adenocarcinoma (AC) and squamous cell carcinoma (SCC). This model included age (A), gender (G), Adjuvant treatment (T) tumor-grade (Gr), stage I subgroup (S1), and tumor-location (Loc), and DNA expression measures. We use the functional notation LR(x,y) to indicate the variable(s) within the logistic regression model used to predict SCC. The unit and reference (ref) and covariate (Cov) ORs are also provided for each model with 95% confidence intervals (CIs).

Model	Cov	Unit/RefS	Versus	Cov OR	Az
LR(A, P_I)=SCC	Age	7.967	NA	1.75 (1.17, 2.62)	0.652
	P_I	0.605	NA	0.82 (0.57, 1.18)	
LR(A, P_P)=SCC	Age	7.967	NA	1.70 (0.14, 2.53)	0.637
	P_P	18.082	NA	0.98 (0.68, 1.42)	
LR(A, P_S)=SCC	Age	7.967	NA	1.74 (1.17, 2.60)	0.643
	P_S	75.787	NA	0.87 (0.60, 1.25)	
LR(A, K_P)=SCC	Age	7.967	NA	1.70 (1.15, 2.52)	0.639
	K_P	3.601	NA	0.97 (0.67, 1.40)	
LR(A, K_S)=SCC	Age	7.967	NA	1.73 (1.16, 2.58)	0.652
	K_S	19.677	NA	0.75 (0.51, 1.12)	
LR(G, P_I)=SCC	Gender	Female	Male	1.99 (0.96, 4.14)	0.605
	P_I	0.605	NA	0.85 (0.59, 1.21)	
LR(G, P_S)=SCC	Gender	Female	Male	1.95 (0.94, 4.03)	0.604
	P_S	75.787	NA	0.91 (0.63, 1.30)	
LR(G, K_P)=SCC	Gender	Female	Male	1.93 (0.93, 4.02)	0.604
	K_P	3.601	NA	0.95 (0.66, 1.37)	
LR(G, K_S)=SCC	Gender	Female	Male	2.24 (1.05, 4.77)	0.623
	K_S	19.677	NA	0.72 (0.50, 1.05)	

Table 9 (Continued)

LR(A, Gr, P_I)=SCC	Age	7.967	NA	1.82 (1.20, 2.76)	0.660
	Grade	1	NA	1.42 (0.79, 2.58)	
	P_I	0.605	NA	0.82 (0.57, 1.18)	
LR(A, Gr, P_P)=SCC	Age	7.967	NA	1.77 (0.18, 2.67)	0.653
	Grade	1	NA	1.44 (0.79, 2.60)	
	P_P	18.082	NA	0.96 (0.66, 1.39)	
LR(A, Gr, P_S)=SCC	Age	7.967	NA	1.82 (1.20, 2.75)	0.660
	Grade	1	NA	1.44 (0.79, 2.61)	
	P_S	75.787	NA	0.86 (0.60, 1.24)	
LR(A, Gr, K_P)=SCC	Age	7.967	NA	1.77 (1.18, 2.66)	0.651
	Grade	1	NA	1.44 (0.79, 2.60)	
	K_P	3.601	NA	0.96 (0.66, 1.38)	
LR(A, Gr, K_S)=SCC	Age	7.967	NA	1.82 (1.20, 2.75)	0.667
	Grade	1	NA	1.50 (0.82, 2.74)	
	K_S	19.677	NA	0.73 (0.49, 1.10)	
LR(A, G, P_I)=SCC	Age	7.967	NA	1.70 (1.13, 2.56)	0.680
	Gender	Female	Male	1.80 (0.85, 3.81)	
	P_I	0.605	NA	0.80 (0.55, 1.15)	
LR(A, G, P_P)=SCC	Age	7.967	NA	1.65 (1.11, 2.47)	0.662
	Gender	Female	Male	1.71 (0.82, 3.59)	
	P_P	18.082	NA	0.97 (0.67, 1.41)	
LR(A, G, P_S)=SCC	Age	7.967	NA	1.70 (1.13, 2.55)	0.670
	Gender	Female	Male	1.77 (0.84, 3.73)	
	P_S	75.787	NA	0.84 (0.58, 1.22)	
LR(A, G, K_P)=SCC	Age	7.967	NA	1.66 (1.11, 2.47)	0.665
	Gender	Female	Male	1.75 (0.93, 3.72)	
	K_P	3.601	NA	0.93 (0.64, 1.35)	

Table 9 (Continued)

LR(A, G, K_S)=SCC	Age	7.967	NA	1.68 (1.12, 2.53)	0.681
	Gender	Female	Male	2.04 (0.94, 4.43)	
	K_S	19.677	NA	0.69 (0.45, 1.05)	
LR(Gr, G, P_I)=SCC	Grade	1	NA	1.14 (0.64, 2.02)	0.604
	Gender	Female	Male	1.93 (0.92, 4.07)	
	P_I	0.605	NA	0.85 (0.59, 1.22)	
LR(Gr, G, K_S)=SCC	Grade	1	NA	1.17 (0.65, 2.10)	0.624
	Gender	Female	Male	2.17 (1.01, 4.66)	
	K_S	19.677	NA	0.72 (0.49, 1.05)	
LR(A, Gr, G,P_I)=SCC	Age	7.967	NA	1.75 (1.16, 2.66)	0.678
	Grade	1	NA	1.30 (0.70, 2.40)	
	Gender	Female	Male	1.67 (0.77, 3.61)	
	P_I	0.605	NA	0.80 (0.56, 1.16)	
LR(A,Gr,G,P_P)=SCC	Age	7.967	NA	1.71 (1.13, 2.59)	0.666
	Grade	1	NA	1.32 (0.72, 2.45)	
	Gender	Female	Male	1.59 (0.74, 3.39)	
	P_P	18.082	NA	0.96 (0.66, 1.39)	
LR(A,Gr,G, P_S)=SCC	Age	7.967	NA	1.75 (1.15, 2.66)	0.672
	Grade	1	NA	1.32 (0.72, 2.44)	
	Gender	Female	Male	1.64 (0.76, 3.52)	
	P_S	75.787	NA	0.84 (0.58, 1.22)	
LR(A,Gr,G, K_P)=SCC	Age	7.967	NA	1.71 (1.13, 2.58)	0.665
	Grade	1	NA	1.33 (0.72, 2.44)	
	Gender	Female	Male	1.63 (0.75, 3.52)	
	K_P	3.601	NA	0.92 (0.63, 1.34)	
LR(A,Gr,G, K_S)=SCC	Age	7.967	NA	1.75 (1.15, 2.66)	0.683
	Grade	1	NA	1.36 (0.73, 2.53)	
	Gender	Female	Male	1.89 (0.85, 4.16)	
	K_S	19.677	NA	0.68 (0.45, 1.04)	

Table 10: Gender association analysis. In this analysis we use the subgroup-2 dataset to predict gender. This model included age (A), Adjuvant treatment (T) tumor-grade (Gr), histology (H) including adenocarcinoma (AC) and squamous cell carcinoma (SCC), stage I subgroup (S1), tumor-location (Loc), and the DNA expression measures. We use the functional notation LR(x,y) to indicate the variable(s) within the logistic regression model used to predict male gender. The unit and reference (ref) and covariate (Cov) ORs are also provided for each model.

Model	Cov	Unit/Ref	Versus	Cov OR	Az
LR(K_P)=MALE	K_P	3.601	NA	1.46 (0.98, 2.17)	0.625
LR(K_S)=MALE	K_S	19.677	NA	2.03 (1.08, 3.82)	0.635
LR(A, P_I)=MALE	Age	7.967	NA	1.32 (0.93, 1.88)	0.607
	P_I	0.605	NA	1.26 (0.89, 1.80)	
LR(A, P_P)=MALE	Age	7.967	NA	1.35 (0.95, 1.92)	0.585
	P_P	18.082	NA	1.08 (0.76, 1.52)	
LR(A, P_S)=MALE	Age	7.967	NA	1.32 (0.93, 1.88)	0.600
	P_S	75.787	NA	1.22 (0.86, 1.73)	
LR(A, K_P)=MALE	Age	7.967	NA	1.33 (0.93, 1.91)	0.654
	K_P	3.601	NA	1.43 (0.96, 2.11)	
LR(A, K_S)=MALE	Age	7.967	NA	1.33 (0.93, 1.92)	0.666
	K_S	19.677	NA	1.96 (1.05, 3.65)	
LR(Gr, P_I)=MALE	Grade	1	NA	1.88 (1.07, 3.29)	0.623
	P_I	0.605	NA	1.34 (0.94, 1.91)	

Table 10 (Continued)

LR(Gr, P_P)=MALE	Grade	1	NA	1.82 (1.05, 3.17)	0.584
	P_P	18.082	NA	1.08 (0.76, 1.52)	
LR(Gr, P_S)=MALE	Grade	1	NA	1.84 (1.06, 3.21)	0.624
	P_S	75.787	NA	1.27 (0.89, 1.81)	
LR(Gr, K_P)=MALE	Grade	1	NA	1.80 (1.03, 3.14)	0.646
	K_P	3.601	NA	1.43 (0.97, 2.10)	
LR(Gr, K_S)=MALE	Grade	1	NA	1.75 (1.00, 3.07)	0.655
	K_S	19.677	NA	1.94 (1.05, 3.60)	
LR(H, P_I)=MALE	Histology	AC	SCC	2.00 (0.96, 4.15)	0.607
	P_I	0.605	NA	1.34 (0.94, 1.92)	
LR(H, P_P)=MALE	Histology	AC	SCC	1.90 (0.92, 3.90)	0.573
	P_P	18.082	NA	1.10 (0.78, 1.56)	
LR(H, P_S)=MALE	Histology	AC	SCC	1.95 (0.94, 4.03)	0.610
	P_S	75.787	NA	1.29 (0.91, 1.83)	
LR(H, K_P)=MALE	Histology	AC	SCC	1.93 (0.93, 4.01)	0.660
	K_P	3.601	NA	1.47 (0.98, 2.19)	
LR(H, K_S)=MALE	Histology	AC	SCC	2.14 (1.01, 4.55)	0.681
	K_S	19.677	NA	2.05 (1.10, 3.80)	
LR(A, Gr, P_I)=MALE	Age	7.967	NA	1.41 (0.97, 2.05)	0.655
	Grade	1	NA	2.04 (1.14, 3.64)	
	P_I	0.605	NA	1.28 (0.89, 1.84)	

Table 10 (Continued)

LR(A, Gr, P_P)=MALE	Age	7.967	NA	1.45 (1.00, 2.10)	0.652
	Grade	1	NA	2.00 (1.13, 3.56)	
	P_P	18.082	NA	1.04 (0.73, 1.48)	
LR(A, Gr, P_S)=MALE	Age	7.967	NA	1.42 (0.98, 2.05)	0.652
	Grade	1	NA	2.00 (1.13, 3.56)	
	P_S	75.787	NA	1.21 (0.85, 1.73)	
LR(A, Gr, K_P)=MALE	Age	7.967	NA	1.43 (0.98, 2.07)	0.685
	Grade	1	NA	1.96 (1.10, 3.49)	
	K_P	3.601	NA	1.38 (0.94, 2.02)	
LR(A, Gr, K_S)=MALE	Age	7.967	NA	1.42 (0.98, 2.07)	0.696
	Grade	1	NA	1.90 (1.07, 3.39)	
	K_S	19.677	NA	1.84 (1.01, 3.35)	
LR(A, H, P_I)=MALE	Age	7.967	NA	1.24 (0.86, 1.78)	0.634
	Histology	AC	SCC	1.80 (0.85, 3.83)	
	P_I	0.605	NA	1.30 (0.91, 1.87)	
LR(A, H, P_P)=MALE	Age	7.967	NA	1.28 (0.89, 1.83)	0.621
	Histology	AC	SCC	1.70 (0.81, 3.57)	
	P_P	18.082	NA	1.08 (0.76, 1.53)	
LR(A, H, P_S)=MALE	Age	7.967	NA	1.24 (0.86, 1.79)	0.627
	Histology	AC	SCC	1.76 (0.83, 3.72)	
	P_S	75.787	NA	1.25 (0.87, 1.78)	

Table 10 (Continued)

LR(A, H, K_P)=MALE	Age	7.967	NA	1.26 (0.87, 1.82)	0.681
	Histology	AC	SCC	1.74 (0.82, 3.69)	
	K_P	3.601	NA	1.44 (0.97, 2.14)	
LR(A, H, K_S)=MALE	Age	7.967	NA	1.24 (0.85, 1.81)	0.698
	Histology	AC	SCC	1.94 (0.89, 4.20)	
	K_S	19.677	NA	1.99 (1.07, 3.68)	
LR(Gr, H, P_I)=MALE	Grade	1	n/a	1.85 (1.05, 3.26)	0.643
	Histology	AC	SCC	1.94 (0.92, 4.08)	
	P_I	0.605	NA	1.37 (0.95, 1.97)	
LR(Gr, H, P_P)=MALE	Grade	1	NA	1.79 (1.02, 3.13)	0.620
	Histology	AC	SCC	1.84 (0.88, 3.82)	
	P_P	18.082	NA	1.08 (0.76, 1.53)	
LR(Gr, H, P_S)=MALE	Grade	1	NA	1.81 (1.03, 3.18)	0.641
	Histology	AC	SCC	1.89 (0.90, 3.96)	
	P_S	75.787	NA	1.29 (0.90, 1.85)	
LR(Gr, H, K_P)=MALE	Grade	1	NA	1.76 (1.00, 3.10)	0.676
	Histology	AC	SCC	1.87 (0.89, 3.92)	
	K_P	3.601	NA	1.43 (0.97, 2.12)	
LR(Gr, H, K_S)=MALE	Grade	1	NA	1.71 (0.97, 3.01)	0.695
	Histology	AC	SCC	2.07 (0.96, 4.43)	
	K_S	19.677	NA	1.95 (1.07, 3.56)	

Table 10 (Continued)

LR(A,Gr, H,P_I)=MALE	Age	7.967	NA	1.33 (0.91, 1.95)	0.667
	Grade	1	NA	1.99 (1.11, 3.57)	
	Histology	AC	SCC	1.69 (0.79, 3.65)	
	P_I	0.605	NA	1.32 (0.91, 1.90)	
LR(A,Gr,H,P_P)=MALE	Age	7.967	NA	1.38 (0.95, 2.02)	0.666
	Grade	1	NA	1.95 (1.09, 3.49)	
	Histology	AC	SCC	1.59 (0.75, 3.39)	
	P_P	18.082	NA	1.05 (0.73, 1.49)	
LR(A,Gr,H,P_S)=MALE	Age	7.9679	NA	1.34 (0.91, 1.96)	0.662
	Grade	1	NA	1.95 (1.09, 3.49)	
	Histology	AC	SCC	1.65 (0.77, 3.54)	
	P_S	75.787	NA	1.24 (0.86, 1.78)	
LR(A,Gr,H,K_P)=MALE	Age	7.967	NA	1.35 (0.92, 1.99)	0.703
	Grade	1	NA	1.91 (1.06, 3.42)	
	Histology	AC	SCC	1.63 (0.76, 3.51)	
	K_P	3.601	NA	1.39 (0.95, 2.05)	
LR(A,Gr,H,K_S)=MALE	Age	7.967	NA	1.33 (0.90, 1.97)	0.717
	Grade	1	NA	1.84 (1.02, 3.30)	
	Histology	AC	SCC	1.81 (0.82, 3.97)	
	K_S	19.677	NA	1.86 (1.03, 3.36)	

In summary, the favorable outcome modeling produced little when using these novel DNA repair markers.

