

# ScholarWorks@GSU

## Effect of Risk and Prognosis Factors on Breast Cancer Survival: Study of a Large Dataset with a Long Term Follow-up

Authors	Wang, Hongwei
Citation	Wang, Hongwei. "Effect of Risk and Prognosis Factors on Breast Cancer Survival: Study of a Large Dataset with a Long Term Follow-up." 2012. Thesis, Georgia State University. <a href="https://doi.org/10.57709/2928391">https://doi.org/10.57709/2928391</a>
DOI	<a href="https://doi.org/10.57709/2928391">https://doi.org/10.57709/2928391</a>
Download date	2026-06-06 23:25:10
Link to Item	<a href="https://hdl.handle.net/20.500.14694/10415">https://hdl.handle.net/20.500.14694/10415</a>

# Effect of Risk and Prognosis Factors on Breast Cancer Survival: Study of a Large Dataset with a Long Term Follow-up

by

Hongwei Wang

Under the Direction of Dr. Jun Han

## ABSTRACT

The main goal of this study is to seek the effects of some risk and prognostic factors contributing to survival of female invasive breast cancer in United States. The study presents the survival analysis for the adult female invasive breast cancer based on the datasets chosen from the Surveillance Epidemiology and End Results (SEER) program of National Cancer Institute (NCI). In this study, the Cox proportional hazard regression model and logistic regression model were employed for statistical analysis. The odds ratios (OR), hazard ratios (HR) and confidence interval (C.I.) were obtained for the risk and prognosis factors. The study results showed that some risk and prognosis factors, such as the demographic factors (race and age), social and family factor (marital status), biomedical factors (tumor size, disease stage, tumor markers and tumor cell differentiation level etc.) and type of treatment patients received had significant effects on survival of the female invasive breast cancer patients.

INDEX WORDS: SEER, Breast cancer, Risk factors, Survival, Cox regression model, Logistic regression analysis

Effect of Risk and Prognosis Factors on Breast Cancer Survival: Study of a Large Dataset with a Long Term  
Follow-up

by

Hongwei Wang

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2012

Copyright by  
Hongwei Wang  
2012

Effect of Risk and Prognosis Factors on Breast Cancer Survival: Study of a Large Dataset with a Long Term  
Follow-up

by

Hongwei Wang

Committee Chair: Dr. Jun Han

Committee: Dr. Yichuan Zhao  
Dr. Xu Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2012

## ACKNOWLEDGEMENTS

It's my great pleasure to express my gratitude to all the people who helped me in my study and research at Georgia State University.

First of all, I want to thank my thesis supervisor, Dr. Jun Han. Without his valuable advice, guidance and supervision, I could not complete my thesis. During the writing of my thesis, Dr. Han helped me in each step with his time and wealthy knowledge. From him, I learned a lot of knowledge about analyzing complex data. Secondly, it's a great honor to have Drs. Yichuan Zhao and Xu Zhang to my thesis defense committee members and attend my thesis defense. Thanks for their time and the valuable comments on my thesis. I would also like to extend my thanks to Drs. Gengsheng Qin, Yuanhui Xiao, Ruiyan Luo, Jiawei Liu and all other faculty and staff members in the department of Mathematics and Statistics, for their wonderful lectures and generous help during my master degree study. My special thanks are reserved for my wife, parents and siblings, who gave me generous encouragement and support. Finally, I would like to convey my thanks to all of my colleagues and friends who gave me supports and helped me during the study period at Georgia State University.

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b> .....	iv
<b>LIST OF TABLES</b> .....	vii
<b>LIST OF FIGURES</b> .....	viii
<b>1 INTRODUCTION</b> .....	<b>1</b>
<b>1.1 Purpose of study</b> .....	1
<b>1.2 Expected results</b> .....	2
<b>2 STUDY SCENARIO</b> .....	<b>3</b>
<b>2.1 Data resource</b> .....	3
<b>2.2 Study design</b> .....	3
<b>2.3 Risk and prognostic factors for survival of female breast cancer patients</b> .....	4
<b>2.4 Dependent variable</b> .....	5
<b>2.5 Statistical analysis method</b> .....	5
<b>3 RESULTS</b> .....	<b>7</b>
<b>3.1 Descriptive statistical analysis</b> .....	7
<b>3.1.1 Demographic factors: age and race/ethnicity</b> .....	7
<b>3.1.2 Social/Family factor: marital status</b> .....	11
<b>3.1.3 Biomedical and pathological factors</b> .....	13
<b>3.1.4 Clinical treatment factors: surgery and/or radiotherapy</b> .....	17
<b>3.2 Survival analysis</b> .....	19
<b>3.2.1 The effect of demographic factors on survival and hazard function</b> .....	19
<b>3.2.2 The effect of social factor on survival and hazard function</b> .....	24
<b>3.2.3 The effect of biomedical prognosis factors on survival hazard function</b> .....	25

<b>3.2.4 The effect of treatments on survival and hazard function</b> .....	32
<b>3.3 Cox proportional hazard model and hazard ratio</b> .....	35
<b>3.4 Logistic regression analysis for 5-year survival status and odds ratio</b> .....	40
<b>4 RESEACH DEFICIENCY AND FURTHER PROSPECT</b> .....	46
<b>5 CONCLUSIONS</b> .....	46
<b>REFERENCES</b> .....	48
<b>APPENDICES</b> .....	50

## LIST OF TABLES

Table 1. Frequency and percentage of participants in different race/ethnicity group .....	7
Table 2. Frequency and percentage of participants in different age groups .....	8
Table 3. Age at diagnosis and the survival/death status two-way table .....	8
Table 4. Race/ethnicity and survival/death status two-way table.....	10
Table 5. Marital status of participants .....	11
Table 6. Marital status by different race/ethnicity .....	11
Table 7. Tumor cell differentiation grade .....	13
Table 8. Tumor size .....	13
Table 9. Lymph node examination result .....	13
Table 10. Tumor extension .....	14
Table 11. Tumor cell differentiation grade and tumor extension two-way table .....	15
Table 12. Tumor marker ERA.....	16
Table 13. Tumor marker PRA.....	16
Table 14. Surgery treatment.....	17
Table 15. Radiation treatment .....	17
Table 16. Summary of the treatments .....	18
Table 17. The effect of age race/ethnicity and marital status on hazard ratio.....	36
Table 18. The effect of histological and pathological factors on hazard ratio .....	37
Table 19. The effect of treatments on hazard ratio.....	38
Table 20. Relative odds ratio of demographic factors and social factor.....	40
Table 21. Relative odds ratio of histopathologic prognosis factors .....	42
Table 22. Relative odds ratio of clinical treatments factors.....	43
Table 23. The raw odds ratio and adjusted odds ratio for some risk and prognosis factors .....	44

**LIST OF FIGURES**

Figure 3.1 The effect of age at diagnosis on overall survival and hazard function .....	21
Figure 3.2 The effect of age at diagnosis on breast cancer specific survival and hazard function .....	22
Figure 3.3 The effect of race/ethnicity factor on survival and hazard function .....	23
Figure 3.4 The effect of marital status on survival and hazard function .....	24
Figure 3.5 The effect of cell differentiation grade on survival and hazard function .....	25
Figure 3.6 The effect of tumor size on survival and hazard function .....	27
Figure 3.7 The effect of lymph node metastasis on survival and hazard function.....	28
Figure 3.8 The effect of tumor extension on survival and hazard function .....	29
Figure 3.9 The effect of tumor markers on survival and hazard function .....	30
Figure 3.10 The effect of surgery treatment on survival and hazard function .....	32
Figure 3.11 The effect of radiation treatment on survival and hazard function.....	33
Figure 3.12 The effect of treatments on survival function.....	34

## 1 INTRODUCTION

### 1.1 Purpose of study

Breast cancer is the second most common cancer among American women (CDC 2007). In recent years, more than 200,000 new cases of breast cancer were diagnosed in the United States in each year. Almost 40,000 women may die due to breast cancer in 2012. Only lung cancer accounts for more cancer deaths in women (American Cancer Society 2011, SEER 2011).

Age and Race are factors commonly considered to have effect on prognosis and survival of breast cancer. The incidence and death rate of breast cancer significantly increases with age (CDC 2007). Over 90% of new breast cancer cases and almost 100% deaths occur in American women over 40 year-old age. Almost 80% of invasive breast cancer cases occur in 50 year- old or above female breast cancer patients. Among adult American females, 20-24 year-old females have the lowest breast cancer incidence rate; women over 65 year-old have the highest incidence rate. The median breast cancer diagnosis age is 61 year-old (American Cancer Society 2011). Secondly, in non-Hispanic white females, the breast cancer incidence rate was reported higher than African American women and other racial and ethnic populations (American Cancer Society 2011). However, some previous publications reported that African American breast cancer patients have a lower survival rate compared to White female patients. According to SEER Cancer Statistics Review, in the United States, among all race/ethnicity groups, the females' lifetime risk of dying from breast cancer from year 2006 to year 2008 for Whites was 2.76 (95% C.I. 2.74-2.77), for African American was 3.25 (95% C.I. 3.20-3.30), for Asian/Pacific Islanders was 1.69 (95% C.I. 1.59-1.81), for Hispanics was 2.03 (95% C.I. 1.96-2.11) and for American Indian/Alaska Natives was 1.82 (95% C.I. 1.58-2.13). Moreover, the female age-adjusted death rate for breast cancer from year 2004 to year 2008 for Whites was 22.8/100,000, for African Americans was 32.0/100,000, for

Asian/Pacific Islanders was 12.2/100,000, for Hispanics was 17.2/100,000, and for American Indian/Alaska Natives was 14.3/100,000 (Bassett 1986, American Cancer Society 2011).