3.3.3 Survival Analysis

The DNA repair protein expression survival findings are shown in Table 11. None of these measures showed significance. The P_I measure showed the strongest trend [HR = 1.49]. Approximately 64% of the patients in the lower P_I group survived past 5 years, whereas 45% of the patients in the upper P_I group survived past this time. These findings (i.e. no relationship to hazard) reinforce the principle discussed earlier that the favorable outcome findings parallel the survival analysis findings for a given variable.

Table 11: DNA repair expression measures and survival. This table provides the hazard ratios (HRs) with 95% confidence intervals, the Wilcoxon (Wil), Chi-square (Chi-sq), and Log-rank (LgR) test p-values and the percentage of patients surviving (Sur) past 3, 5, and 7 years for the various DNA expression measures. Patients were dichotomized by their respective expression distribution median values (i.e described as low and high). The number of patients in each stratification belonging to the censored group (n_c), incident group (n_I) and totals (n) for each experiment are also provided.

Model / Group	**N (n_I, n_C)**	**Wil Chi-Sq (p-val)**	**LgR Chi-Sq (p-val)**	**HR (95% CI)**	**3 Year % Sur**	**5 Year % Sur**	**7 Year % Sur**
Survival K_P	162 (61,101)	0.21 (0.64)	0.02 (0.87)	0.96 (0.58,1.6)			
Low-K_P	67 (26, 41)				69.95 %	55.28 %	29.94 %

Table 11 (Continued)

High-K_P	95 (35, 60)				72.25 %	56.41 %	23.81 %
Low-K_S	68 (26, 42)				70.40 %	55.64 %	29.67 %
High-K_S	94 (35, 59)				71.95 %	56.18 %	23.71 %
Survival P_I	162 (61,101)	1.01 (0.31)	2.23 (0.13)	1.49 (0.88, 2.5)			
Low-P_I	76 (27, 49)				74.77 %	64.13 %	33.23 %
High-P_I	89 (34, 52)				69.69 %	44.65 %	33.49 %
Survival P_P	162 (61, 101)	0.52 (0.46)	0.15 (0.69)	0.90 (0.55, 1.5)			
Low-P_P	72 (30, 42)				67.91 %	53.60 %	31.26 %
High-P_P	90 (31, 59)				74.00 %	57.74 %	22.50 %
Survival P_S	162 (61,101)	0.26 (0.61)	0.61 (0.43)	1.23 (0.7, 2.07)			
Low-P_S	77 (30, 47)				72.39 %	59.95 %	31.06 %
High-P_S	85 (31, 54)				70.26 %	48.96 %	36.72 %

The analysis of this population of stage I NSCLC patients is continued in the following chapter using the novel statistical learning (SL) methods.

Chapter 4: A Fusion of Statistical Learning Techniques with Accepted Epidemiologic Applications

4.1 Background

Statistical learning (SL) techniques with kernel mappings can provide benefits when addressing complicated decision problems [26-28]. These techniques are capable of capturing non-linear input-output characteristics, operating on small datasets with feature correlation, and do not require modeling or distribution assumptions. These attributes are not derived without tradeoffs. These methods do not provide an output that has a useful epidemiologic interpretation and their training often requires specialized techniques. In contrast, logistic regression (LR) modeling, Kaplan-Meier analysis, and Cox regression provide important epidemiologic interpretations and are used extensively due to their availability.

The goal of this work primarily was to demonstrate and evaluate a method of fusing SL with accepted epidemiologic practice using this dataset of lung cancer patients as an example and test-bed. The work in this chapter was excerpted from Behera et al [19]. To meet this objective, it was necessary to develop a platform to implement kernel based SL methods efficiently, which will also support future studies. A technique validated here, (applied in Chapter 2) was used as an efficiency gain. This shows that either nonparametric Az analysis (i.e. without modeling) or LR in conjunction with Az analysis can be used as a sifting or filtering mechanism to find variables that influence the probability of survival characterized by either HRs or Kaplan-Meier comparisons. Although the LR model fusion with SL provides an important interpretation and application in its own right, it is relatively simple to estimate Az within our programming language, whereas incorporating Cox regression or Kaplan Meier analysis as intermediary steps within our processing routines would require considerable code development. Thus, the use of binary separation analysis (used in the favorable outcome modeling) in conjunction with SL based survival analysis also represents an import efficiency step in the hybrid analysis.

To adapt SL methodology for epidemiologic application, a probabilistic neural network (PNN) [29] was combined with LR modeling and survival analyses (i.e. Kaplan-Meier analysis and Cox regression) to demonstrate the concept. This hybrid approach combines the strengths of the SL methodology with these important epidemiologic techniques. The PNN is a statistically inspired neural network [29] that uses a kernel mapping [30, 31] to estimate the underlying probabilities. The PNN was adapted to provide a patient score, which is different use than its intended classification application. For the LR modeling comparisons, the favorable and unfavorable group analysis presented in Chapter 2 was used to dichotomize the patient population for the LR analysis. Raw clinical variables were used to form a new patient score variable with the modified PNN. Additionally, the PNN output (i.e. the patient score) was used as the input variable for survival analysis. There are weight parameters within the PNN (i.e. the kernel sigma-weights) that must be estimated properly. Differential Evolution (DE) was used for this optimization problem [32]. DE is an evolutionary computing strategy for global optimization tasks. Because the dataset was limited, stochastic methods were developed to provide feedback to the DE optimization and to derive the patient PNN scores. This new system was also evaluated with the simulated datasets and methods described previously [33].

4.2 Methods: Modeling Techniques

4.2.1 Favorable Outcome and Survival Analysis

The non-interaction LR model [34] was used to predict favorable and unfavorable survival outcome (explained in detail in Chapter 2). Although the work in this chapter is presented after the work in chapters 2-3, it was performed at an earlier date [19]. The dataset was constructed by considering those patients that has complete ascertainment for smoking status, age, grade, and gender. For the LR analysis, we formed incident (n=59) and censored (n=92) groups as in Chapter 2. This dataset was similar with the various subgroups described in Chapter 2.   Three variables were used in the analysis. Age and grade were combined with the PNN to form a hybrid variable labeled as the patient-score. Gender was incorporated as a controlling variable. The patient-score was used as the input to LR and survival analysis (Cox regression and Kaplan Meier Analysis). The reasons for combing age and grade are as follows. Grade showed weak association in the previous analysis (i.e. trend only). This could imply either the association is truly weak or it exists and could not be captured by a linear technique. Moreover, age is continuous, and grade can be considered as a three-state continuous variable both amenable for probability modeling, whereas the other categorical variables do not (strictly) lend themselves to probability modeling. The hybrid variable modeling was compared with the accepted approaches (LR and survival analysis) using age, grade, and gender as the inputs. In this survival analysis, patient strata were formed by choosing the median age and median PNN score as the separation points. The other relevant variables were introduced with both age and PNN score to evaluate their influence on the respective survival probability curves.

As previously, for the LR modeling comparisons, ORs were used to assess measurement association with 95% CIs and Az was used to assess predictive capability.   For age and PNN score (i.e. the continuous variables), the LR model coefficients were re-scaled to provide ORs per SD change for each variable. The ORs for grade were cited in per unit increase. The Az was estimated with three methods. First, to assess the SL training and patient scores, the definition of Az was applied [35] using the respective distributions. Secondly, the Az quantities for the LR models were generated within SAS as described in Chapter 2 using the output of the LR model (same interpretation as provided by the first method). For the Kaplan-Meier analysis, chi-square Wilcoxon (more sensitive to shorter term survival differences) and log-rank (more sensitive to longer term survival differences) tests were used for differences in stratification. Hazard ratios with 95% CIs were estimated with Cox Regression. Thirdly, Az was also derived from Cox regression and is a measure of the agreement between the model and actual time-to-event outcome [36], which is a different interpretation.

4.2.2 Probabilistic Neural Network and Kernel Methods

A variation of the PNN was implemented using a Gaussian kernel, although there are many kernels meeting the established criteria [37]. Paralleling our earlier work [38], the distance metric for a d dimensional input vector (i.e. the relevant patient variables) is given by

$$D_i(\mathbf{w}) = \sum_{j=1}^{d} \frac{(w_j - w_{ij})^2}{\sigma_j^2}, \qquad \text{Eq. (1)}$$

where i is the patient index, w_{ij} is the j^{th} component of the i^{th} sample's input vector, and w_j is the j^{th} component of a prospective test sample's input vector **w**. The sigma-weights, σ_j , were estimated with DE optimization. Specifically, d = 2, with w_{i1} = age, and w_{i2} = grade for the i^{th} patient. The probability density estimation [30, 31] for **w** with n training samples is expressed as

$$g(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \exp[-D_i(\mathbf{w})] = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{w}, \mathbf{w}_i). \qquad \text{Eq. (2)}$$

Normalization factors (ignored) are discussed below. The PNN was constructed with the above formulism for each group. For group-1, the density for **w** is given by

$$g_1(\mathbf{w}) = \frac{1}{n_1}\sum_{i=1}^{n_1} k(\mathbf{w}, \mathbf{w}_i) \quad . \qquad \text{Eq. (3)}$$

For a given **w**, the sum on $\mathbf{w}_i$ is taken over group-1 samples only with $n = n_1$. The $g_2(\mathbf{w})$ density was estimated the same way by restricting the sum on $\mathbf{w}_i$ to the group-2 samples with $n = n_2$. In both the g_1 and g_2 estimations, **w** included samples from both groups. Equation (3) [i.e. the kernel mapping] also represents a function mapping of the vectors **w** and $\mathbf{w}_i$, where each element (for fixed i) of the summation represents the inner product of the mapped vectors [28], rendering a nonlinear problem tractable with the proper choice of kernel. Assuming prior probabilities and misclassification costs are equal, the PNN classifier [29] is expressed as

$$\frac{g_1(\mathbf{w})}{g_2(\mathbf{w})} > c \quad , \qquad \text{Eq. (4)}$$

where c is a constant. For classification when this condition is met, **w** belongs to group-1. Because we were interested in developing a score for each patient (not classification), we formed a score with the above expression given by

$$\text{patient-score} = \frac{g_1(\mathbf{w})}{g_2(\mathbf{w})} \, . \qquad \text{Eq. (5)}$$

The multivariate normalization factors were not important for this application because both g_1 and g_2 contained the same sigma-weights. These scores were used with LR modeling and the survival analysis. Because the above expression is always positive and can be large, we used z = ln (patient-score) in the analyses as the PNN derived patient score and performed a range compression technique to reduce statistical outlier interference in the LR modeling.

4.2.3 Probabilistic Neural Network Training and Operation

A stochastic cross-validation technique was developed in combination with DE to estimate the sigma-weights for the kernel in the PNN. DE is a stochastic global optimization strategy that is self-organizing via feedback and represents an evolutionary process. An algorithm described by the founders of DE [32] was developed and their notation is used in this work. Important points underlying DE were discussed in our previous work [38] and are briefly discussed here. A uniform crossover Cr = 0.9 and scale factor F = 0.2 were used as starting points. The zero-generation vector population (i.e. NP = 40 vectors) was initialized with uniformly distributed random variables with components constrained to this range [0.01, 1.5]. The vector's components are the two sigma-weights. For a given generation, the DE process constructs a mutant vector (or $\mathbf{v}_g$) by stochastic perturbation from the current population of **x**, where g is the generation index. From this, a candidate vector (or $\mathbf{u}_g$) is constructed that competes with a given current generation vector, $\mathbf{x}_g$, selected at random in such a way that it was not involved with the $\mathbf{v}_g$ (or $\mathbf{u}_g$) construction. Possible solutions ($\mathbf{x}_g$ and $\mathbf{u}_g$) compete against each other using feedback from the optimization problem. The winner moves to the next generation of **x** (i.e. the g+1 generation). For a given generation, there are NP competitions. In this DE application, Az was the feedback measure using the two patient-score distributions (i.e. for the censored and incident groups) derived from Eq. (5). The feedback to the DE was formed by ensemble averaging derived with bootstrap sampling [39]. For one DE generation, N_t bootstrap populations were generated. To form a given bootstrap population, n_2 samples were selected randomly from group-1 and from group-2 with replacement. We keyed on n_2 as not to bias the sampling to the larger population. One sample from each class was selected randomly and used as **w** in Eq. (5) to generate the respective patient-score quantities. The remaining samples were used to build the respective $\mathbf{w}_i$ populations in Eq. (5). We refer to this process as a leave two-out stochastic cross-validation technique. When $N_t = 1$, the process is somewhat similar to the conventional leave two-out approach using different realizations of the population. This process was then repeated N_t (i.e. training) times and the average Az was used as feedback for one DE generation. The process was terminated after G generations. The weights that provided the largest Az were carried over to the analysis and used to generate z for each patient using stochastic methods and ensemble averaging. For a given **w**, a bootstrap population was generated from the $\mathbf{w}_i$ population and the respective z was generated for all n_1 and n_2 patients. Each patient's z was derived from ensemble averaging by repeating this process for N_{sc} times. The training process and final score generation flow are shown in the Figure 6 schema. The software for the PNN and DE applications was developed recently [19] using the IDL (ITT Visual Information Solutions, Boulder CL) programming language.

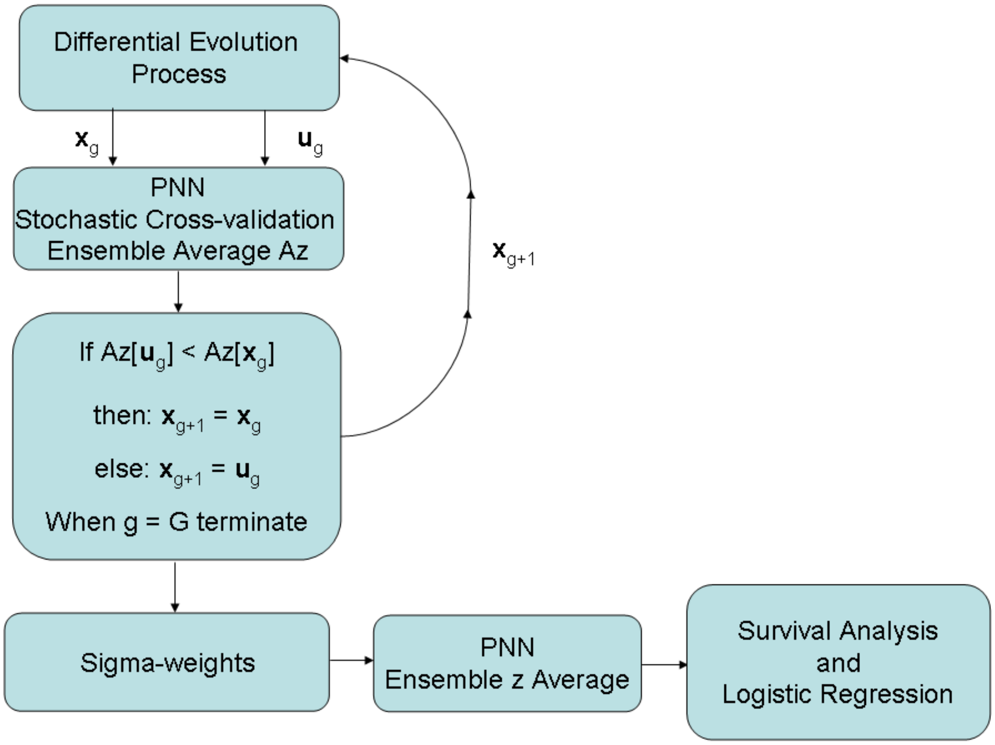

Differential Evolution Process
$\mathbf{x}_g$
$\mathbf{u}_g$
PNN Stochastic Cross-validation Ensemble Average Az
$\mathbf{x}_{g+1}$
If Az[$\mathbf{u}_g$] < Az[$\mathbf{x}_g$]
then: $\mathbf{x}_{g+1}$ = $\mathbf{x}_g$
else: $\mathbf{x}_{g+1}$ = $\mathbf{u}_g$
When g = G terminate
Sigma-weights
PNN Ensemble z Average
Survival Analysis and Logistic Regression

Figure 6: Modified probabilistic neural network (PNN) stochastic training and z generation. This schema shows the PNN training for the Differential Evolution (DE) sigma-weight vector construction, competition, and feedback from the g to the g+1 populations. The sigma-weight vectors $\mathbf{x}_g$ and $\mathbf{u}_g$ compete for the next generation. The receiver operating characteristic curve area (Az) from the stochastic cross-validation is derived with ensemble averaging to reduce the chance of passing outliers back to the vector competition. When g = G, the evolution stops and the sigma-weights are used in the PNN to generate z for each patient stochastically with ensemble averaging. The z quantities are then passed to the survival and logistic regression analyses.

4.3 Results

4.3.1 Favorable Outcomes

The patient characteristics for this specific data subset were statistically similar to those described in Chapter 2 and are not shown. For specific comparison reference, the associations from the LR model (i.e. accepted approach) with age, grade, and gender were estimated for this data subset and are provided in top portion of Table 12 for easy reference, which are similar to those provided in Chapter 2. In the univariate age model (Az = 0.636), the age OR = 0.60 was significant. In the grade adjusted model (Az = 0.657), the OR for age was similar and the grade OR = 0.68 was not significant. In the grade and gender adjusted model (Az = 0.703), the age OR = 0.63 and gender OR = 0.38 were significant, whereas the grade OR = 0.73 was not significant.

Table 12: Odds ratios. The odds ratios (ORs) and 95% confidence intervals are provided parenthetically for the variables used in the logistic regression modeling. The ORs for the continuous variables (age and z) are cited per standard deviation (SD) increase in the respective variable or as a unit increase (grade) while controlling for the other variables when applicable. The z variable includes grade and age simultaneously. The ORs for the other covariates (Cov) are listed in the column to the right. The area under the receiver operating characteristic curve (Az) is also provided for each model.

Model	**SD**	**Age OR**	**Az**	**Cov**	**Unit**	**Cov OR**
Accepted						
Age	8.681	0.60 (0.42,0.86)	0.636			
Grade adjusted	8.681	0.58 (0.40, .83)	0.657	Grade	1	0.68 (0.40, 1.15)
Grade and Gender adjusted	8.681	0.63 (0.43, 0.91)	0.703	Grade	1	0.73 (0.42, 1.25)
				Gender	Male vs Female	0.38 (0.19, 0.78)
Model	**SD**	**ln(z) OR**	**Az**	**Cov**	**Unit**	**Cov OR**
Hybrid						
z (Age and Grade)	1.695	4.15 (2.15, 8.01)	0.763			
Gender adjusted	1.695	3.67 (1.88, 7.16)	0.778	Gender	Male vs. Female	0.50 (0.24, 1.05)

The DE training for the modified PNN resulted in two sigma-weights with σ_1 = 0.013610961 and σ_2 = 0.35805283 for age and grade, respectively. Using N_t = 1 produced training Az values between 0.700-0.830. Choosing N_t = 5 gave consistent findings and was used in the analysis. The stochastic cross-validation performance coinciding with these weights gave Az = 0.710 with SE=0.03 after three generations (G = 3). These parameters were used to generate z for each patient with N_{sc}= 20. Processing age and grade separately through the PNN gave Az = 0.656 for age and Az = 0.538 for grade, which are similar to the Az values when assessing these variables individually with LR modeling (shown in Chapter 2).

The continuous hybrid LR findings are shown in the bottom of Table 12. The combined effect shows that for a SD increase in z (SD=1.69), the respective patient is about 4.15 times more likely to experience a favorable survival outcome (or incident group member is 0.24 more likely to experience a favorable outcome) with Az = 0.763, which was significantly larger (p = 0.0062) than that provided by the respective age and grade LR model. Due to the way the PNN was defined, increasing z was protective, whereas increasing in age was not. Adjusting for gender increased the predictive capability of the model with Az = 0.778 (SE = 0.03), although the gender OR lost significance. Gender also reduced the association for z with OR = 3.67 per standard deviation increase, which was a stronger association than provided by age in the corresponding model. The Az derived from the hybrid model (z and gender) was significantly greater than that of the corresponding LR model with age, grade, and gender (p = 0.0173).

To evaluate the effect of the kernel mapping on age and grade, the LR model outputs for the two models were plotted as a function of grade and age. The left side of Figure 7 shows the grade plots for the LR (accepted approach with age and grade) model. The respective grade plots for the hybrid LR model using z (age and grade combined) are shown on the right side of Figure 7. In these plots, black was used to denote censored group samples and red to denote incident group samples. The grade 1 plots for both models exhibit similar behavior for the lower ages and show that patients 65 years of age and younger are more likely to be in censored group. The hybrid model separates some older grade 1 patients in contrast with the accepted LR model. A comparison of the grade 2 plots shows that the hybrid model provides separation for the younger, middle age, and some upper age patients, whereas the respective accepted LR model produces confusion between the groups. In the grade 3 plots, both models provide separation for lower age patients, whereas the hybrid model shows group separation in the middle-age range as well. Because z is a composite variable and difficult to interpret, the associations between age, grade, z, and group status shown in Figure 7 are also summarized in Table 13. This provides the average values for age and the z variables separated by grade and group.

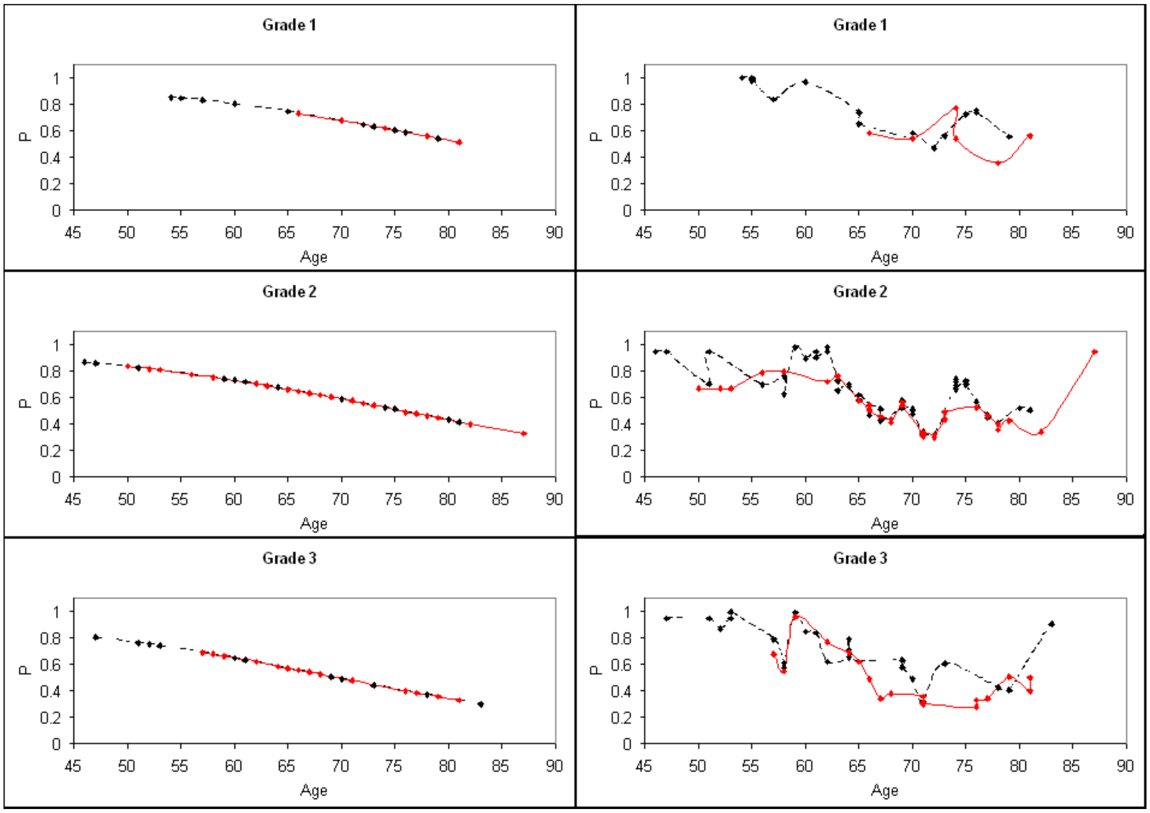


Figure 7: Logistic regression model output plots for each tumor-grade (1-3). The plots on the left show the logistic regression model probabilities (P) using the age and grade variables as the model inputs for each tumor grade. The plots on the right show the respective hybrid logistic regression model probabilities (P) using the variable z (i.e. age and grade combined with the probabilistic neural network) as the model input. Because there are overlapping points (patients with the same grade and age), some points are not distinguishable. The censored group (black) is compared with the incident group (red). The curves were fitted with a cubic spline.

Table 13: Age and z relationships. This table gives the mean values for age and z as a function of tumor-grade (Grade) and censored/incident group status and combined total. The number (n) of patients in each category is also provided.