In addition, tumor size, tumor extension and tumor histopathologic characteristics, such as cancer cell differentiation degree, tumor markers, are highly associated with survival of cancer patients and are considered as important prognosis factors (Elston and Ellis 1991). The extension of the disease, known as the extent or spread of the malignant tumor when it is diagnosed will significantly affect the seriousness of invasive breast cancer (SEER 2011). Some other factors, such as clinical treatments the patients received, also contribute to the prognosis and survival of breast cancer (Early breast cancer trialists' collaborative group 2005).

The purpose of this study is to explore the effects of risk and prognosis factors on survival or death of the adult female invasive breast cancer patients in the United States. To achieve this purpose, the Cox proportional hazard regression models and logistic regression models were constructed for the Surveillance Epidemiology and End Results data (SEER) to calculate the hazard ratios (HR), odds ratios (OR) and confidence interval (CI) for risk and prognosis factors.

## **1.2 Expected results**

The expected results: many risk factors and prognosis factors, such as demographic factors (age at diagnosis and race/ethnicity), social factors (marital status), and biomedical factors (tumor histopathologic characteristics) have significant effects on survival of female breast cancer patients. The treatments patients received are also expected to prolong the survival times. This will provide important information about early detection, screening and treatment of the invasive breast cancer and help to improve the survival rate of American women with breast cancer.

## **2 STUDY SCENARIO**

### **2.1 Data source**

The data used in this study were obtained (in TXT format) from the Surveillance Epidemiology and End Results (SEER) program of National Cancer Institute (NCI), a premier source for cancer statistics. The original data sets downloaded from the SEER's website contain cases from 12 geographic registry areas in the United States diagnosed from year 1973 to 2008. The adult female (20 years old and above) primary invasive breast cancer patients diagnosed from year 1990 to 2002 with active follow-up were selected for analyzing the effects of demographic factors, social/family factor and biomedical prognosis factors on breast cancer survival.

### **2.2 Study design**

In this study, the dataset including the adult female patients with primary invasive breast cancer diagnosed from 1990 to 2002 was used to analyze the effects of demographic factors (age and race/ethnicity), social factor (marital status), biomedical and histopathologic factors (tumor cell differentiation grade, tumor size, tumor extension, lymph node metastasis and treatments patients received etc.) on survival or death among female primary invasive breast cancer patients. The dataset in this study contains total 370133 female breast cancer patients in 12 geographic areas of the United States diagnosed from year 1990 to year 2002, and at the end of the year 2008, 221094 (59.73%) subjects were still alive at the cut-off point, 67740 (18.30%) subjects were dead due to breast cancer and other 81299 (21.96%) participants were dead due to other diseases during the time period. Logistic regression models and Cox proportional hazard regression models were constructed to analyze the effect of all risk and prognosis factors and calculate the odds ratio and hazard ratio. In this study, the SAS 9.2 statistical software procedures (logistic procedure, lifetest procedure and phreg procedure etc.) were used to do data analysis.

### **2.3 Risk and prognostic factors for survival of female breast cancer patients**

The survival and prognosis of breast cancer is affected by many factors, and those factors that are associated with prognosis of disease are considered as risk and prognosis factors. Risk and prognosis factors in this study include demographic factors, social factor and prognosis factors include biomedical or histopathologic factors.

Demographic factors of participants include age and race/ethnicity. Age was considered as the most important risk factor for female breast cancer patients (American Cancer Society 2011). The classification of those ages at diagnosis of the patients in this study was expanded to 4 groups: 20-39 years old, 40 to 49 years old, 50 to 64 years old and 65 years old and above. All the patients in this study were classified into 6 race/ethnicity groups: non-Hispanic Whites, African American, Hispanic, Asian and Pacific Islander, American Indian/Alaskan Native and other races.

Patients' marital status was considered as a social relations factor. In this study, patients' marital status is determined using responses to the study program, "now married", "widowed, divorced or separated", or "never married".

Biomedical factors include tumor size, tumor markers, tumor cell differentiation grade, lymph node examination result and tumor extension. Tumor size at the time of diagnosis was classified into 5 classes: tumor less than 2cm, 2-5cm, over 5cm, unclear cases and Paget's disease of nipple with no demonstrable tumor; tumor cell differentiation grade includes five groups: well differentiated, moderately differentiated, poorly differentiated, undifferentiated and unknown cases; local lymphatic metastasis includes three categories: negative examination result, positive result and unclear cases; tumor markers (include the estrogen receptor assay, ERA and progesterone receptor assay, PRA) were classified into positive marker group, negative marker group, borderline group and unclear group, respectively; the tumor extension was classified into four groups: localized, regional, distant and extension unknown. In addition, the treatment methods: surgery and radiotherapy were also considered as prognosis factors

for survival and prognosis of breast cancer patients. Treatments were categorized into the following groups: no treatment, treatment and treatment unknown.

#### **2.4 Dependent variables**

In survival analysis of the SEER data, the survival time (months of survival from baseline to death or to December 31, 2008) and survival status (alive or censoring and dead) were used as the dependent variables. In logistical model, the categorical variable, survival status (dead and alive or censoring) was set to be the outcome variable.

#### **2.5 Statistical analysis method**

At the beginning of the study, the differences in frequency and percentage of demographic factors, social factors, biomedical and histological factors were examined by using the SAS freq procedure. A chi-square analysis was also used for the difference of survival status among age groups and race/ethnicity groups. The significance of difference of the incidence rates among all age groups was tested as well.

In survival analysis study, the SAS lifetest procedure can be used to calculate the nonparametric estimates of the survival functions, and it also can be used to do comparison among survival curves and the rank tests computation for association of the event time with covariates (SAS Institute Inc. 2012). In this study, the lifetest procedure was used to check the effect of demographic factors, social factors, biomedical and histological factors on survival and hazard functions and test the Cox proportionality assumption via visual examinations of survival curves and “log-log” plots curves. In addition, the Cox’s proportional hazards model was constructed for this complex study dataset. If  $t$  is the observed time of event occurrence,  $T$  represent the random variable of survival time, which is defined on the domain  $t \in [0, \infty)$ . Then  $S(t)$  is the survival function,  $S(t) = p(T > t)$ , i.e. the probability of surviving at least to time  $t$ . As a probability, it has range  $S(t) \in [0, 1]$ .  $S(t) = 1 - F(t) = \int_t^{\infty} f(\tau) d\tau$ . The hazard function represents the concept of the risk of dying at the time  $t$ , given the subject having survived to time  $t$ , which is defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{p(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = f(t) / S(t).$$

In the Cox's proportional hazards model, we can start with a baseline hazard function and set it as  $h_0(t)$ . We can model the effect of some covariates on the hazard function which is given by

$$h(t|Z) = h_0(t)\exp(\beta^T Z)$$

when  $h_0(t)$  is the unspecified baseline hazard function. If  $h(t|Z)$  denotes the hazard rate at time  $t$  for an individual with risk vector  $Z$ . The Cox model is

$$h(t|Z) = h_0(t)c(\beta^T Z) \text{ (Cox, 1972),}$$

with the parameter vector  $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_i)^T$  (Cox, 1972). In Cox proportional hazard model, if we compare two subjects with covariate values  $Z$  and  $Z^*$ , the ratio of the two hazard rates is given by

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t)\exp[\sum_{k=1}^p \beta_k Z_k]}{h_0(t)\exp[\sum_{k=1}^p \beta_k Z_k^*]} = \exp[\sum_{k=1}^p \beta_k (Z_k - Z_k^*)],$$

which is the relative risk of one subject with the risk factor  $Z$  having an event compared to a subject with risk factor  $Z^*$  (Klein and Moeschberger 2003). The SAS phreg procedure was used to construct the model and the hazard ratio and its confidence interval (CI) for each factor were also obtained through this procedure (SAS Institute Inc. 2012).

In addition, a multivariate logistic regression model was used to check the effect of the risk and prognosis factors on breast cancer 5-year relative survival and calculate the odds ratio (OR) and its confidence interval (CI) for each factor. The logistic regression model can be used to investigate the relationship between a discrete response variable and a series of explanatory variables. The SAS logistic procedure can be used to fit a linear logistic regression model for a discrete survey response variable by maximum likelihood method (SAS Institute Inc. 2012).

In this study, the SAS software (version 9.2, SAS Institute, Cary, NC) was used in data analysis and p-value less than significance level 0.05 was considered to be statistical significant.

### 3 RESULTS

#### 3.1 Descriptive statistical analysis

In this section, the differences in frequency and percentage of demographic factors, social factors, biomedical and histological factors were examined by using the SAS freq procedure (SAS 9.2, SAS Institute Inc.), and also a chi-square analysis was used for the difference of survival status among age groups, race/ethnicity groups and treatments groups. The tables in this thesis were created using Word, Excel and SAS ODS (Output Delivering System).

##### 3.1.1 Demographic characteristics: age and race/ethnicity

Table 1. Frequency and percentage of participants in different race/ethnicity group

Race	Frequency	Percentage
Non-Hispanic White	287818	77.76%
African American	31144	8.41%
Hispanic	26589	7.18%
Asian	22832	6.17%
American Native	1601	0.43%
Other	149	0.04%
Total	370133	100.00%

Total 370133 adult female patients with primary invasive breast cancer diagnosed from 1990 to 2002 in SEER data resource were selected in this study. 287818 (77.76%) patients were non-Hispanic Whites; 31144 (8.41%) were African-American females; 26589 (7.18%) were Hispanic females; 22832 (6.17%) were Asians and 1601 cases (0.43%) were American Indians/Alaska Natives. Only 173 (0.04%) cases were other races/ethnics.

In this study, the subjects were separated into 4 different age groups (Table 2): 20-39 years old (21608, 5.84%) and 40 to 49 years old (64122, 17.32%), 50 to 64 years old (119346, 32.24%) and 65-year old and above (165057, 44.59%). The mean and median age at diagnosis were 61.8 year-old and 62 year-old, respectively (SAS means Procedure, SAS institute inc. 2012).