Censored	**Grade** 1	**Grade** 2	**Grade** 3	**Total**
n	17	49	26	92
Age (mean)	66.41	66.27	63.19	65.42
z (mean)	2.11	3.91	3.12	3.36
Incident	**Grade** 1	**Grade** 2	**Grade** 3	**All**
n	6	34	19	59
Age (mean)	73.83	68.88	69.47	69.58
z (mean)	0.26	1.37	-0.07	0.8

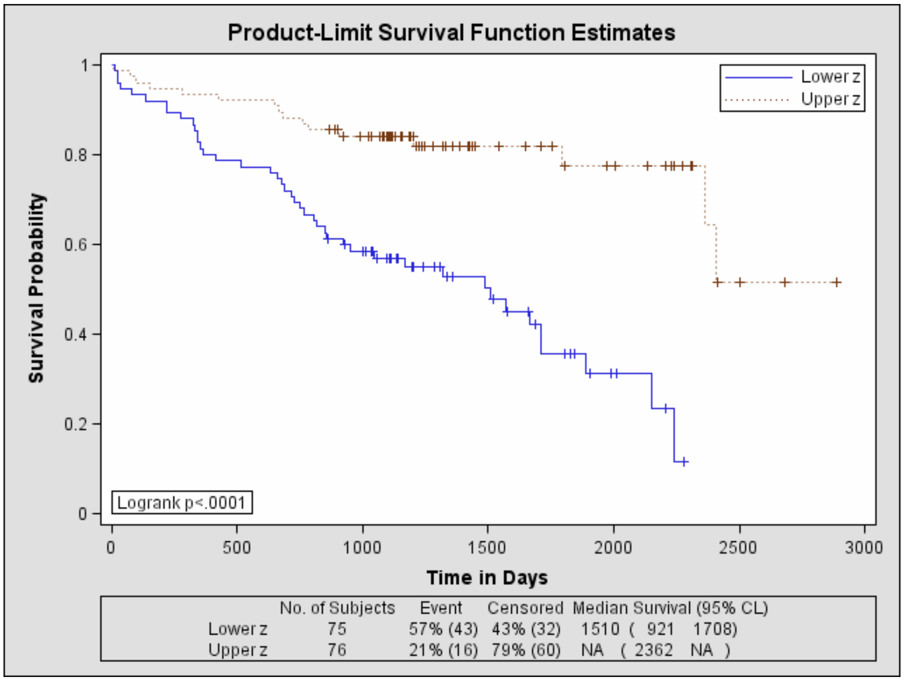
As in the previous chapters, the overall survival (OS) and censor times were used to form two groups because of the separation between the respective distribution means. The favorable group had a mean censor time of 3.97 years (i.e. mean known OS time, which is a low-side limit assuming these patients did not expire the day after study-contact), whereas the incident group had a mean OS time of 2.20 years (data not shown). The minimum censor time (2.35 years) is greater than the mean OS time for the incident group indicating validity of the dichotomization method, which is discussed in more detail in Chapter 6.

4.3.2 Survival Analysis

As above, the results from the standard methods of survival analysis are presented first. The relevant findings for comparison are in Table 14. The hazard for age was HR = 1.72 indicating that upper-age group membership is significantly more hazardous than lower-age group membership. The longer-term survival is significantly different between the two age groups ($p < 0.050$). Including grade induced a greater hazard with HR = 1.78, but the change in the survival curves (not shown) when controlling for grade was not significant in either the short term ($p = 0.074$) or the longer-term ($p = 0.091$). The addition of gender caused a significant change in the survival curves compared with age alone for both the short term ($p < 0.002$) and long-term survival ($p < 0.005$) but the HR = 1.64 lost significance (curves not shown). The grade and gender adjusted hazard for age was HR = 1.68 (also lost significance). The statistical test findings for age and gender are provided in Table 14 (top rows). The survival probability curves for z are shown in Figure 8 and the HRs are provided in Table 14. The findings from these accepted approaches follow those from Chapter 2 and are presented here for reference.

Table 14: Hazard relationships for dichotomous age and z. For the age and z variables, two groups were formed using the respective distribution median as the cut-point and compared. The hazard ratios (HRs) are provided with 95% confidence intervals parenthetically. The area under the receiver operating characteristic curves (Azs) derived from Cox regression models are also provided. Because age and z translate inversely with respect to hazard, increased age confers a greater hazard while decreased z confers a greater hazard. To make HR comparisons of z with age, the reciprocal of the z HR is required.

Model	**Age HR**	**Az**
Accepted		
Dichotomous Age	1.72 (1.02, 2.90)	0.5792
Grade adjusted	1.78 (1.06, 3.02)	0.606
Gender adjusted	1.64 (0.96, 2.78)	0.669
Grade Gender adjusted	1.68 (0.99, 2.85)	0.677
Model	**z HR**	**Az**
Hybrid		
Dichotomous z	0.25 (0.14, 0.47)	0.691
Gender adjusted	0.28 (0.15, 0.53)	0.738



	No. of Subjects	Event	Censored	Median Survival (95% CL)
Lower z	75	57% (43)	43% (32)	1510 (921 1708)
Upper z	76	21% (16)	79% (60)	NA (2362 NA)

Figure 8: Survival probability curves for z. The upper and lower-z groups were formed by dichotomizing the total collection of patients at their median z value. The upper-z group (upper brown curve) exhibits better survival characteristics than the lower-z group (bottom blue curve). These findings incorporate tumor-grade with age via the probabilistic neural network combination.

The hybrid survival analysis is to be compared with the accepted approach. There is a significant survival difference between these upper and lower-z groups (i.e combination of age and grade) both in the short term (p < 0.0001) and long term (p < 0.0001) with HR = 0.25 indicating those in the upper-z group are at a significantly reduced hazard compared with those in the lower z group (i.e. the hazard for those in the lower-z membership was HR = 4.0). About 52% of the upper-z group survived past 7 years, whereas as about 11% of the lower-z group survived past this time. In comparison, 37% of the lower-age (when controlling for grade) group survived past 7 years, whereas about 29% of the upper-age group survived past this time. The addition of gender also produced a significant change in both short term (p = 0.0146) and the longer term (p = 0.0319) with HR = 0.28 (HR = 3.57 for lower-z membership). The associated statistical comparisons for z and gender are provided in Table 15 (bottom two rows). As shown in Table 14, the hybrid Cox model (i.e. using z) showed greater concordance (Az = 0.691) with the outcome than that of the Cox model (accepted approach) using age and grade (Az = 0.606), but the difference in Az was a trend (p = 0.056). Likewise, the Az comparison between the hybrid Cox model using z and gender (Az = 0.738) with the Cox model using age, grade, and gender (Az = 0.677) showed a similar trend (p = 0.0747).

Table 15: Survival probability statistical test summaries. The statistical tests findings (Chi-square and p-values) for the various age and z related survival probability curves are provided with the degrees of freedom (DF). When comparing more than two survival curves (*), the hypothesis that all the curves were the same was tested against the alternative that at least one curve was different.

Model	**Test**	**Chi-Sq**	**DF**	**p-val**
Accepted				
Dichotomous Age over Strata	Log-Rank	4.178	1	0.0409
	Wilcoxon	3.407	1	0.0649
Dichotomous Age and Gender over Strata	Log-Rank	12.738	3	0.0052*
	Wilcoxon	13.511	3	0.0043*
Hybrid				
Dichotomous z over Strata	Log-Rank	22.759	1	<0.0001
	Wilcoxon	14.941	1	0.0001
Dichotomous z and Gender over Strata	Log-Rank	28.186	3	<0.0001*
	Wilcoxon	22.488	3	<0.0001*

4.3.3 Additional Validation Analysis

4.3.3.1 Simulation Evaluation

A simulation was performed to assess the training, optimization, and patient scoring system shown in Figure 4 under ideal conditions. The same simulation described previously [33] was used as the input to the modified PNN. This is a two-class simulation with two correlated input measurements and a non-linear separation boundary. Two hundred samples per class were generated giving 400 samples total as previously [33] for the training dataset. The training dataset was used to estimate the sigma-weights using the algorithm described above (Figure 6). We used the same stochastic averaging (N_t= 5, and N_{sc} = 20) and bootstrap methods. We stopped the differential evolution optimization for G=3 as above, which gave two sigma-weights (0.291156797, 0.0872920) with a training Az = 0.987. The training dataset was then used for $\mathbf{w}_i$ in the score generation using independent data. We then simulated an evaluation dataset of the same dimension (200 per class giving 400 samples total) that was not used in the sigma-weight generation. These new samples were then used as **w** in the stochastic score generation and evaluated. This evaluation gave Az = 0.979. This shows in principle, the system is viable and that the training distribution must be representative of the population. It is also worth noting that the separation provided by this modified PNN system was larger than that described previously using a different statistical learning system when processing the same type of simulated datasets (i.e. Az ≈ 0.950).

4.3.3.2 Traditional Holdout Cross-Validation

To assess the internal validity of the PNN score approach, we used the schema shown in Figure 6 with one main difference. Two patient samples (one sample from each group) were selected at random and held out (i.e. leave two-out cross-validation) of the training process. To slow the DE convergence, we set Cr = 0.1. The system comprised of the remaining n-2 patients was trained for 20 DE generations for each holdout pair. These n-2 samples were used for training and for generating training z scores (age combined with grade with the PNN) and Azs. For each DE generation, a bootstrap population was generated from the fixed n-2 population and an Az was generated. The weights that gave the largest Az for the 20 DE generations were used to generate the z scores for the two samples (holdout pair). We used stochastic averaging for the output scores, where 20 bootstrap populations were generated from the fixed n-2 training samples (generated 20 scores for each of the two left out samples). This process cycled (i.e. choosing another pair at random leaving a new n-2 training population for the next 20 DE generations) until all patients received a score. The resulting leave two out cross-validations gave Az = 0.700 (i.e. similar to the Az estimated from the score generation), indicating the approach was internally valid, when generating scores for samples not observed by the system during the training process.

4.4 Brief Statistical Learning Summary

An SL methodology comprised of DE optimization, a kernel mapping, and stochastic ensemble averaging was presented as an illustration to generalize widely used analysis techniques. The technique gives the SL methodology an epidemiologic interpretation. Although a specific example was used in this work, the framework applies to all situations where LR modeling and survival analysis are appropriate. The approach can be easily modified to include as many input variables as required and new samples can be added into the training procedure with the proper clinical feedback indicating the system can learn continually without computer processing demands due to its relative simplicity. The system will require further evaluation with different datasets before it can be applied in practice.

Chapter 5: A System for Automated Measurements and Analysis of Tissue Microarray Image Data

There are two aims of this tissue microarray (TMA) investigation. The first aim is to develop a method of preparing and preprocessing the raw image data to allow automated measurements for the analysis of the multispectral TMA images. The second aim is to investigate whether various automated measures are related to some pre-defined endpoints. These endpoints may be survival, histology type, or tumor-grade for example. That is, the goal is to explore the possibility of capturing different information automatically than what a pathologist distills from tissue sections by inspection. This work analyzes high-throughput TMA arrays stained with routine antibodies used in pathology. This is used as developmental work and a test-bed to address the first aim. In meeting the first aim, a framework or protocol is established for preparing TMA images to investigate novel protein stains as potential biomarkers in the near future. Due to technical limitations, low resolution TMA images were investigated at this time. The intent of this preliminary investigation is to develop methods that scale easily and are applicable to higher resolutions.

5.1 TMA Technology

Traditionally, pathology evaluation was performed with stained whole sections on slides. Tissue microarray technology, digital pathology, and virtual microscopes are newer additions to the pathology toolset and are widely used for research purposes. Briefly, the

TMA construction is initiated by the pathologist's input. Areas of interest are identified on the H&E stained tissue sections by a pathologist. Tissue cores are obtained from the corresponding areas of the originating formalin fixed paraffin embedded tissue block using a semi-automated tissue microarrayer. Multiple tumor tissue cores (i.e. core biopsies using a needle) are often taken to account for possible tumor heterogeneity. Tissue cores are then transferred to a recipient wax block. Each recipient block contains tissue cores from multiple patients, arranged in an array pattern. Sections of the recipient microarray block are cut using a microtome and analyzed using standard histological immunohistochemical stains (see Figure 9).

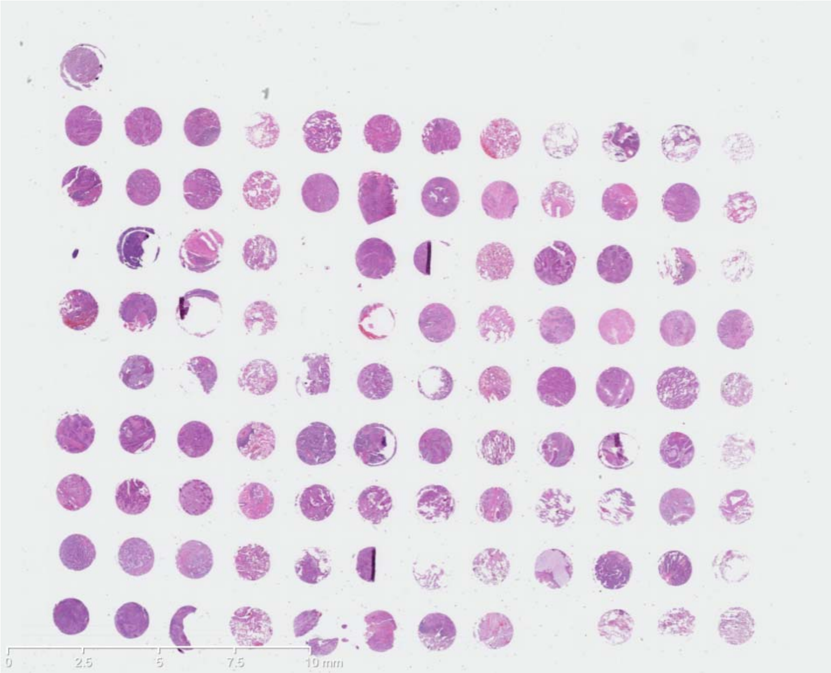


Figure 9: Raw TMA image. Three tumor cores and one normal core were taken from each patient. This array contains cores from 27 patients. Each full row corresponds to three patients (i.e. 12 core samples with four samples per patient). The lone sample at the top left is an arbitrary spotter marker defining the proper orientation of the image.

In this work, the digitized scans of TMA sections were obtained using the Nanozoomer 2.0 HT whole slide scanner (Hamamatsu, USA). The NanoZoomer is capable of high magnification up to 40x objective and can process up to 210 slides at once. The NanoZoomer uses 3-chip TDI (Time Delay Integration) line scanning method, which accurately reproduces sample colors. The line scanning method enables the quick production of high resolution digital slides.