Table 2. Frequency and percentage of participants in different age groups

Age	Frequency	percentage
20-39	21608	5.84%
40-49	64122	17.32%
50-64	119346	32.24%
65 and above	165057	44.59%
Total	370133	100%

Table 3. Age at diagnosis and the survival/death status two-way table

Frequency Percentage Row Pct Column Pct	Death due to breast cancer	Death due to other reasons	Alive	Total
20-39	5955 1.61 27.56 8.79	770 0.21 3.56 0.95	14883 4.02 68.88 6.73	21608 5.84
40 to 49	12333 3.33 19.23 18.21	2911 0.79 4.54 3.58	48878 13.21 76.23 22.11	64122 17.32
50 to 64	20673 5.59 17.32 30.52	11955 3.23 10.02 14.70	86718 23.43 72.66 39.22	119346 32.24
65 and above	28779 7.78 17.44 42.48	65663 17.74 39.78 80.77	70615 19.08 42.78 31.94	165057 44.59
Total	67740 18.30	81299 21.96	221094 59.73	370133 100.00

From table 2 we found that the proportion of 20-39 year-old patients in this study was only 5.84%, and among all the other older age groups, the percentage increased with age. The difference of percentages among all age groups was significant ( $P < 0.0001$ ). This indicates that breast cancer incidence rate increases with age (American Cancer Society 2011).

When age at diagnosis increased, the overall death rate due to all the reasons also increased (Table 3). The "50 to 64 year-old" and the "65 year-old and above" age groups had the higher death rate among all age groups ( $P < 0.0001$ ). However, in this study, the 20-39 years old age group had a relatively higher breast cancer specific death rate: 27.56% of patients in this age group died of breast cancer. In the 40-49 year-old age group, the breast cancer specific death rate was 19.23%, and in the 50-64 year-old and 65 year-old and above age groups, the breast cancer specific mortalities were only 17.32% and 17.44%, respectively.

Some previous studies reported that African American breast cancer patients had a lower survival rate compared to White female patients due to the lower social-economical status (Bassett, 1986; American Cancer Society 2011). From table 4 below, we also can found that African American patients had highest breast cancer specific death rate (29.32%) and overall lowest long-term survival rate (50.28%) among all races/ethnics ( $P < 0.0001$ ).

Table 4. Race/ethnicity and survival/death status two-way table

Frequency Percentage Row Pct Column Pct	Death due to breast cancer	Death due to other reasons	Alive	Total
Non-Hispanic Whites	49031	67482	171305	287818
	13.25	18.23	46.28	77.76
	17.04	23.45	59.52	
	72.38	83.00	77.48	
African American	9130	6355	15659	31144
	2.47	1.72	4.23	8.41
	29.32	20.41	50.28	
	13.48	7.82	7.08	
Hispanic	5630	4132	16827	26589
	1.52	1.12	4.55	7.18
	21.17	15.54	63.29	
	8.31	5.08	7.61	
Asian	3559	3026	16247	22832
	0.96	0.82	4.39	6.17
	15.59	13.25	71.16	
	5.25	3.72	7.35	
Native	383	292	926	1601
	0.10	0.79	0.25	0.43
	23.92	18.24	57.84	
	0.57	0.36	0.42	
Other	7	12	130	149
	0.00	0.00	0.04	0.04
	4.70	8.05	87.25	
	0.01	0.01	0.06	
Total	67740	81299	221094	370133
	18.30	21.96	59.73	100.00

Note: Pct-percentage

**3.1.2 Social/Family factor: marital status**

Table 5. Marital status of participants

Marital Status	Frequency	Percentage
Married	200286	54.11%
Ever Married	114687	30.99%
Single	42129	11.38%
Unknown	13031	3.52%
Total	370133	100.00%

Table 6. Marital status by different race/ethnicity

Frequency Percentage Row Pct Column Pct	Married	Ever married	Single	Unknown	Total
Non-Hispanic Whites	158945 42.94 55.22 79.36	90979 24.58 31.61 79.33	27934 7.55 9.71 66.31	9960 2.69 3.46 76.43	287818 77.76
African American	11117 3.00 35.70 5.55	11433 3.09 36.71 9.97	7176 1.94 23.04 17.03	1418 0.38 4.55 10.88	31144 8.41
Hispanic	14595 3.94 54.89 7.29	7037 1.90 26.47 6.14	4044 1.09 15.21 9.60	913 0.25 3.43 7.01	26589 7.18
Asian	14799 4.00 64.82 7.39	4828 1.30 21.15 4.21	2735 0.74 11.98 6.49	470 0.13 2.06 3.61	22832 6.17
Native	743 0.20 46.41	385 0.10 24.05	214 0.06 13.37	259 0.07 16.18	1601 0.43

	0.37	0.34	0.51	1.99	
Other	87 0.02 58.39 0.04	25 0.01 16.78 0.02	26 0.01 17.45 0.06	11 0.00 7.38 0.08	149 0.04
Total	200286 54.11	114687 30.99	42129 11.38	13031 3.52	370133 100.00

As shown in table 5, among total 370133 patients in this study, there were 200286 (54.11%) married participants; 42129 (11.38%) participants had never married; 114687 (30.99%) participants were widowed, divorced or separated and in this study, they were defined as ever married. The marital status of other 13031 (3.52%) patients was unclear.

From Table 5 and 6 we found that African American females relatively had a significantly higher rate of single (23.04%) or currently single (ever married) ship (36.71%), and lower marriage rate (35.70%) compared with other races/ethnicities ( $P < 0.0001$ ).

### 3.1.3 Biomedical and pathological factors

Table 7. Tumor cell differentiation grade

Tumor Cell Differentiation Grade	Frequency	Percentage
well differentiated	55510	15.00%
Moderately differentiated	125402	33.88%
Poorly differentiated	110689	29.91%
Undifferentiated	8459	2.29%
Unknown	70073	18.93%
Total	370133	100.00%

From table 7, we found that among all the invasive breast cancer cases in this study, 55510 (15.00%) subjects were tumor cell well differentiated cases; 125402 (33.88%) were moderately differentiated cases; 110689 (29.91%) were poorly differentiated, and 8459 (2.29%) cases were the rare undifferentiated grade; other 70073 (18.93%) were unknown cases.

Table 8. Tumor size

Tumor Size	Frequency	Percentage
2cm or less in diameter	217061	58.64%
Paget's disease	138	0.04%
2-5cm in diameter	98375	26.58%
Over 5cm in diameter	24665	6.66%
Unknown	29894	8.08%
Total	370133	100.00%

Table 9. Lymph node examination result

Lymph Node Examination	Frequency	Percentage
Negative	193011	52.15%
Positive	111855	30.22%
Unknown	65267	17.63%
Total	370133	100.00%

Table 10. Tumor extension

Tumor Extension	Frequency	Percentage
Localized	319500	86.32%
Regional	19175	5.18%
Distant	20464	5.53%
Unknown	10994	2.97%
Total	370133	100.00%

Table 8 and 9 showed that in more than half cases (58.64%) the patients' tumor sizes at the diagnosis were less than 2cm, and 85.22% of tumors were less than 5cm. There were 193011 (52.15%) patients had no regional lymph node metastasis at the time of diagnosis while 111855 (30.22%) patients had nodal involvement. Other 65267 (17.63%) cases did not have clear lymph node metastasis information. A small size tumor without lymph node metastasis indicates an early disease stage and usually has a better prognosis and survival chance (American Cancer Society 2011).

Invasive breast cancer extension was divided into three different histological stages in this study. The Localized extension indicates the tumor is confined entirely within the original organ; Regional extension indicates the tumor has extended into the surrounding organs or tissues and/or into the regional lymph nodes; the distant extension stage indicates the tumor has extended into the remote organ(s) with or without the lymph nodes metastasis (SEER 2011). Table 10 indicated that the majority of participants in this study (319500, 86.32%) was in the localized extension stage; 19175 (5.18%) cases were in the regional stage; 20464 (5.53%) patients were in the relative later stage, the distant stage; and other cases were unclear (10994, 2.97%). The long-term survival of breast cancer patients is highly affected by the disease extension stage (Farley and Flannery 1989, SEER 2011). Tumor size less than 2cm in diameter, lymph node negative status, tumor markers positive status, localized tumor and local extended tumor are considered as low-risk tumor characteristics with relative better prognosis and survival chance (Anderson and Jatoi et al. 2005). Table 11 indicated that a well differentiated tumor is usually

associates with an earlier extension stage ( $P < 0.0001$ ). For example, 95.89% well differentiated tumor were found without regional or distant extension. However, a decline of 12% was observed in the poorly differentiated tumor group.

Table 11. Tumor cell differentiation grade and Tumor extension two-way table

Tumor cell differentiation grade and Tumor extension two-way table					
Frequency Percentage Row Pct Column Pct	Localized	Regional	Distant	Unclear	Total
Well differentiated	53226	1227	596	461	55510
	14.38	0.33	0.16	0.12	15.00
	95.89	2.21	1.07	0.83	
	16.66	6.40	2.91	4.19	
Moderately differentiated	114107	5898	4011	1386	125402
	30.83	1.59	1.08	0.37	33.88
	90.99	4.70	3.20	1.11	
	35.71	30.76	19.60	12.61	
Poorly differentiated	92603	7881	8265	1940	110689
	25.02	2.13	2.23	0.52	29.91
	83.66	7.12	7.47	1.75	
	28.98	41.10	40.39	17.65	
Undifferentiated	7078	549	662	170	8459
	1.91	0.15	0.18	0.05	2.29
	83.67	6.49	7.83	2.01	
	2.22	2.86	3.23	1.55	
Unknown	52486	3620	6930	7037	70073
	14.18	0.98	1.87	1.90	18.93
	74.90	5.17	9.89	10.04	

	16.43	18.88	33.86	64.01	
--	-------	-------	-------	-------	--

Table 12. Tumor marker ERA

Tumor Marker ERA	Frequency	Percentage
positive	225625	60.96%
negative	66030	17.84%
Borderline	1835	0.50%
unknown	76643	20.71%
Total	370133	100.00%

Table 13. Tumor marker PRA

Tumor Marker PRA	Frequency	Percentage
positive	188293	50.87%
negative	94573	25.55%
Borderline	2651	0.72%
unknown	84616	22.86%
Total	370133	100.00%

Table 12 and table 13 showed the tumor markers examination result. ERA (estrogen receptor assay) and PRA (progesterone receptor assay) are two different tumor markers. More than 60% of patients had a positive ERA examination result and more than half of the patients (50.87%, table 13) had a positive PRA examination result. Positive expressions of tumor markers are associated with a relative better prognosis and also a higher survival chance (Anderson and Jatoi et al. 2005).