5.2 TMA Image Description and Preprocessing

The TMA arrays were scanned with 20x magnification and digitized with 0.453μm per pixel spatial resolution. A free software (Olympus, Richmond Hill, Ontario) was used to convert the original NDPI format (i.e Hamamatsu format) images to a format (i.e. Tagged Image File Format or TIFF) compatible with our programming language (IDL Version 8.1, ITT Visual Information Solutions, Colorado). The converted images are multispectral data comprised of red, green, and blue (RGB) component lower resolution images each with 8 bits per-pixel dynamic range with 21.1μm per pixel spatial resolution. Each image is 1280 ×1024 pixels in dimension. Each disc (i.e tissue core) is approximately 55-60 pixels in diameter and the disc spacing (separation) is about 30 pixels. Each slide of TMA contains multiple core sections from 27 patients, as compared to the traditional method where a section of tissue from a patient is on a single slide and reviewed by the pathologist. The TMA allows a pathologist to evaluate several patients with a single slide viewing.

To address aim one, operator input was combined with automated processing to prepare or preprocess the TMA images for automated analysis. The operator input serves as a quality control mechanism. In principle, the arrays have four stained sections for each patient most of which have thee malignant tissue sections (samples) and one normal tissue control section. However, not all samples allowed for a normal control tissue sample, leaving four malignant sections for some patients. Moreover, not all sections were viable for automated analysis for various reasons (e.g. missing sections or over contrast). The sections for a given patient are adjacent row-wise (in the horizontal direction) to each other with 12 sections (i.e. cores) per row (three patients) with 9 rows on a given full TMA image array. Our database has seven full TMA images and one partially filled image. The rows do not necessarily run parallel to the image borders (as exemplified in Figure 9) in most images. Samples for some patients were missing due to poor staining and some sections for given patient were missing. The patient labeling that relates to the sections on the image is maintained in a spreadsheet (secondarily to the web-based database described previously). That is, the images do not contain patient identification information. Thus, a correspondence between the image data and patient identification (IDs) must be established.

The first step in connecting the image with the patient was to orient (or check) the images in the proper representation and then inspect the images. The one isolated section on the top left in Figure 9 is a spotter core (i.e an arbitrary sample) indicating that the image is in the proper orientation. The red image (arbitrary) was used for the preprocessing analysis. The layout of a given image was sketched on paper and each section was labeled in accordance with the key spreadsheet containing the patient IDs. A given image was inspected with the aim of finding one viable tumor section per patient, because many of the normal sections were not viable. The first choice was to select a section that was intact (appeared as a uniform disc) without background interference (i.e. without over contrast due to missing tissue). The core sections were counted across each row for a given image, and the viable sections were marked on the sketch.

The next step was automated. The background (i.e. non core sections of the image) was segmented from the tumor section images. Segmentation was achieved with a static pixel value threshold = 220, determined empirically. In this step, all pixel values less than 220 were set to zero. After this step, the tumor sections were separated from background along with some stray smaller sections scattered about. In the next step, a label region algorithm (IDL routine) was applied to the segmented image that sets each contiguous (i.e. blobs) non-zero region within the image to a specific but arbitrary value. Thus, all contiguous sections (all pixels within the section) in a given image now have a unique but arbitrary value.

The following step was semi-automated. An interactive program was developed where the operator views the image and selects the one (pre-selected) section per-patient (using the sketch as guide) by clicking on the section. At this time (after the respective mouse click), the computer program reads the pixel values from the selected region and makes a mask by replacing the original pixel values for the selected region with values equal to the patient's ID (respective IDs were coded into the program). This process is repeated until reaching the last sample on the respective TMA image; all of the non-selected regions are discarded during this process. This results in a mask (shown in Figure 10, left), where each patient's tumor section is labeled with their ID number (all the pixels within the section have the patient's ID number as their value). The mask is then inspected by sampling regions from it and making comparisons with the original array and the sketch. Once developing the protocol described above, this procedure takes about thirty minutes per microarray image. The same mask works for the green and blue images because the three images are spatially aligned.

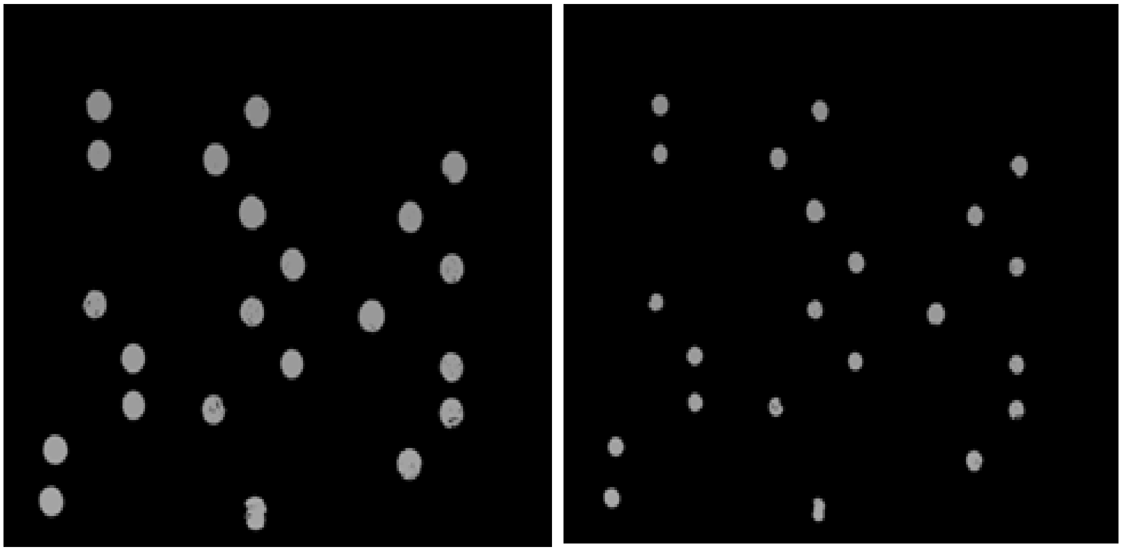


Figure 10: The mask images of the TMA image shown in Figure 9. Each patient's tumor section is labeled with the patient's identification number (all the pixels within the section have the patient's identification number as their value) to form the binary mask (left). To reduce edge effects further, each section was eroded in the binary-mask to account for the filter kernel length (right). These images are over contrasted (all non-zero pixel values set to a constant) for illustration purposes so all samples can be observed simultaneously.

5.3 Automated Image Analysis Methods

When analyzing a given image, the patient-mask is used as a guide to acquire the information for a given patient automatically. This works well for calculations such as averages, standard deviations or pixel differences. Modifications are required when applying filter kernels to analyze textures because of edge effects. Filter kernels of length five were used (described below). To reduce edge effects, the raw microarray image was first filtered with the background intact, because the background was closer to the section intensity level than that of the zero background. To achieve the final selection, two related masks were used. From the patient-ID-mask, a binary mask was formed by converting all non-zero valued pixels to unity giving a binary-mask. To reduce edge effects further, each section was eroded in the binary-mask (contains the selected samples only) to account for the filter kernel length. The final patient labeled mask (shown in Figure 10, right) is given by: *final patient-mask* = *eroded binary mask* × *the patient ID labeled mask*. The raw TMA images (red, green and blue) are then filtered separately. The portion of the filtered image used in the analysis is then given by: *eroded binary mask* × *filtered raw TMA image*. It was noted that the sections in the raw image are spaced about thirty pixels apart indicating that the five element kernels cannot span across two sections simultaneously. For a given filtered section, the standard deviation was used as the summary measure. If kernels of larger extent were used a simple modification is required: one section could be segmented in isolation by embedding (center) it in a sufficiently large zero-background image, apply the filter, and then use the same steps above modified to operate on one section at a time.

The automated image analysis considered measures derived from the raw images such as the average (M), standard deviation (SD), and differences in pixel values between the multispectral images. For reference each pixel has red, green, and blue components which are defined as the vector (x_1, x_2, x_3) = (red pixel value, green pixel value, blue pixel value). Laws texture filters [40] were also applied to these images to assess whether textures are related to the various outcomes. For completeness, the Laws filter set is comprised of five one dimensional kernels. Each kernel has five elements expressed as: L= (1,2,6,4,1). E=(-1.-2,0,2,1), S=(-1,0,2,0,-1), and W=(-1,2,0,-2,1), R=(1,-4,6,-4,1) referred to as level (L) , edge (E) , spot (S), wave (W), and ripple (R), respectively. This naming convention describes the textures captured by each kernel. A half-band filter was also applied defined as B_h = (-1,1). These filters are generically referred to as **f** below and are k × 1 (k = 5 or k= 2) element column vectors. The two-dimension filter kernels are obtained by the outer product of any two filters given they have the same number of elements in this application. The two dimensional kernels are defined as,

$$\mathbf{H} = \mathbf{f}_i \mathbf{f}_j^T , \qquad \text{Eq. (6)}$$

where T indicates transpose and the indices define arbitrary filters (i.e. i = E and j= E, or i = E and j = W]. For this work, we investigated two dimensional filters for i = j only. These filters were applied to the red, green, and blue components for given section independently. For a given binary output (i.e. AC or SQ for example), the T-test is used to compare the measures across the two groups. After applying the filter operation, a given component section was summarized as the standard deviation of the filtered pixel values.

Because the multispectral data is correlated, Principal Component Analysis (PCA) [41] was also applied to each section to reduce the three component images to one image. For a given section, PCA was applied at pixel level and at the summary level, where the summary measures across all samples were included. At the section or pixel level, (x_1, x_2, x_3) were used from a given section as the input. Assume there are m pixels in a given section then the matrix **X** is an m × 3 matrix with the rows defined as $\mathbf{x}_i$= $(x_1, x_2, x_3)_i$ for i =

1 through m rows (i.e. number of pixels within the specific section). The respective column mean is first removed from each column, and the column centered matrix is redefined as **X**. The covariance matrix of **X** is given by

$$\mathbf{C} = \frac{1}{m}\mathbf{X}\mathbf{X}^{T} \qquad \text{Eq. (7)}$$

The Eigenvalues and Eigenvectors of **C** were calculated. The vector with the largest Eigenvalue = $\mathbf{E}_p$ (i.e. the principal component) was used to form a new section from the three component images by the inner product: *new section* = $< \mathbf{x}_{i\,,}\ \mathbf{E}_p>$ for all i. At the summary level, the same analysis applies with the proper relabeling. Now we let m = the number of patient samples and define the vector $\boldsymbol{\sigma}_i$ =($\sigma_{1,}\, \sigma_{2,}\, \sigma_3$,)$_i$ where i is the patient index. The components [i.e. $\sigma_{1,}\, \sigma_{2,}$ and σ_3] represent the standard deviation calculated from a given filter output or from a mixture of filter outputs for a given section. The matrix **X** is formed with the rows defined as $\boldsymbol{\sigma}_i$ and the same analysis is performed.

5.4 Results

5.4.1 Tumor Histology Subtype Analysis

Because the previous work showed a survival advantage for those with AC (n=72) compared with SCC (n=39), the measures between patients with these two histological subtypes were compared. This data subset was developed by considering patients with AC and SCC histology that had full ascertainment for the TMA data. Image examples for AC and SCC are shown in Figure 11. The results from raw image analysis are summarized in Table 16.

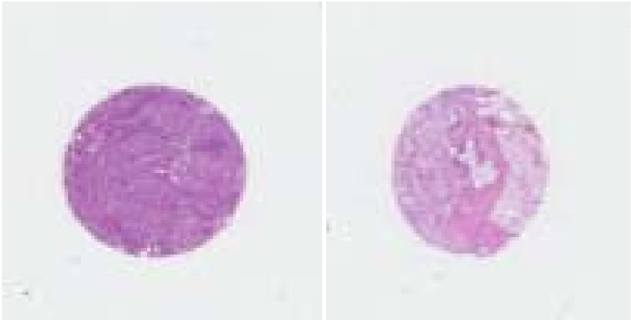


Figure 11: Raw images of SCC and AC. This shows single tumor cores from SCC (left) and AC (right) patients from the TMA.

Table 16: Feature analysis. This gives the p-values from the T-test when comparing patients with AC and SC. From left to right, we compared the mean intensity (M), the standard deviation (SD) and four x-y symmetric Laws filters (i.e. the outer product of a given filter with its transpose) and the x-y symmetric half-band filter. The left column indicates the spectral component image that was filtered.

	M	**SD**	**EE**	**SS**	**WW**	**RR**	**B_hB_h**
x_1	0.2637	0.8047	0.4310	0.1163	0.0430	0.0398	0.0514
x_2	0.3118	0.7715	0.2779	0.0466	0.0184	0.0243	0.0224
x_3	0.3109	0.5260	0.2903	0.0765	0.0382	0.0402	0.0448

The band pass filters from right to left are ordered from high-frequency to low-frequency band-pass with RR and B_hB_h both having similar high-frequency band-pass characteristics. The findings show that the higher frequency filters (WW, RR, and B_hB_h) produce a measure that shows a difference across the two histological types with the x_2

(green) component showing greater differences. The M and SD measures showed no significance from any of the component images or when applying PCA. The differences in x_1-x_2, x_1-x_3, and x_3-x_2, were also investigated and showed no significance with or without PCA (not shown). Examples of related textures are shown in Figure 12. Textures that showed significance were generated synthetically for demonstration by filtering white-noise with the respective filter kernels. A filter is sensitive to the texture it generates (i.e. filter texture reciprocity). When comparing the raw images in Figure 11 with the related textures shown in Figure 12, it is not obvious that there are distinguishable textural differences between the two histology subtypes.