### 3.1.4 Clinical treatment factors: surgery and/or radiotherapy

Table 14. Surgery treatment

Surgery Treatment	Frequency	Percentage
Non-Surgery	21225	5.73%
Surgery	348313	94.10%
Unknown	595	0.16%
Total	370133	100.00%

Table 15. Radiation treatment

Radiation Treatment	Frequency	Percentage
Non-Radiation	204600	55.28%
Radiation	154698	41.80%
Unknown	10835	2.93%
Total	370133	100.00%

In breast cancer, local treatment can significantly affect the risk of local or regional recurrence and long-term survival rate (Early Breast Cancer Trialists' Collaborative Group 2005). In this study, the effects of surgery and/or radiotherapy treatments the patients received were also considered as important prognosis factors. From table 14, 15 and 16, we found that more than 90% of the patients received surgery treatment. However, only 154698 (41.8%) patients received the radiotherapy. 150332 (40.62%) participants received both surgery treatment and radiotherapy treatment.

Table 16. Summary of the treatments

Surgery and Radiotherapy Two-Way Table				
	no radiation	radiation	unknown	Total
no surgery	16345	4337	543	21225
	4.42	1.17	0.15	5.73
	77.01	20.43	2.56	
	7.99	2.80	5.01	
surgery	187973	150332	10008	348313
	50.79	40.62	2.70	94.10
	53.97	43.16	2.87	
	91.87	97.18	92.37	
unknown	282	29	284	595
	0.08	0.01	0.08	0.16
	47.39	4.87	47.73	
	0.14	0.02	2.62	
Total	204600	154698	10835	370133
	55.28	41.80	2.93	100.00

## 3.2 Survival analysis

### 3.2.1 The effect of demographic factors on survival and hazard function

In this part, the effects of all risk and prognosis factors on breast cancer survival and hazard functions were examined by using SAS lifetest procedure.

Figure 3.1 showed the effect of age at diagnosis on overall survival distribution function and hazard function. The age group “65 and above” had the worst survival function and the highest hazard probability ( $P < 0.0001$ ). In addition, compared to other age groups, the overall survival probability of the “65 and above” age group decreased faster over time while all other survival curves showed a slow decreasing pattern. The “40-49” year-old age group had the highest overall survival probability and the lowest hazard probability among all age groups. However, although the “20-39” year-old age group had the minimum number of patients in this study, the overall survival probability of this group was only higher than the “65 and above” age group, but lower than the “40-49” year-old group and the “50-64” year-old group. Figure 3.2 showed the effect of age at diagnosis on breast cancer specific survival distribution function and hazard function. The age group of “20-39” had the worst survival function and the highest hazard probability ( $P < 0.0001$ ) compared to other age groups. This is probably because the “high-risk” breast cancers (early age at onset, larger tumor size, axillary lymph node metastasis, worse histologic differentiation and negative tumor markers expression etc.) have more occurrences among the younger patients (Anderson and Jatoi 2005, Albain and Allred et al. 1994, Diab and Elledge et al. 2000). The breast cancer specific hazard curve of the “65 and above” group showed a slow decreasing pattern. For other age groups, the peaks of the breast cancer specific hazard curves all appeared around 20-25 months after diagnosis and then slowly declined. As indicated by Jatoi and Tsimelzon, the highest risk of breast cancer death occurs around 2-3 years after the diagnosis (Jatoi and Tsimelzon et al. 2005).

Figure 3.3 showed that the African American female patients had the worst breast cancer specific survival chance and the highest hazard rate comparing to non-Hispanic Whites, Asians and Hispanics. At the beginning, the African American group had a higher hazard rate and the peak was shown around 20 months and then slowly decreased with time. Non-Hispanic Whites and Asians had better survival probability and the survival curves had a relative flat pattern. However, the hazard rate curves of American Indian/Alaska Native and other races were unstable due to the relatively small sizes during a long-term follow-up time period.

Figure 3.1 The effect of age at diagnosis on overall survival and hazard function

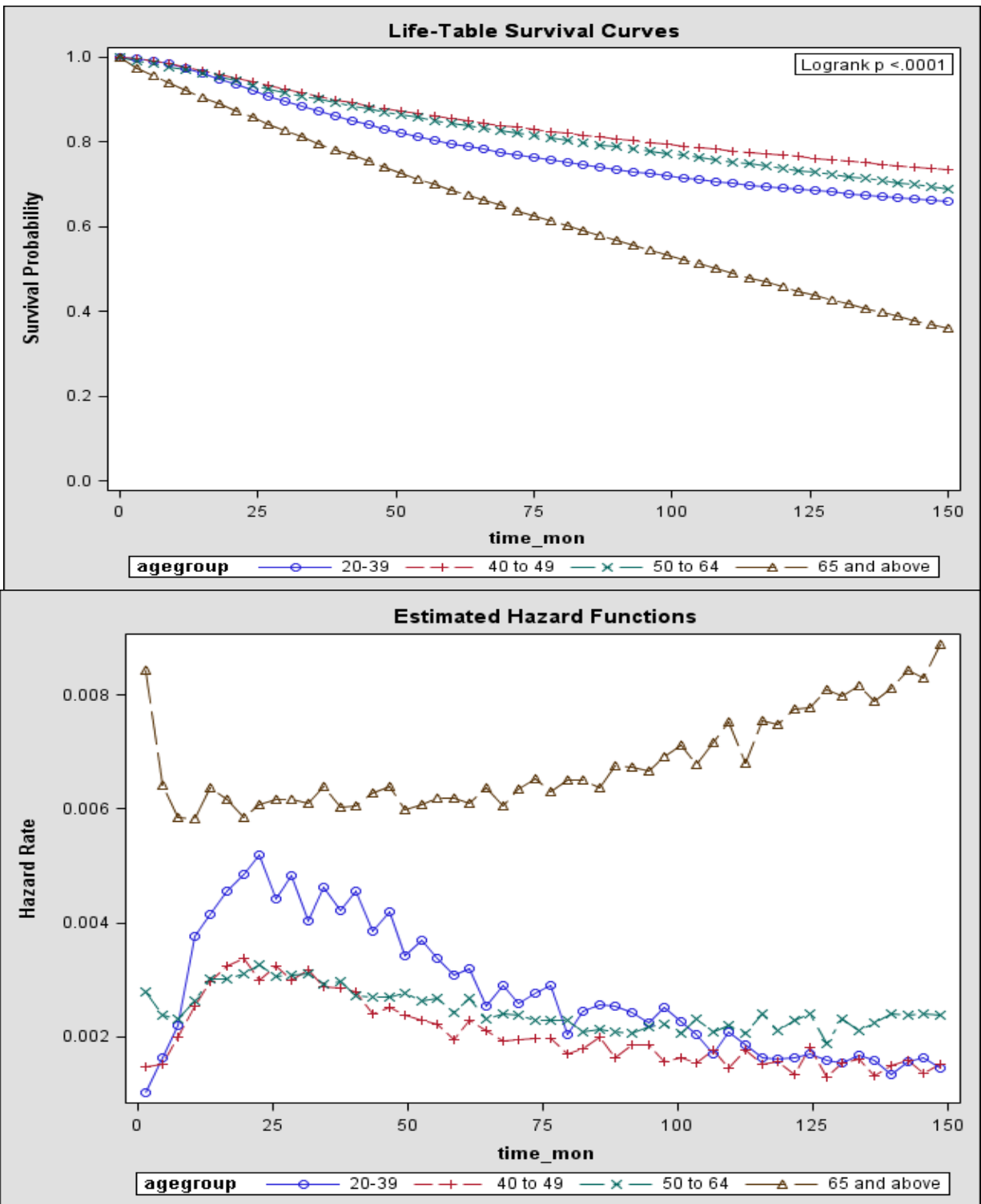


Figure 3.2 The effect of age at diagnosis on breast cancer specific survival and hazard function

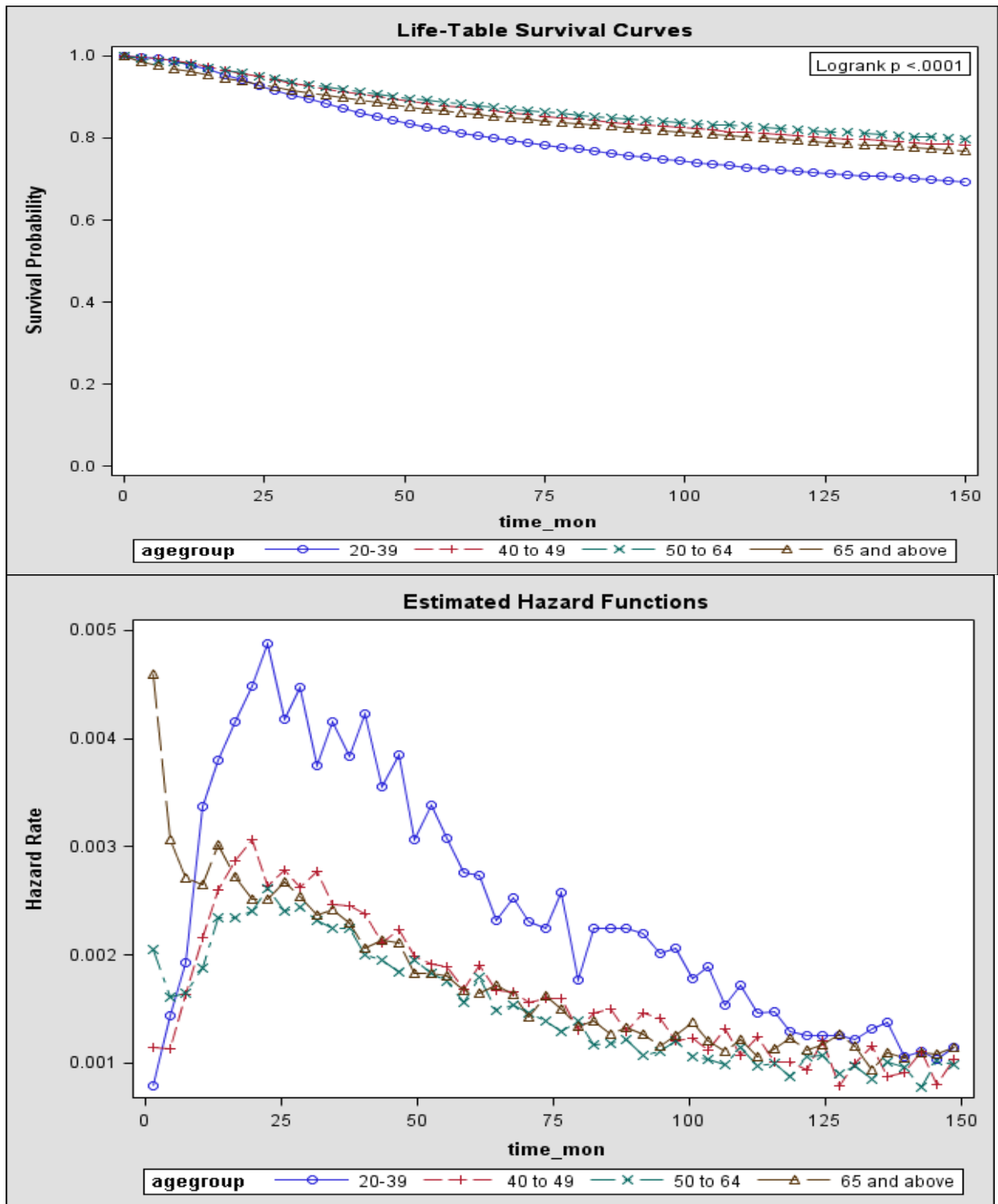
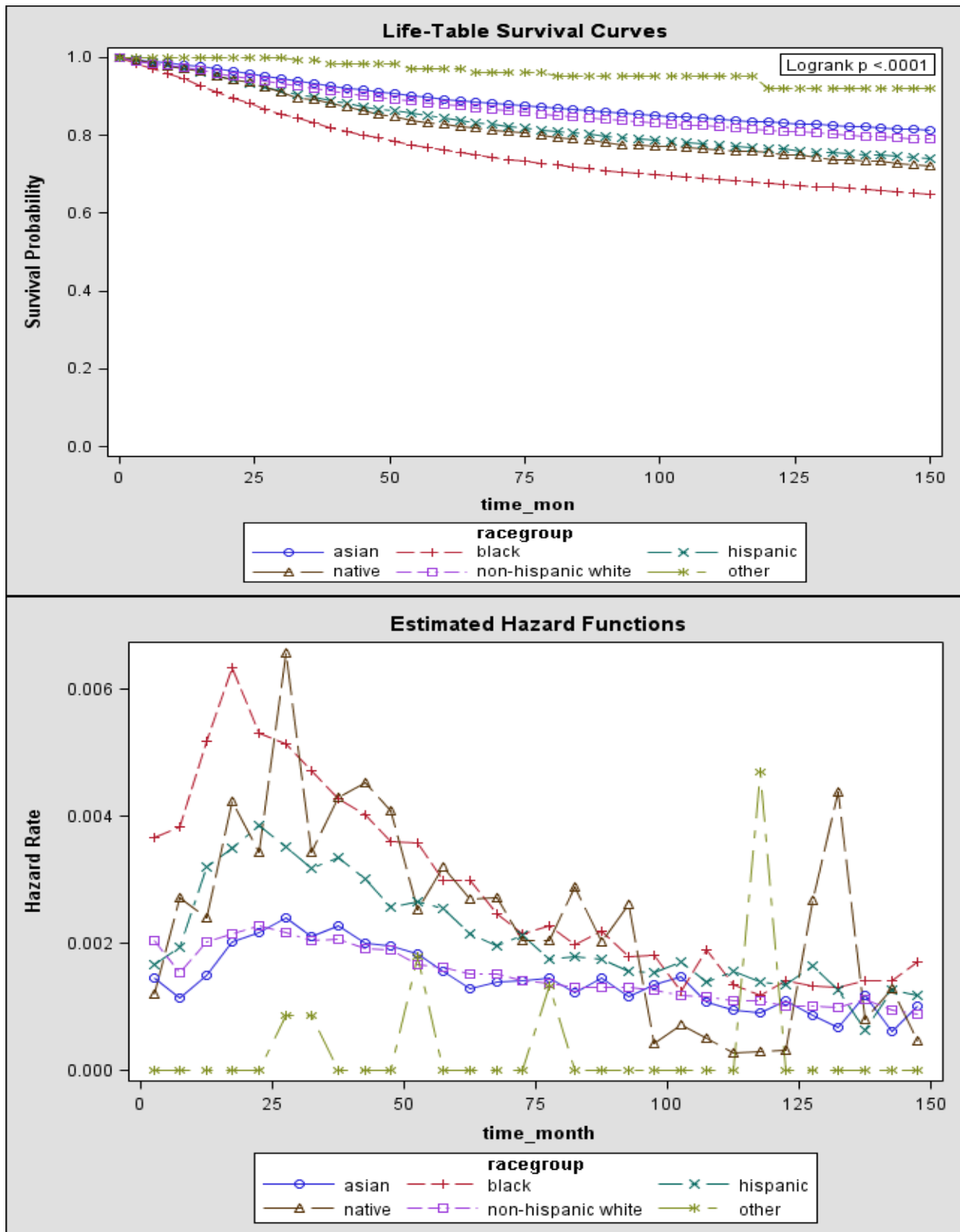


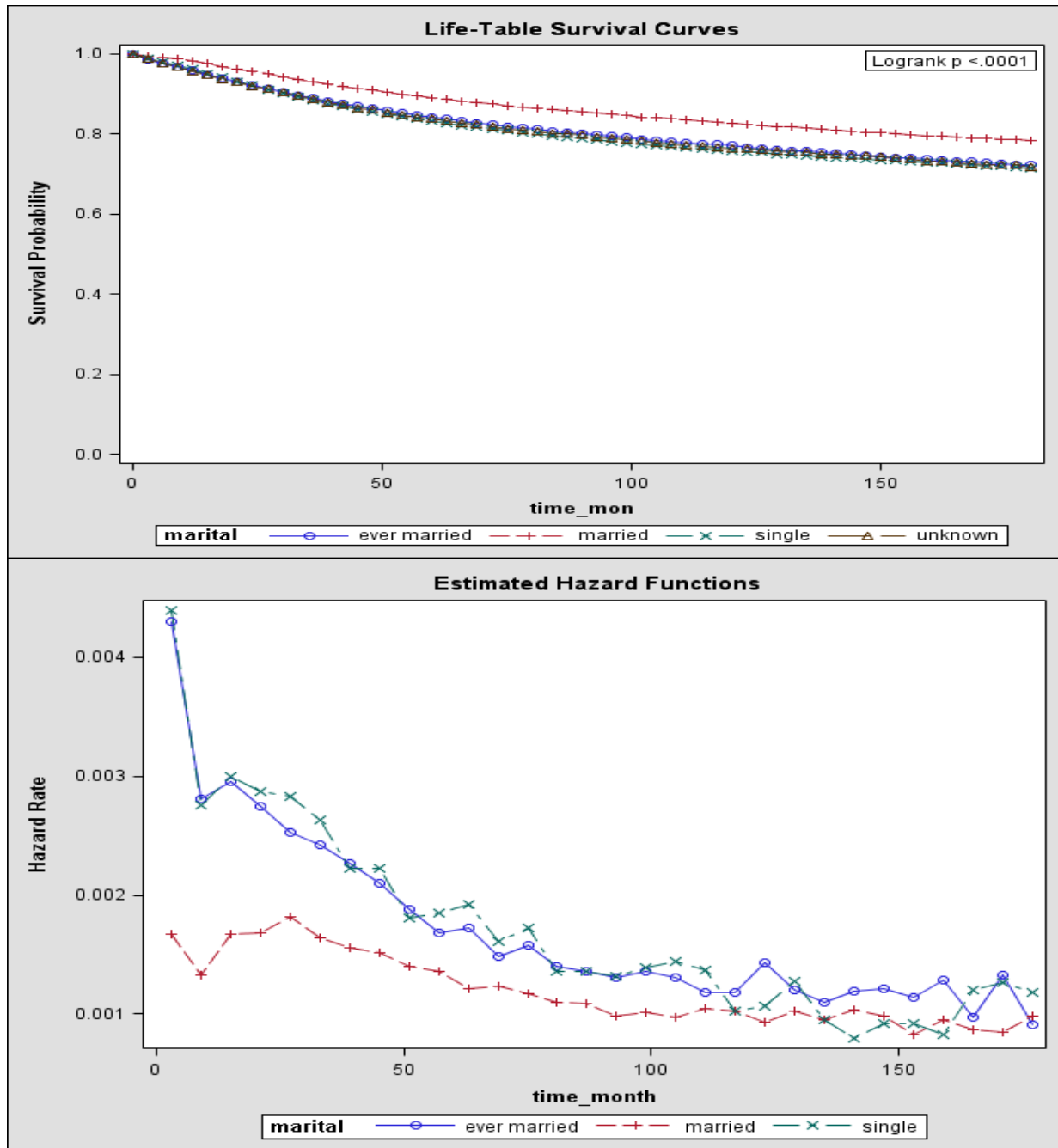
Figure 3.3 The effect of race/ethnicity factor on survival and hazard function