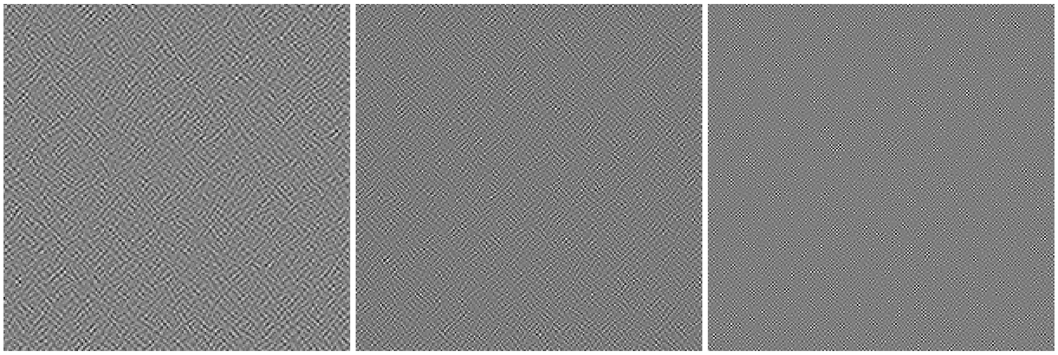


Figure 12: Examples of Laws filters. Laws filters were applied to white-noise fields to demonstrate three textures: SS (left), WW (middle), and RR (right).

To assess whether the measures bring additional information to the process, the x_2 results from WW and RR filtering were used in Cox regression. Using this dataset and dichotomizing by histology subtype gave HR= 1.77 (0.995, 3.139) showing a baseline trend in that those with SC are at elevated hazard. Controlling for WW(x_2) [i.e. the WW x_2 component] with the two tumor types gave HR=1.95, (1.04, 3.65) indicating the WW texture measure at this scale complements the tumor categorical variable. Adding RR

(x_2) with WW(x_2) as controlling factors did not influence findings. Cox regression was performed by dichotomizing WW(x_2) at its median distribution value (65.8) which gave HR=1.91 (1.05, 3.47), showing (i) patients with an above median WW(x_2) measure have a significantly elevated hazard, and (ii) the x_2 measure at the WW scale is related to survival.

Because all three of the WW scale measures showed significance, we restricted the pixel level PCA analysis to this measure and reduced the three images to one image and applied the WW filter which produced little association (p=0.0391) when comparing across histology subtypes. Using Cox regression by dichotomizing at the median, which gave HR = 0.929 (0.522,1.652). We then applied PCA at the summary level and combined the three standard deviation measures derived from WW which separated the two tumor types (p=0.035), but was not better than that provided by WW(x_2) in isolation (see Table 17). We then combined W(x_2), WW(x_3) and RR(x_2) with PCA, which also separated the two groups (p=0.0295), but was not better than provided WW(x_2) in isolation.

In summary, a measure from the TMA green image filtered with the WW or RR Laws two-dimensional filter kernels was related to both histology subtypes and survival.

5.4.2 Tumor-Grade Associations

The analyses discussed above were repeated on the same dataset to compare the groups of patients belonging to the three different categories of tumor-grade (i.e.

grades1-3). Raw tumor image examples for each grade are shown in Figure 13. The results of SS filtering are shown in Table 17. None of the raw pixel measures (SD, M, or spectral component differences) showed significance (not shown) or other filters gave significant results (not shown). In this table for example $x_1(1,2)$ indicates a comparison using grade 1 samples with grade 2 samples from the red image.

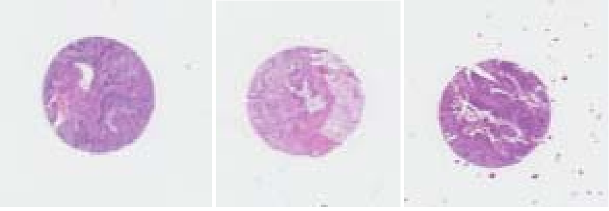


Figure 13: Tumor-grade examples. This shows tumor-grade examples for grade 1 (left), grade 2 (middle), and grade 3 (right) tumors.

Table 17: Texture analysis with the SS filter. This gives the p-values from the T-tests by comparing patients with the various grades with the red (x_1), green (x_2) and blue (x_3) components separately. For example, $x_1(1, 3)$ is comparing grade 1 and 3 in the red component.

Red	$x_1(1, 2)$	$x_1(1, 3)$	$x_1(2, 3)$
p-value	0.22	0.03	0.06
Green	$x_2(1, 2)$	x2(1, 3)	x2(2, 3)
p-value	0.16	0.02	0.08
Blue	$x_3(1, 2)$	$x_3(1, 3)$	$x_3(2, 3)$
p-value	0.25	0.02	0.04

As shown in Table 17, most significant differences were found when comparing grade 1 samples with grade 3 samples in all three components ($P<0.05$). However, the difference between grade 2 and 3 was significant only in the blue component giving p-value= 0.04.

Based on the above results, the measurements of the three SS filtered components (red, green and blue) were modeled with logistic regression (LR) using a two category grade system. Grade 1 and grade 2 samples were combined to form one composite grade as the reference outcome and grade 3 samples formed the other outcome**.** The measures of x_2 (green) and x_3 (blue) components were found to be significantly associated with the grade: x_2 OR = 0.62, (CI: 0.38 0.78) and x_3 OR= 0.61, (CI: 0.36 0.88) per SD increase (SD=23.6 and SD=12.1 respectively). This indicates that a patient with a decreased texture score is more likely in the grade 1-2 group. These results are consistent with the findings from the T-test as shown in Table 17. No association was found with survival outcomes when stratifying by these two modified grade groups (not shown). When comparing the images shown Figure 14, textural differences are not distinguishable by observation. As above, PCA was not productive.

The texture analysis using grade as the endpoint found measures related to a hybrid grading scale. This suggests the stage 1 grade may be better categorized as a two-state problem. It is important to note the most significant grade findings resulted from the blue and green images using the SS scale. In contrast, the most significant tumor-type findings were derived from the WW and RR scales from the green images.

5.5 Brief Image Analysis Summary

In summary, a method of preparing TMA images for automated processing was developed and evaluated. Relationships were established relating various image textures to histology, survival, and tumor-grade indicating the system developed here for automated TMA analysis is a valid approach.

Chapter 6: Discussion

Various aspects of stage I NSCLC were investigated using a diverse dataset with a multifaceted analysis. This evaluation included survival analysis and favorable outcomes modeling using clinical and pathological variables with accepted epidemiologic approaches. Novel repair protein expression DNA measures were investigated as biomarkers for survival. Multivariate models were investigated to determine measures related to recurrence. A method was developed to integrate kernel based statistical learning methods with accepted epidemiologic practice, essentially fusing the strengths of both approaches. A system was developed to analyze tissue microarrays with automated image processing methods. The various methodologies and findings are summarized below.

6.1 Baseline Clinical and Pathological Variable Evaluation and Modeling Strategies

For patients with stage I NSCLC, the favorable outcome modeling showed that younger age and female gender were associated with a favorable survival. These findings were also observed previously [2, 15, 42]. This analysis demonstrated that the histological subtype in combination with gender and recurrence provided the greatest predictive value for a subset of patients. The four category histology-subtypes, tumor-location, and smoking status were not significant. We also found that increasing age was related to the SCC histology subtype, and increasing tumor-grade was related to male gender. The work also showed that none of the clinical or pathological factors (described in Chapter 2) were related to recurrence.

For the favorable outcomes analysis, a method was established of dichotomizing the patient population by incident and censored group membership [19], which is a different approach than taken by other researchers [22]. This approach may be better suited for limited datasets because it does not require discarding data (see Chapter 2). The approach can also serve as intermediate step to find variables that may be related to survival as evaluated with Cox-regression or Kaplan Meier analysis (see Chapter 2 and Chapter 4).

Notwithstanding the dichotomization approach, it could be argued that the LR favorable modeling was suboptimal because the time-to-event variable resolution was reduced to a coarser dichotomous variable. However for a specific set of variables, the LR output provides a different metric (i.e. probability of having a favorable outcome) than that provided by Cox regression (i.e. instantaneous relative risk). Moreover, Kaplan-Meier analysis provided a population level evaluation and is therefore not suited for individual predictions. Thus, the resolution reduction is the price paid for an alternative output. The LR modeling (or more generally binary classification) may be more suitable for clinical based predictions at the patient level because the output is easily interpretable.

The survival analysis (Kaplan Meier and Cox regression) of the standard clinical and pathological variables (Chapter 2) clearly showed that younger age, AC histology subtype, negative for disease recurrence, and female gender confer longer survival. Differences within stage I subgroups were not directly related to survival. However, when considering patients with certain clinical and pathological factors with stage IA, their survival prospects were better. Younger age patients with both the AC histology subtype and Stage IA disease have better survival outcomes. This work showed that tumor-grade was not a significant variable for influencing these patients' survival outcome when using accepted approaches, which is in contrast with other findings [2] that also used accepted approaches. The increased hazard for SCC patients in comparison with AC, male gender, and increasing age were significantly greater than those found in related work [2], which may be due to either population or timeframe differences. The survival findings with adjuvant therapy were consistent with a recent meta-analysis that documented an increased hazard ratio for adjuvant chemotherapy in patients with stage IA NSCLC [43], although the findings in this dissertation showed only a trend.

The fact that adjuvant chemotherapy did not influence survival in this cohort of patients is not surprising. In randomized studies for early stage NSCLC, adjuvant chemotherapy improved survival for patients with stage II and IIIA disease [44]. For patients with stage IA disease, a meta-analysis demonstrated a hazard ratio of 1.4 with adjuvant chemotherapy, suggesting that this group did not benefit. For patients with stage IB disease, the available evidence indicates that patients with tumor size greater than 4 cm might benefit from cisplatin-based chemotherapy.

6.2 Novel Biomarkers

With the advent of molecular and genomic approaches in cancer research, there has been increased interest in the study of protein biomarkers in clinical tumor tissues [45]. Genomic instability is a significant attribute of cancer cells that facilitates tumor progression [46]. DNA repair capacity has been shown to be a prognostic factor in NSCLC patients with resected tumors. Patients with tumors that showed a high ERCC1 expression (i.e. a repair mechanism) had a more favorable prognosis and therefore did not benefit from adjuvant cisplatin-based chemotherapy [23]. Other proteins involved in DNA damage repair mechanisms such as Ku86 and PARP have not been studied to any extent in NSCLC. Thus, it was hypothesized that these repair mechanisms may be related to survival or recurrence. To the contrary as shown in Chapter 3, PARP and Ku86 expression did not provide significant associations with favorable outcome, survival, associations with other variables, or recurrence with the exception that Ku86 expression showed a significant association with male gender. No prognostic significance was found for these markers in this study population, perhaps because of the retrospective nature of the study. Nonetheless the observed association between Ku86 expression and patient gender can inform testable hypotheses to be evaluated using larger sample sizes and an appropriate matched control group in a prospective setting. It is notable, agents that modulate PARP and the homologous recombination repair are currently in clinical studies. Discovering predictive biomarkers for these agents requires a thorough characterization of expression of the target in the tumor tissue and other related pathway markers. For this reason, the results of this study are relevant for drug development efforts with agents that modulate DNA damage repair pathways.

In summary, the novel biomarker analysis resulted in important negative outcome experiments. The DNA repair measures were modeled with the clinical and pathological factors and no significant relationships with survival were found. Moreover, none of these factors or combination of factors provided significant associations with recurrence. The prognostic value of recurrence in predicting favorable (or unfavorable) outcome or survivability may be limited in general because about 64% of the incident and 92% of censored group patients were negative for recurrence (or unknown for the censored group). However, positive recurrence is a sure indicator of an unfavorable outcome. A better understanding of those factors that can predict recurrence is required.

6.3 Statistical Learning and Epidemiology

Cox regression, Kaplan-Meier analysis, and logistic regression (LR) are important and widely used epidemiologic techniques. These methods are readily available in many software packages and they provide important epidemiologic interpretations. Cox regression and LR modeling are parametric and by nature assume a specific output and input variable covariate forms. The forms can be changed by user imposition. On the other hand, kernel based statistical learning methods assume no specific data form and can operate on small datasets with complicated relationships between the covariates and output. However, SL approaches do not provide a readily interpretable epidemiologic output and they can require more specialized training techniques that are often not commercially available.

A technique for incorporating SL methods with epidemiologic analyses was developed and evaluated in Chapter 4 [19]. The approach used ensemble averaging with bootstrap sampling [39] to overcome data limitations. Differential evolution was developed to determine the kernel weights. This is an evolutionary processing technique for global optimization problems based on the work by Storn and Price [47]. Age and tumor-grade were combined with the modified PNN and were used as inputs to logistic regression and survival analysis (i.e. a hybrid approach). This hybrid approach was compared with the accepted methods of using these raw clinical variables as inputs. These findings indicate the hybrid approach provided greater Az in the logistic regression modeling and greater hazard relationships in the survival analyses than that of the accepted approaches using the respective variables. In contrast with the findings discussed in Chapter 2, grade was related to survival outcome when combined with age. The internal validity of these findings are supported by the cross-validation analysis and the simulation methods discussed in the Chapter 4. This approach represents a framework that is easily generalized.

The SL output was used as the input into LR model and survival analysis. This approach combines the strengths of SL and accepted epidemiologic practice. In this capacity, the SL device was operating as frontend preprocessing step for these accepted analysis techniques. Processing the SL output with these approaches provides a mechanism for converting the SL output into epidemiologic metrics, such as ORs and HRs. We used a relatively simple SL device by converting the PPN classifier [29] to give a patient-score to demonstrate the concept with a two-class probability problem. This specific approach can be extended to include more than two classes (e.g. death, greater than three, and five year survival benchmarks). The PNN applies to multiclass problems, as well, and multinomial logistic regression can address multiple level outcomes. The PNN classifier was converted to provide an output score for each patient, which is a different application. For example, the PNN has been used as a classifier in other types of survival and medical research [48, 49].

The approach presented in Chapter 4 based on the PNN represents a simplifying step to demonstrate the main principles. There are more sophisticated methods that can be used within this SL-epidemiologic framework. More generally, the same hybrid approach is applicable for the output of any other type of SL method or decision device, such as support vector machines, kernel based partial least squares, or other types of neural networks [50-54]. Thus, any combination technique can be used in place of the modified PNN shown in the schema in Chapter 4.