### 3.2.2 The effect of social factor on survival and hazard function

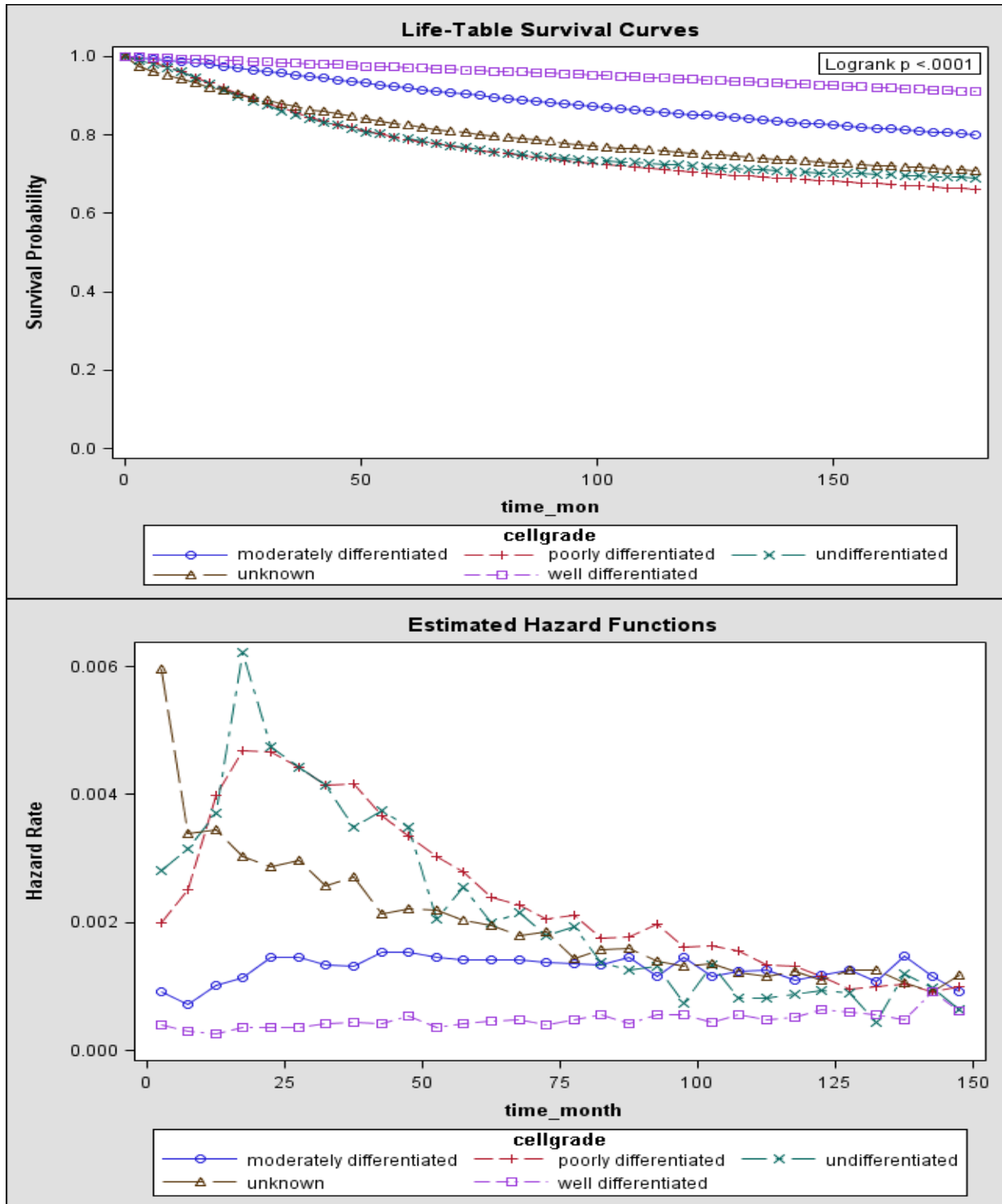
Figure 3.4 showed that married people had the relative better survival ability among married, ever married and single (never married) groups ( $P < 0.0001$ ). There was no significant difference of survival or hazard function between the single group and the ever married group.

Figure 3.4 The effect of marital status on survival and hazard function



### 3.2.3 The effect of biomedical prognosis factors on survival and hazard function

Figure 3.5 The effect of cell differentiation grade on survival and hazard function



Tumor cell differentiation grade had significant effect on survival and hazard functions ( $P < 0.0001$ ). Figure 3.5 showed that “well differentiated” group had the best survival probability; “poorly differentiated” and “undifferentiated” groups had lower survival probability. The peaks of hazard rate of both poorly differentiated and undifferentiated groups were shown around 20 months after diagnosis. However, the well differentiated group and the moderately differentiated groups had the flat pattern, and the well differentiated group had the lowest hazard rate.

Figure 3.6 showed that tumor size is an important prognosis factor and had significant effect on female breast cancer survival ( $P < 0.0001$ ). Tumor size less than 2cm and Paget’s disease cases had the best survival ability and the lowest hazard rate with relatively flat patterns as shown on figure 3.6. Tumor size over 5cm patients had the highest hazard rate during all the follow-up time period and the peak appeared around 20 months after diagnosis.

Figure 3.7 demonstrated the negative lymph node status (without regional lymph node metastasis) had relatively better survival chance and the lowest hazard rate compared to the positive lymph node group ( $P < 0.0001$ ). The hazard rate of positive lymph node groups reached the peak after 20 months and then decreased slowly.

The hazard function for distant extension cases had the highest hazard rate in most time of follow-up period. The localized extension group had the best survival ability and lowest hazard rate with a flat pattern curve ( $P < 0.0001$ , figure 3.8). Figure 3.9 showed that patients with positive tumor marker expression had a better survival chance.

Figures 3.6 to 3.9 indicated that all the histological and pathological factors had significant effect on both survival ability and hazard rate. The lifetest procedure results showed that smaller tumor size, well differentiated tumor, positive tumor markers (figure 3.9), negative lymph node metastasis and early histological (localized tumor) disease stage had better survival ability. The “high-risk” tumors, such as

tumors with larger size, positive lymph node, negative tumor markers expression, poorly differentiated or undifferentiated tumors, regional or distant extension had lower survival probability.