Generalizations of the LR model, incorporating kernel based techniques, and neural networks into epidemiologic studies represents a diverse field of inquiry. Neural networks have been adapted to survival analysis by predicting survival time intervals for intraocular melanoma [52]. Earlier research used a PNN and LR modeling to predict survival in early stage NSCLC but did not fuse the models [55]. Logistic regression is a member of a family of generalized linear models. Replacing the LR argument with various forms of smooth functions has provided benefits in the study of colon-cancer [56], heart-disease [57] and infant mortality [58]. Other researchers have incorporated univariate kernel density estimations for studying prostate-cancer [59], health disparities [60], and nutrient intake [61]. Similarly, univariate kernel density estimations have been used to estimate summary measures that were incorporated into LR modeling in fast-food consumption studies [62]. Our work differs from this other work in that the PNN (or kernel mapping) application makes no assumption concerning the functional relationship of the variables under study and we incorporated the measures into LR. Many of the medical uses of neural networks are reviewed elsewhere [63].

6.4 Image Processing and TMA Data

Tissue microarrays (TMAs) are emerging as an important research tool. These arrays allow for the analysis of DNA, RNA, and protein expression on a large number of clinical samples (i.e. tumor samples) simultaneously [64] using laboratory assays such as IHC or in-situ hybridization [65]. The use of TMAs in comparison with standard whole tumor sections on slides offers several benefits and major savings in terms scientific resources, such as use of laboratory reagents, technician effort, pathologist effort etc. Moreover, recent research with TMAs has led to advancements in quantifying various biomarkers [66]. The cost for these expanded datasets is that the quality of data in the TMA images may be degraded compared to conventional pathology slides, as discussed below. Similarly, different forms of analysis may be required to extract the information from these expanded dataset. The manufactures of the TMA imaging equipment, as well as other developers, also provide software for image analyses that is often operator-guided. For example, work by Behera et al [67] quantified IHC stained TMAs with an operator-guided approach (Aperio, Vista, CA) that resulted in associations between tumor-nuclei with grade and histology. The novel biomarker analysis presented in Chapter 3 also used operator-guidance for the measurements. As discussed in this review [64] scoring methods for TMA images are often subjective, although automated methods are under development. For example, recent work in lung carcinoma showed that an automated quantitative scoring system for TMAs was significantly correlated with the pathologist scoring [68].

In this dissertation, a different approach was taken to analyze TMA images. A customized method was developed to first preprocess the raw images to allow subsequent automated measurements from the entire dataset simultaneously without operator imposition or discretion. The preprocessing is essentially a quality control (QC) procedure that requires operator-input. Various automated image measures were evaluated. Significant relationships were found with tumor type, survival, and tumor-grade resulting from various Laws filters that capture specific texture characteristics. The WW (wave) and RR (ripple) filters showed the strongest relationships with the survival when applied to the green component image. This showed that those with increased filter-scores were at a significantly elevated hazard. When investigating tumor-grade, the SS (spot) filter applied to either the blue or green component images showed a significant relationship between grade 3 patients compared with the combined group of grade 1 and 2 patients; the blue component image provided the greater associations. These findings suggest that grade may be better categorized as either a two-state problem or perhaps as a continuous variable instead of three states for stage 1 patients. Although grade plays an important role in many tumor types, its prognostic significance in lung AC has not been established [69]. This is a significant deficiency because the majority of NSCLC patients have AC. In general, there is higher degree of pathology *concordance* for better differentiated tumors and there are no clear standards for describing moderate and poor differentiation for NSCLC among pathologists [70]. These texture findings are consistent with the known uncertainty in NSCLC grading. The various relationships were found at lower resolution and therefore represent preliminary findings. We hypothesize that stronger relationships will result when analyzing these images at higher spatial resolutions.

6.5 Limitations of the Study

There are several limitations associated with our findings, the most significant of which is the retrospective data collection and limited number of patient samples. This limitation resulted in incomplete case ascertainment such that the analytic data samples differed between various evaluations. In the parametric LR modeling, we did not consider interaction terms in the favorable outcome analysis to limit over-fitting and the presentation length. We were able to construct logistic regression models with increased predictive capability by limiting the work to two histology-subtypes, which limits the model's applicability. The recurrence variable was unknown for many of the censored patients, suggesting the related findings are preliminary. The adjuvant treatment findings should also be qualified because the specific treatment type and regimen were unknown.

We dichotomized the favorable outcome analysis by censored or incident group membership, representing a novel separation methodology [19] that will require further evaluation. This approach reduces the uncertainty in the status for those patients that did not survive but there are likely patients in the unfavorable group that survived past some censored group patients. If we assume that the censored group patients did not expire the day after losing study contact, their censored time is a conservative estimate (i.e. left-limit) of their overall survival time, indicating that the time separation (mean censored and incident times) between the two groups is greater than that specified by the separation of the censored and incident group means. This indicates that associations found (ORs or Az) are more likely conservative estimates.   Another approach [22] (that could be considered at the standard approach to dichotomizing time-to event data) is to use a survival time cut-point to dichotomize the patient population (i.e. no possibility of overlapping survival time). This approach cannot accommodate censored patients on the left side of the cut-point (i.e. censored patients are discarded), which is not practical for limited datasets. This approach may create an artificial separation when there are many samples near the boundary. In our initial analyses (not shown), the standard approach using various survival time cut-points was considered, but was not productive with this dataset. The generality of our approach will require further evaluation with different datasets.

Although DE is a robust approach, there is no guarantee that it will converge indicating that the findings may be less than optimal. The generation termination limits were empirically set. Because, we found that letting the process evolve over many generations produced weights that were too finely tuned and did not provide performance consistency between the training evaluation and the final score assessments. Because the dataset was limited, further evaluation using both simulation methods and holdout cross-validation with the patient-score was also provided in chapter 4. The findings from the hybrid modeling will require further evaluations with different datasets to show generalization. In principle to use the system developed in Chapter 4 in practice, the sampled patient population should be representative of stage 1 lung cancer patients in general.

While the TMAs play an important role in cancer research and are widely used for biomarker studies, they are not routinely used in clinical laboratory testing [65]. One of the major concerns associated with TMAs are the small size of the tissue cores and that they may not accurately represent the score obtained from a whole section, due to the heterogeneous nature of tumor tissue [71, 72]. Hence, multiple cores are usually extracted per block, taking the tumor heterogeneity into account. Several studies have validated the use of TMA in various tumor sites [73-75] and have demonstrated concordance between biomarker scores from TMA and whole sections. The TMA technology requires sampling of the tumor tissue at regions containing large amounts of cancer cells. These regions of interest are manually selected by visual assessment of histology slide images by expert pathologists. Because these methods are new, standard automated protocols have not yet been developed to identify the regions of interest in the tumor tissue [66]. Hence, our analyses of the TMA data provided here were experimental in nature and performed with low spatial resolution data. Related work in whole slide image analysis has been performed by other researchers [76, 77].

6.6 Conclusions

The work produced several important findings. Baseline survival characteristic were estimated for a southeastern contemporary population. This is important because lung cancer survival patterns differ regionally and serially. In addition, these measures can serve as a baseline to assess whether new measures bring additional information to the analysis as in Chapter 3. The DNA repair expression variables were not related to survival, although some associations with SCC and gender in models that included Ku86 score were found [14]. The worked showed the significance of recurrence limiting survival, which was known [78, 79]. However, none of the clinical or pathological variables, novel biomarkers, or combinations of these measures were related to recurrence, indicating more work is required to find measured related to recurrence. A platform was established and evaluated to integrate SL techniques with accepted epidemiologic practice. In conjunction, a technique was developed as an efficiency step to assist in finding variables that may be related to survival via Cox regression or Kaplan Meier analysis. A system was developed to analyze TMA images with automated measures that first requires operator-input as a QC procedure. Once developing the QC system, it takes about 30 minutes to prepare one microarray image for automated processing. This system was evaluated with TMA images resulting from standard stains. Measures related to survival, tumor-type, and tumor grade were established with this system.

Standard statistical methods and experimental methods were developed to analyze the clinical and pathological characteristics of this cohort. Several relationships related to survival and tumor characteristics were quantified. These findings may have importance to determine optimal therapy and level of aggression required to manage stage I NSCLC, specifically. Future work includes (i) investigating novel stains and biomarkers with the system developed in Chapter 5 using higher resolution images and extending the SL methods developed and evaluated in Chapter 4 to build more general models, (ii) applying the methods developed here to our parallel work in lung cancer maintenance therapy [7] and the analysis of SCLC [10] as well, and (iii) the continued development and expansion of the web-based database in support of our scientific aims [20]. The longer-term goal is to develop models for use at the individual prediction level for patients with all forms of lung cancer and for understanding lung cancer etiology.

References Cited

1. Manser RL, Irving LB, Byrnes G, Abramson MJ, Stone CA, Campbell DA., Screening for lung cancer: a systematic review and meta-analysis of controlled trials. Thorax, 2003. **58**(9): p. 784-9.

2. Ou SH, Zell JA, Ziogas A, Anton-Culver H., Prognostic factors for survival of stage I nonsmall cell lung cancer patients : a population-based analysis of 19,702 stage I patients in the California Cancer Registry from 1989 to 2003. Cancer, 2007. **110**(7): p. 1532-41.

3. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AAI., Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. Mayo Clinic Proceedings, 2008. **83**(5): p. 584-594.

4. Sher T, Dy GK, Adjei AA., Small Cell Lung Cancer. Mayo Clinic Proceedings, 2008. **83**(3): p. 355-367.

5. American Cancer society: Lung Cancer - Non-Small Cell. http://www.cancer.org/Cancer/LungCancer-Non-SmallCell/index?ssSourceSiteId=null.

6. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD., Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England Journal of Medicine, 2011. **365**(5): p. 395-409.

7. M. Behera, T.K.Owonikoko, Z. Chen, S. A. Kono, F. R. Khuri, C. P. Belani, S. S. Ramalingam. Single-agent maintenance therapy for advanced-stage non-small cell lung cancer: A meta-analysis. in ASCO Annual Meeting. 2011. Chicago: J Clin Oncol 29: 2011 (suppl; abstr 7553), In review for publication in Lung Cancer.

8. T. K. Owonikoko, M.Behera, H. N. Tran, Z. Chen, R. P. Chowdry, N. F. Saba, S. S. Ramalingam and F. R. Khuri. Systematic comparative analysis of efficacy of EGFR tyrosine kinase inhibitors (TKIs) in the frontline versus salvage therapy of NSCLC. in ASCO Annual Meeting. 2011: J.Clin Oncol Vol 29, No 15_suppl (May 20 Supplement), 2011: 7602.

9. T. K. Owonikoko, S.S.Ramalingam, M. Behera, J. C. Brandes, N. F. Saba, C. Bhimani, S. Harichand-Herdt, D. M. Shin, F. R. Khuri, C. Ragin. Survival impact of newly approved therapeutic agents in patients with advanced non-small cell lung cancer (NSCLC): A SEER-Medicare database analysis. in ASCO Annual Meeting. 2010 Chicago: J Clin Oncol 28:15s, 2010 (suppl; abstr 7633).

10. Taofeek K. Owonikoko, Madhusmita Behera, Zhengjia Chen,Chandar Bhimani, Walter J. Curran, Fadlo R. Khuri,Suresh S. Ramalingam, A systematic analysis of efficacy of second line chemotherapy in sensitive and refractory small cell lung cancer. Journal of Thoracic Oncology, accepted January 2012, in publication; Presented at the 2010 Chicago Multidisciplinary Symposium in Thoracic Oncology. Chicago December 9-11, 2010, 2012.

11. Bach, PB., Inconsistencies in findings from the early lung cancer action project studies of lung cancer screening. Journal of the National Cancer Institute, 2011. **103**(13): p. 1002-6.

12. Infante M, Lutman FR, Cavuto S, Brambilla G, Chiesa G, Passera E, Angeli E, Chiarenza M, Aranzulla G, Cariboni U, Alloisio M, Incarbone M, Testori A, Destro A, Cappuzzo F, Roncalli M, Santoro A, Ravasi G; DANTE Study Group., Lung cancer screening with spiral CT: baseline results of the randomized DANTE trial. Lung cancer, 2008. **59**(3): p. 355-63.

13. Pastorino U, Bellomi M, Landoni C, De Fiori E, Arnaldi P, Picchio M, Pelosi G, Boyle P, Fazio F., Early lung-cancer detection with spiral CT and positron emission tomography in heavy smokers: 2-year results. Lancet, 2003. **362**(9384): p. 593-7.

14. Madhusmita Behera, John J.Heine, Gabriel L. Sica, Erin E. Fowler, Ha Tran, Robert W. Fu, Anthony A. Gal, Robert Hermann, William Mayfield, Fadlo R. Khuri, Taofeek K. Owonikoko, Suresh S. Ramalingam., Survival analysis of stage I NSCLC patients using clinical and DNA-repair pathway expression variables In review for publication in Clinical Lung Cancer, 2012.

15. Montesinos J, Bare M, Dalmau E, Saigi E, Villace P, Nogue M, Angel Segui M, Arnau A, Bonfill X., The changing pattern of non-small cell lung cancer between the 90 and 2000 decades. The open respiratory medicine journal, 2011. **5**: p. 24-30.

16. Devesa SS, Bray F, Vizcaino AP, Parkin DM., International lung cancer trends by histologic type: male:female differences diminishing and adenocarcinoma rates rising. International journal of cancer. Journal international du cancer, 2005. **117**(2): p. 294-9.

17. TL Fairley, PhD, E Tai, MD, JS Townsend, MS, SL Stewart, PhD, CB Steele, DO, Div of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion; SP Davis, PhD, Office on Smoking and Health, National Center for Chronic Disease Prevention and Health Promotion; JM Underwood, PhD, EIS Officer, CDC., Racial/Ethnic Disparities and Geographic Differences in Lung Cancer Incidence — 38 States and the District of Columbia, 1998–2006. Morbidity & Mortality Weekly Report, 2010. **59**(44): p. 1434-1438.