Figure 3.6 The effect of tumor size on survival and hazard function

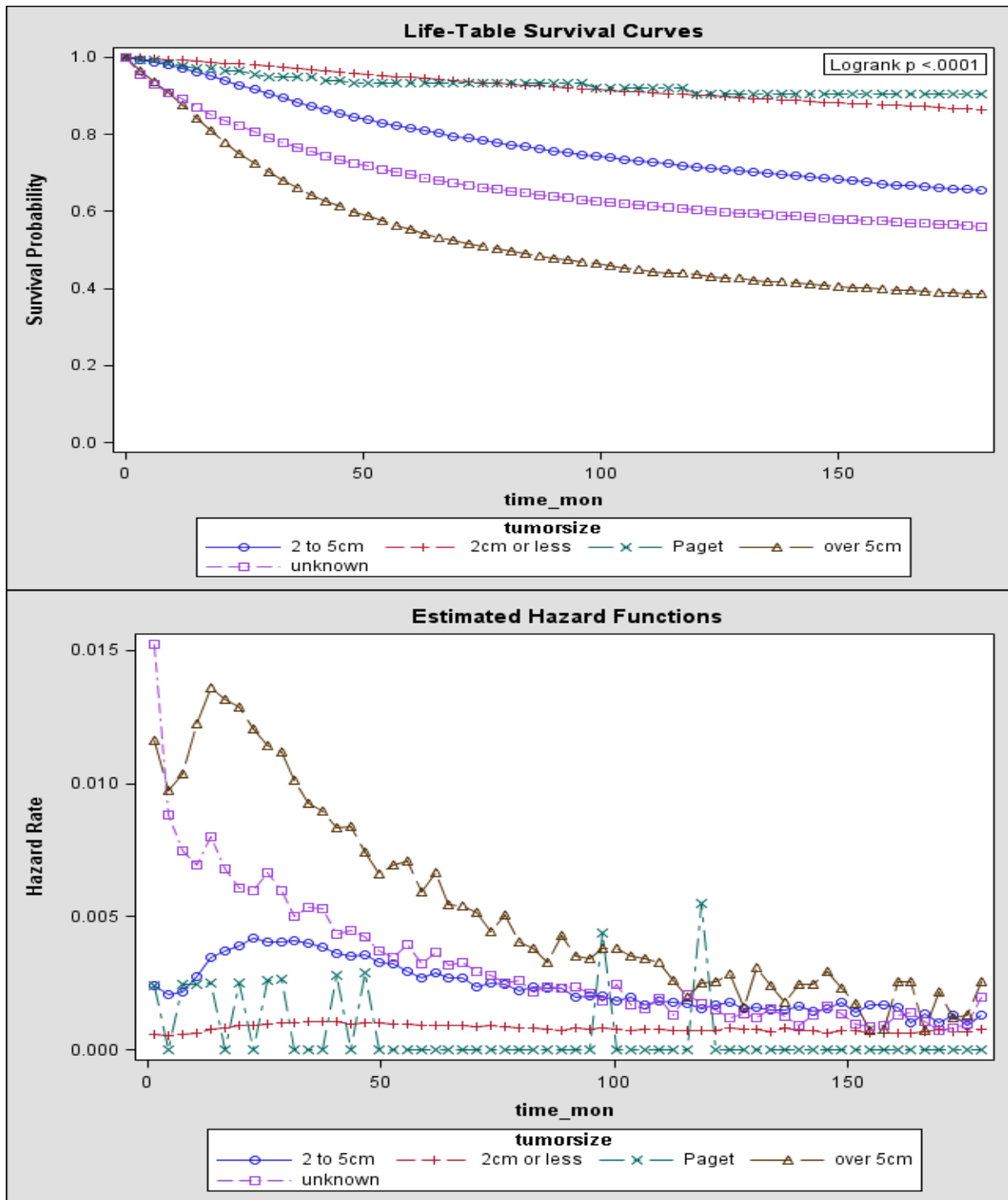


Figure 3.7 The effect of lymph node metastasis on survival and hazard function

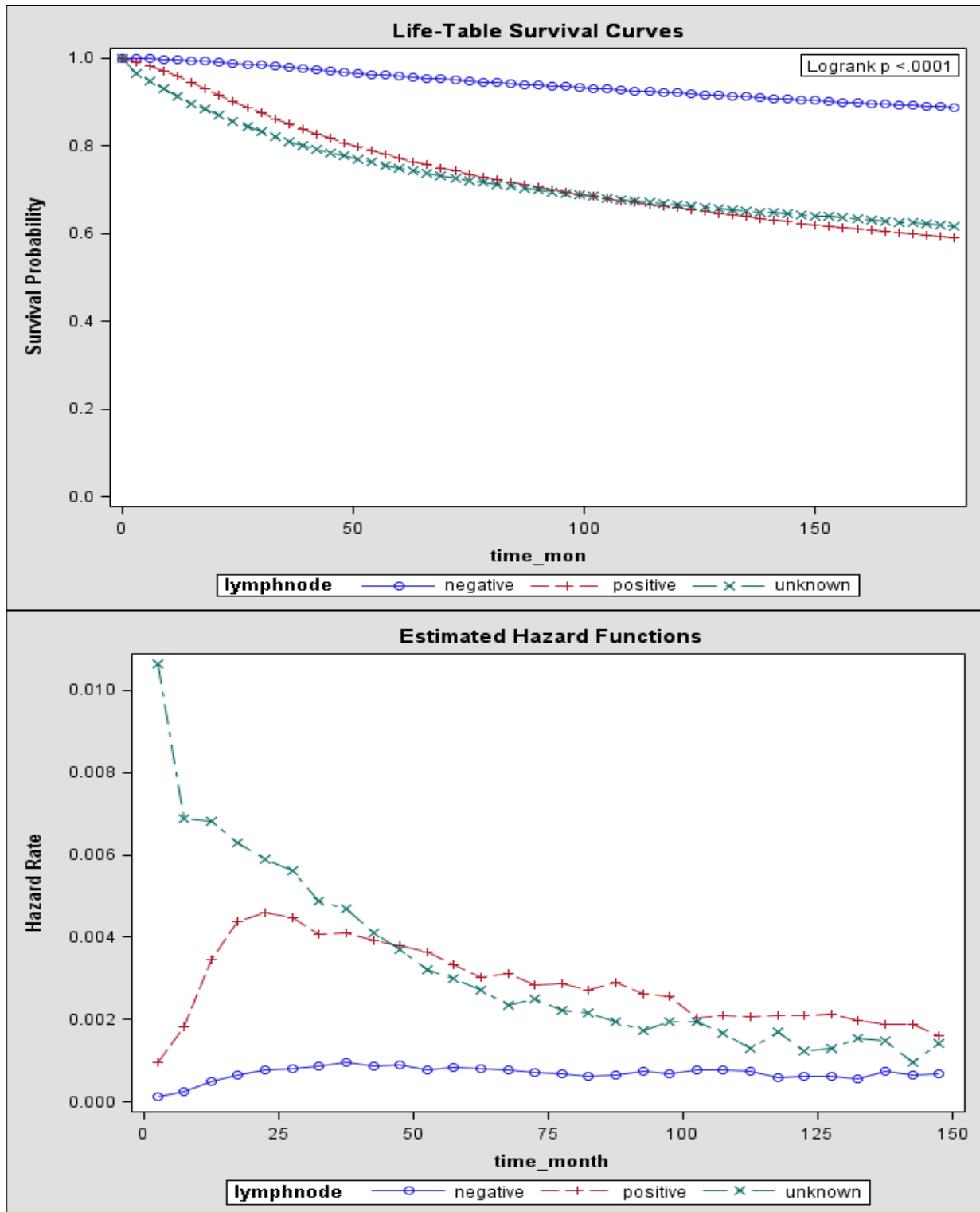


Figure 3.8 The effect of tumor extension on survival and hazard function

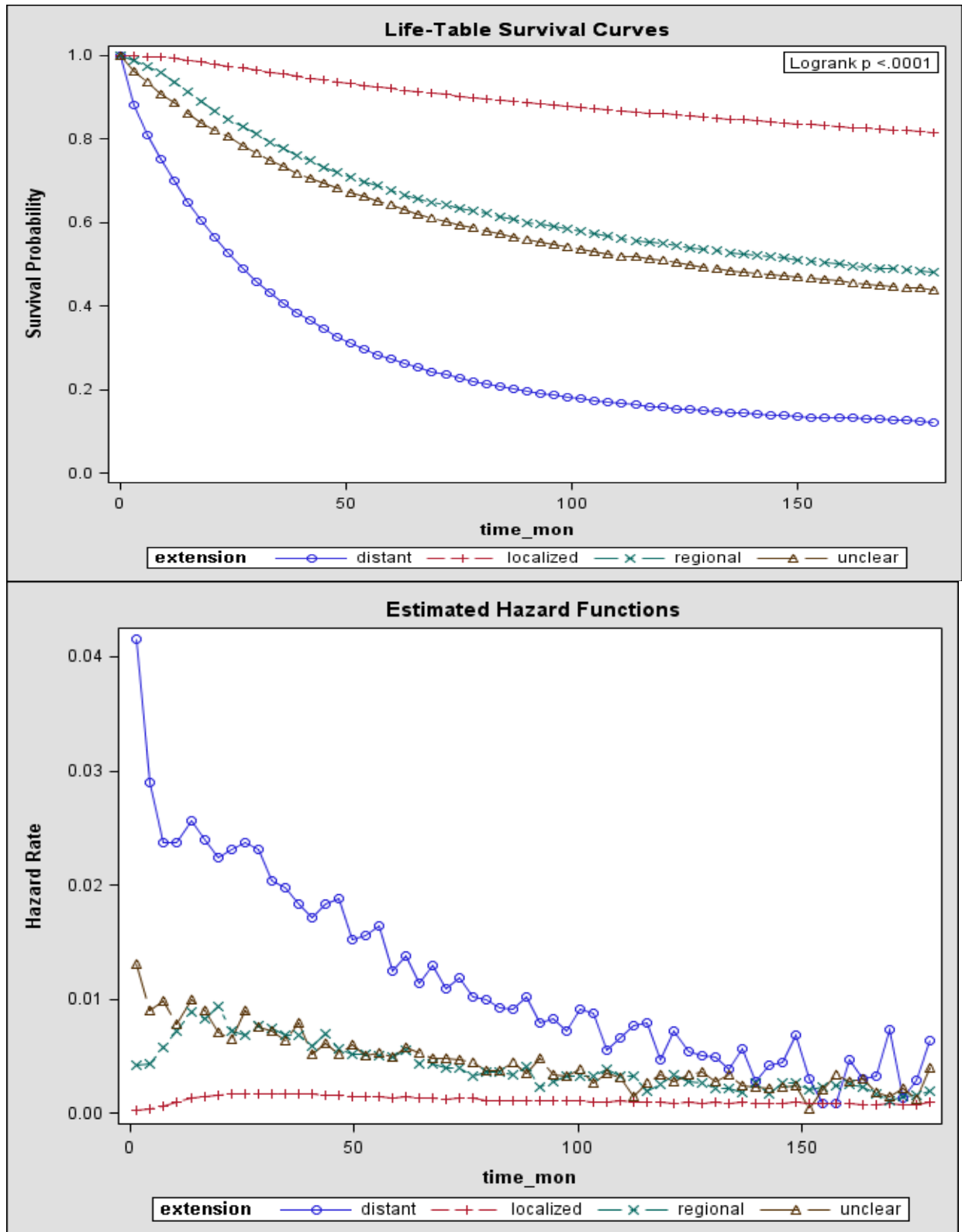
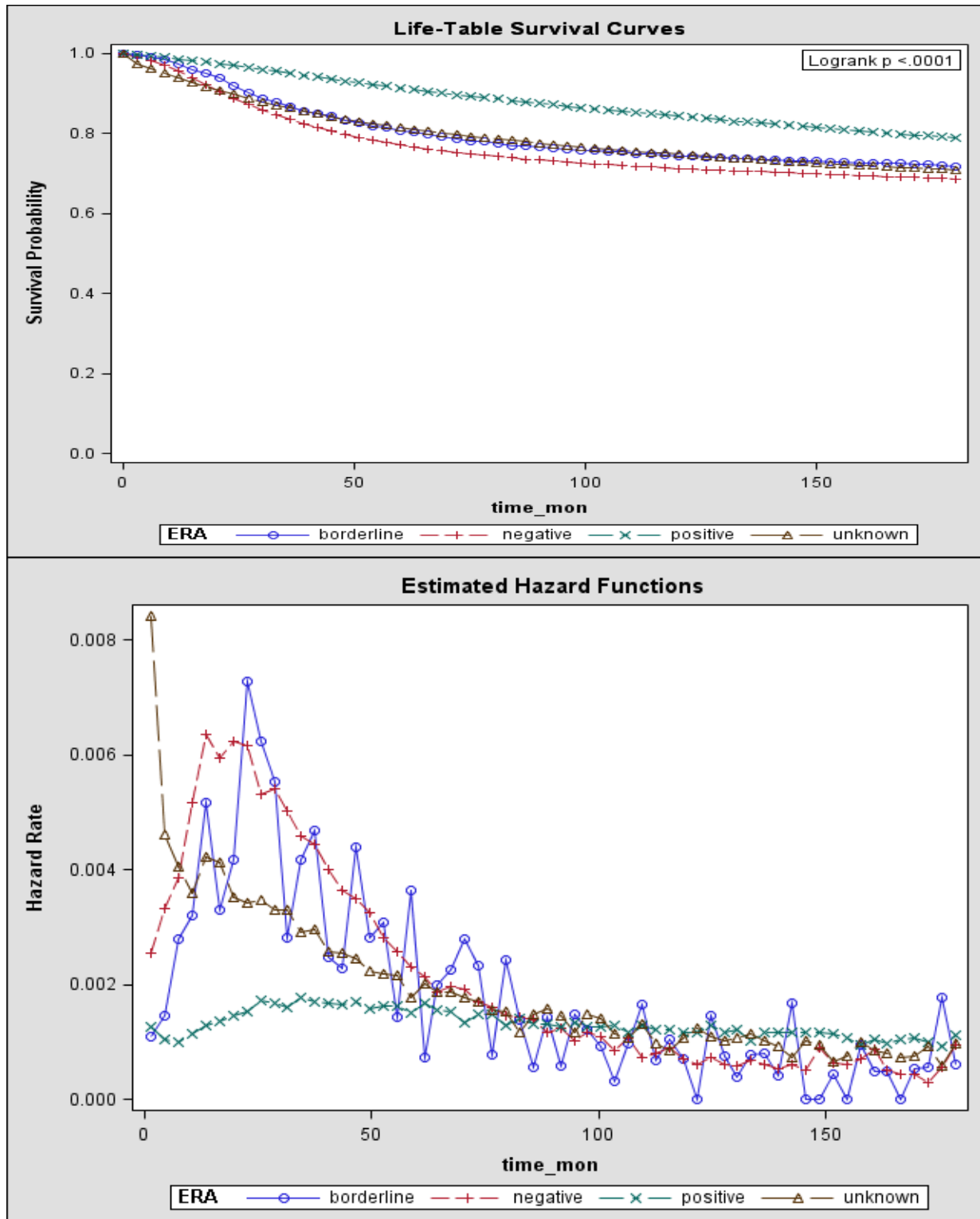