18. Non-small cell lung cancer survival rates by stage;. in American Cancer Society http://www.cancer.org/Cancer/LungCancer-Non-SmallCell/DetailedGuide/non-small-cell-lung-cancer-survival-rates. 2012.

19. Behera M, Fowler EE, Owonikoko TK, Land WH, Mayfield W, Chen Z, Khuri FR, Ramalingam SS, Heine JJ., Statistical learning methods as a preprocessing step for survival analysis: evaluation of concept using lung cancer data. BioMedical Engineering OnLine, 2011. **10**(1): p. 97.

20. Madhusmita Behera, Haibin Wang, Erik Bouzyk, Taofeek K. Owonikoko, Robert Hermann, William Mayfield, Mourad Tighiouart, Anthony Gal, Fadlo R. Khuri, and Suresh S. Ramalingam. Winship Cancer Institute of Emory University Lung Cancer Database (WILCAD). in AMIA Annual Symposium. 2010. Washington, DC.

21. A.E.Smith, S.S. Anand., Patient survival estimation with multiple attributes: Adaptation of Cox's regression to give an individual's point prediction. in in 5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2000) A workshop at the 14th European Conference on Artificial Intelligence(ECAI-2000). 2001. Berlin, Germany.

22. Eschrich S, Yang I, Bloom G, Kwong KY, Boulware D, Cantor A, Coppola D, Kruhøffer M, Aaltonen L, Orntoft TF, Quackenbush J, Yeatman TJ., Molecular Staging for Survival Prediction of Colorectal Cancer Patients. Journal of Clinical Oncology, 2005. **23**(15): p. 3526-3535.

23. Olaussen KA, Dunant A, Fouret P, Brambilla E, André F, Haddad V, Taranchon E, Filipits M, Pirker R, Popper HH, Stahel R, Sabatier L, Pignon JP, Tursz T, Le Chevalier T, Soria JC., DNA Repair by ERCC1 in Non–Small-Cell Lung Cancer and Cisplatin-Based Adjuvant Chemotherapy. New England Journal of Medicine, 2006. **355**(10): p. 983-991.

24. Patz EF Jr, Swensen SJ, Herndon JE 2nd., Estimate of lung cancer mortality from low-dose spiral computed tomography screening trials: implications for current mass screening recommendations. Journal of Clinical Oncology, 2004. **22**(11): p. 2202-6.

25. Pisters KM, Le Chevalier T., Adjuvant chemotherapy in completely resected non-small-cell lung cancer. Journal of Clinical Oncology, 2005. **23**(14): p. 3270-8.

26. Vapnik, V.N., The Nature of Statistical Learning Theory. Second ed. 2000, NY: Springer.

27. Vapnik, V.N., Statistical Learning Theory 1998, NY: John Wiley & Sons, Inc.

28. Shawe-Taylor, J. and Cristianini, N., Kernel Methods for Pattern Analysis. 2004, Cambridge, UK Cambridge University Press.

29. Specht, DF., Probabilistic neural networks. Neural Networks, 1990. **3**: p. 109-118.

30. Parzen, E., On estimation of a probability density function and mode. Annals of Mathematical Statistics, 1962. **33**(3): p. 1065-1076.

31. Cacoullos, T., Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics, 1966. **18**(1): p. 179-189.

32. K. V. Price, R. M. Storn, and J. A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization 2005, Heidelberg: Springer

33. J.J.Heine, W.H. Land, and K.M. Egan, Statistical learning techniques applied to epidemiology: a simulated case-control comparison study with logistic regression. BMC Bioinformatics, 2011. **12**: p. 37.

34. Hosmer, D.W. and Lemeshow,S., Applied Logistic Regression second ed. 2000, NY: John Wiley & Sons, Inc.

35. Hanley, J.A. and McNeil, B.J., The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982. **143**(1): p. 29-36.

36. Pencina, M.J. and D'Agostino, R.B., Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in medicine, 2004. **23**(13): p. 2109-23.

37. Mercer, J., Functions of positive and negative type, and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1909. **209**: p. 415-446.

38. Land, W.H.Jr., Margolis, D., Kallergi, M., Heine, J. J., A Kernel Approach for Ensemble Decision Combinations with two-view Mammography Applications. International Journal of Functional Genomics and Personalised Medicine 2010. **3**(2): p. 157-182.

39. Efron, B. and Tibshirani, R.J., An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57. 1993, Boca Raton, FL: Chapman & Hall.

40. Laws, K., Textured Image Segmentation. 1980, Ph.D Dissertation, University of Southern California.

41. Jolliffe, I.T., Principal Component Analysis. 2nd ed., 2002 ed. 1986, Springer Series in Statistics; 2nd ed., 2002.

42. Albain KS, Crowley JJ, LeBlanc M, Livingston RB., Survival determinants in extensive-stage non-small-cell lung cancer: the Southwest Oncology Group experience. Journal of Clinical Oncology, 1991. **9**(9): p. 1618-26.

43. Früh M, Rolland E, Pignon JP, Seymour L, Ding K, Tribodet H, Winton T, Le Chevalier T, Scagliotti GV, Douillard JY, Spiro S, Shepherd FA., Pooled analysis of the effect of age on adjuvant cisplatin-based chemotherapy for completely resected non-small-cell lung cancer. Journal of Clinical Oncology, 2008. **26**(21): p. 3573-81.

44. Pignon JP, Tribodet H, Scagliotti GV, Douillard JY, Shepherd FA, Stephens RJ, Dunant A, Torri V, Rosell R, Seymour L, Spiro SG, Rolland E, Fossati R, Aubert D, Ding K, Waller D, Le Chevalier T., Lung Adjuvant Cisplatin Evaluation: A Pooled Analysis by the LACE Collaborative Group. Journal of Clinical Oncology, 2008. **26**(21): p. 3552-3559.

45. Warren MV, Chan WY, Ridley JM., Analysis of protein biomarkers in human clinical tumor samples: critical aspects to success from tissue acquisition to analysis. Biomarkers in Medicine, 2011. **5**(2): p. 227-248.

46. Kennedy RD, D'Andrea AD., DNA Repair Pathways in Clinical Practice: Lessons From Pediatric Cancer Susceptibility Syndromes. Journal of Clinical Oncology, 2006. **24**(23): p. 3799-3808.

47. Storn, R. and Price, K., Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization 1997. **11**: p. 341-359.

48. Johansen R, Jensen LR, Rydland J, Goa PE, Kvistad KA, Bathen TF, Axelson DE, Lundgren S, Gribbestad IS., Predicting survival and early clinical response to primary chemotherapy for patients with locally advanced breast cancer using DCE-MRI. Journal of Magnetic Resonance Imaging, 2009. **29**(6): p. 1300-1307.

49. Xu R, Cai X, Wunsch D., Gene Expression Data for DLBCL Cancer Survival Prediction with A Combination of Machine Learning Technologies. in Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. 2005.

50. Maglogiannis I, Zafiropoulos E and Anagnostopoulos I., An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Applied Intelligence, 2009. **30**(1): p. 24-36.

51. Mihir Sewak, Priyanka Vaidya, Chien-Chung Chan., SVM Approach to Breast Cancer Classification. in Computer and Computational Sciences, 2007. IMSCCS 2007. Second International Multi-Symposiums on. 2007.

52. Azzam, F.G.T., Anthony, C.F., and Bertil, E.D., Modelling survival after treatment of intraocular melanoma using artificial neural networks and Bayes theorem. Physics in Medicine and Biology, 2004. **49**(1): p. 87.

53. Ding, B. and Gentleman, R., Classification Using Generalized Partial Least Squares. Journal of Computational and Graphical Statistics, 2005. **14**(2): p. 280-298.

54. Ablitt NA, Jianxin G, Keegan J, Stegger L, Firmin DN, Guang-Zhong Y., Predictive cardiac motion modeling and correction with partial least squares regression. Medical Imaging, IEEE Transactions on, 2004. **23**(10): p. 1315-1324.

55. Marchevsky AM, Patel S, Wiley KJ, Stephenson MA, Gondo M, Brown RW, Yi ES, Benedict WF, Anton RC, Cagle PT., Artificial neural networks and logistic regression as tools for prediction of survival in patients with Stages I and II non-small cell lung cancer. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc, 1998. **11**(7): p. 618-25.

56. Zhao LP, Kristal AR, White E., Estimating relative risk functions in case-control studies using a nonparametric logistic regression. American journal of epidemiology, 1996. **144**(6): p. 598-609.

57. Abrahamowicz M, du Berger R, Grover SA., Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. American journal of epidemiology, 1997. **145**(8): p. 714-29.

58. Gage TB, Fang F, O'Neill E, Stratton H., Maternal age and infant mortality: a test of the Wilcox-Russell hypothesis. American journal of epidemiology, 2009. **169**(3): p. 294-303.

59. Savage CJ, Lilja H, Cronin AM, Ulmert D, Vickers AJ., Empirical estimates of the lead time distribution for prostate cancer based on two independent representative cohorts of men not subject to prostate-specific antigen screening. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2010. **19**(5): p. 1201-7.

60. Osypuk, T.L. and Acevedo-Garcia, D., Are racial disparities in preterm birth larger in hypersegregated areas? American journal of epidemiology, 2008. **167**(11): p. 1295-304.

61. Vercambre MN, Fournier A, Boutron-Ruault MC, Clavel-Chapelon F, Ringa V, Berr C., Differential dietary nutrient intake according to hormone replacement therapy use: an underestimated confounding factor in epidemiologic studies? American journal of epidemiology, 2007. **166**(12): p. 1451-60.

62. Moore LV, Diez Roux AV, Nettleton JA, Jacobs DR, Franco M., Fast-food consumption, diet quality, and neighborhood exposure to fast food: the multi-ethnic study of atherosclerosis. American journal of epidemiology, 2009. **170**(1): p. 29-36.

63. Lisboa PJ, Taktak AF., The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks, 2006. **19**(4): p. 408-415.

64. Hassan S, Ferrario C, Mamo A, Basik M., Tissue microarrays: emerging standard for biomarker validation. Current Opinion in Biotechnology, 2008. **19**(1): p. 19-25.

65. Voduc D, Kenney C, Nielsen TO., Tissue Microarrays in Clinical Oncology. Seminars in radiation oncology, 2008. **18**(2): p. 89-97.

66. Karacali, B. and Tozeren, A., Automated detection of regions of interest for tissue microarray experiments: an image texture analysis. BMC Medical Imaging, 2007. **7**(1): p. 2.

67. Behera M, Park Y, Chisolm CS, Cooper LAD, Chen Z, Fu RW, Sica G, Gal AA, Owonikoko TK, Khuri FR, Marcus AI, Brat DJ, Zhou W, Ramalingam SS. Quantitative digital analysis of LKB1 biomarker in non-small cell lung cancer (NSCLC) patients. in Bio-IT World Conference & Expo. 2011. Boston, MA.

68. Wang CW., Fast quantification of immunohistochemistry tissue microarrays in lung carcinoma. Computer Methods in Biomechanics and Biomedical Engineering, 2011: p. 1-10.

69. Barletta JA, Yeap BY, Chirieac LR., Prognostic significance of grading in lung adenocarcinoma. Cancer, 2010. **116**(3): p. 659-669.

70. http://cancergrace.org/lung/tag/tumour-grade/.

71. Bubendorf L, Nocito A, Moch H, Sauter G., Tissue microarray (TMA) technology: miniaturized pathology archives for high-throughput in situ studies. The Journal of Pathology, 2001. **195**(1): p. 72-79.

72. Horvath L. and Henshall S.,The Application of Tissue Microarrays To Cancer Research. Pathology - Journal of the RCPA, 2001. **33**(2): p. 125-129 10.1080/003130201200338791.

73. Leversha MA, Fielding P, Watson S, Gosney JR, Field JK., Expression of p53, pRB, and p16 in lung tumours: a validation study on tissue microarrays. The Journal of Pathology, 2003. **200**(5): p. 610-619.

74. Torhorst J, Bucher C, Kononen J, Haas P, Zuber M, Köchli OR, Mross F, Dieterich H, Moch H, Mihatsch M, Kallioniemi OP, Sauter G., Tissue Microarrays for Rapid Linking of Molecular Changes to Clinical Endpoints. The American Journal of Pathology, 2001. **159**(6): p. 2249-2256.

75. Jourdan F, Sebbagh N, Comperat E, Mourra N, Flahault A, Olschwang S, Duval A, Hamelin R, Flejou JF., Tissue microarray technology: validation in colorectal carcinoma and analysis of p53, hMLH1, and hMSH2 immunohistochemical expression. Virchows Archiv, 2003. **443**(2): p. 115-121.

76. Cooper LAD, Kong J, Gutman DA, Wang F, Cholleti SR, Pan TC, Widener PM, Sharma A, Mikkelsen T, Flanders AE, Rubin DL, Van Meir EG, Kurc TM, Moreno CS, Brat DJ, Saltz JH., An Integrative Approach for In Silico Glioma Research. Biomedical Engineering, IEEE Transactions on, 2010. **57**(10): p. 2617-2621.

77. Kong J, Sertel O, Shimada H, Boyer KL, Saltz JH, Gurcan MN., Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. Pattern Recognition, 2009. **42**(6): p. 1080-1092.

78. Harpole DH Jr, Herndon JE 2nd, Young WG Jr, Wolfe WG, Sabiston DC Jr., Stage I nonsmall cell lung cancer. A multivariate analysis of treatment methods and patterns of recurrence. Cancer, 1995. **76**(5): p. 787-796.

79. Rena O, Papalia E, Ruffini E, Casadio C, Filosso PL, Oliaro A, Maggi G., Stage I pure bronchioloalveolar carcinoma: recurrences, survival and comparison with adenocarcinoma of the lung. Eur J Cardiothorac Surg, 2003. **23**(3): p. 409-414.