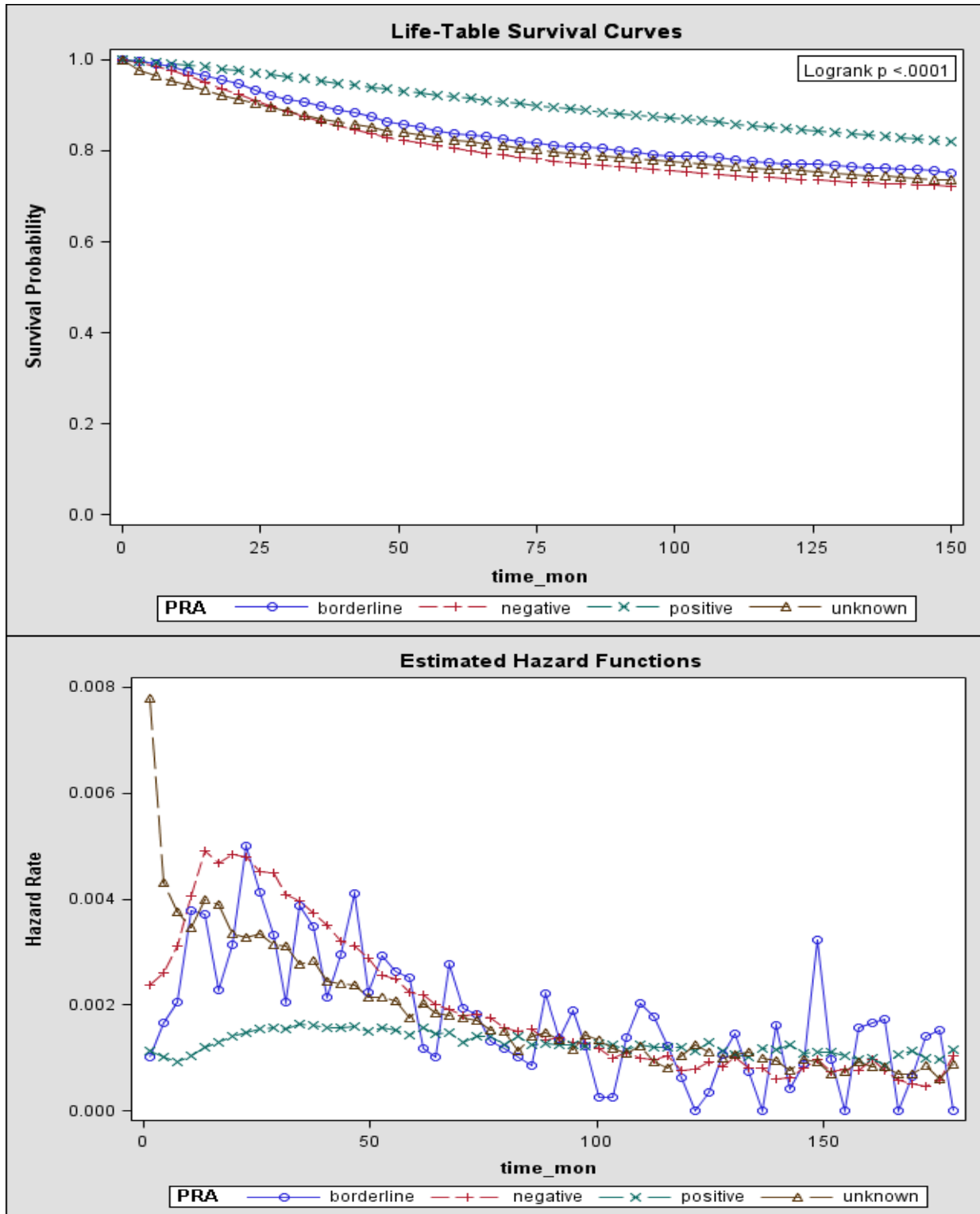


Figure 3.9 The effect of tumor markers on survival and hazard function





### 3.2.4 The effect of treatments on survival function and hazard function

Figure 3.10 The effect of surgery treatment on survival and hazard function

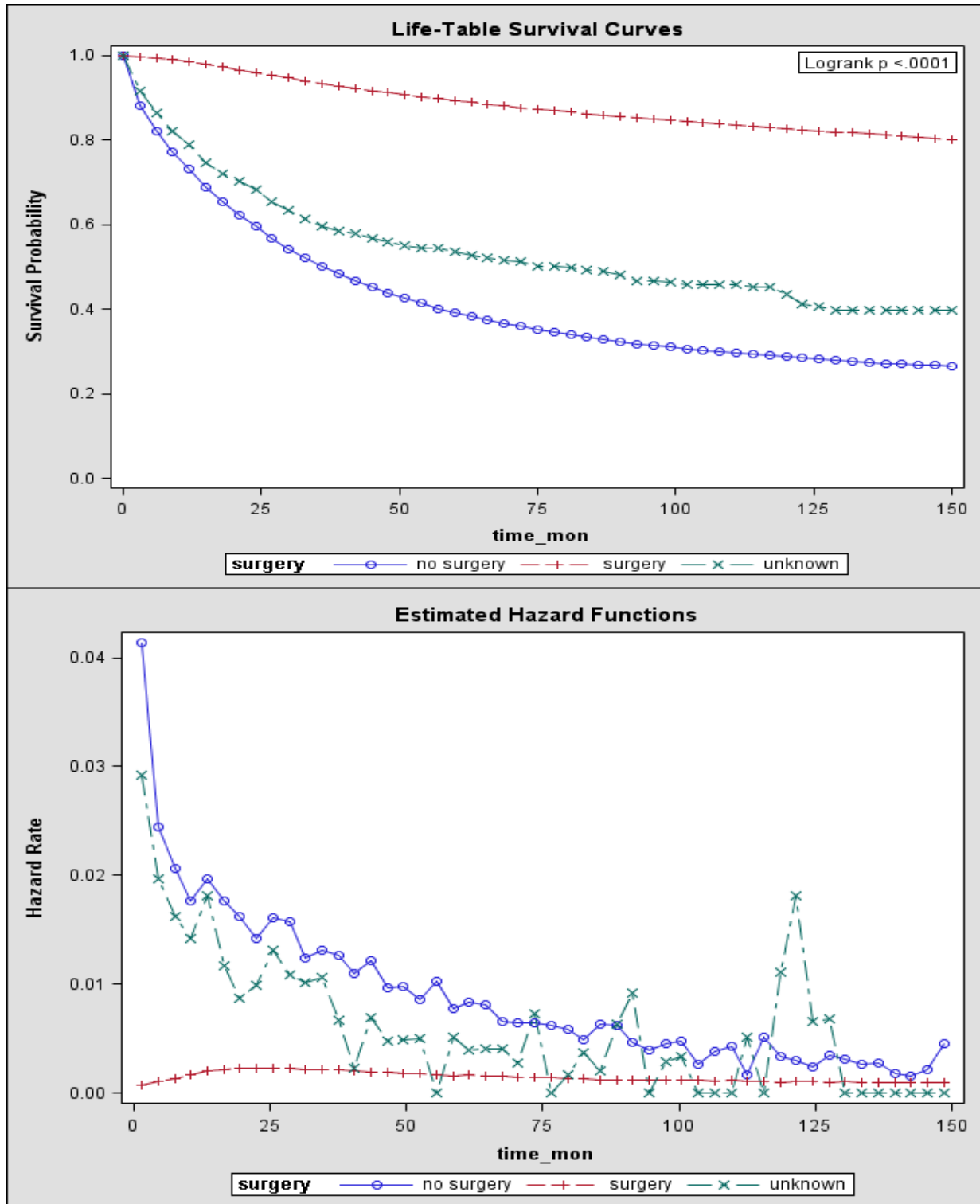


Figure 3.11 The effect of radiation treatment on survival and hazard function

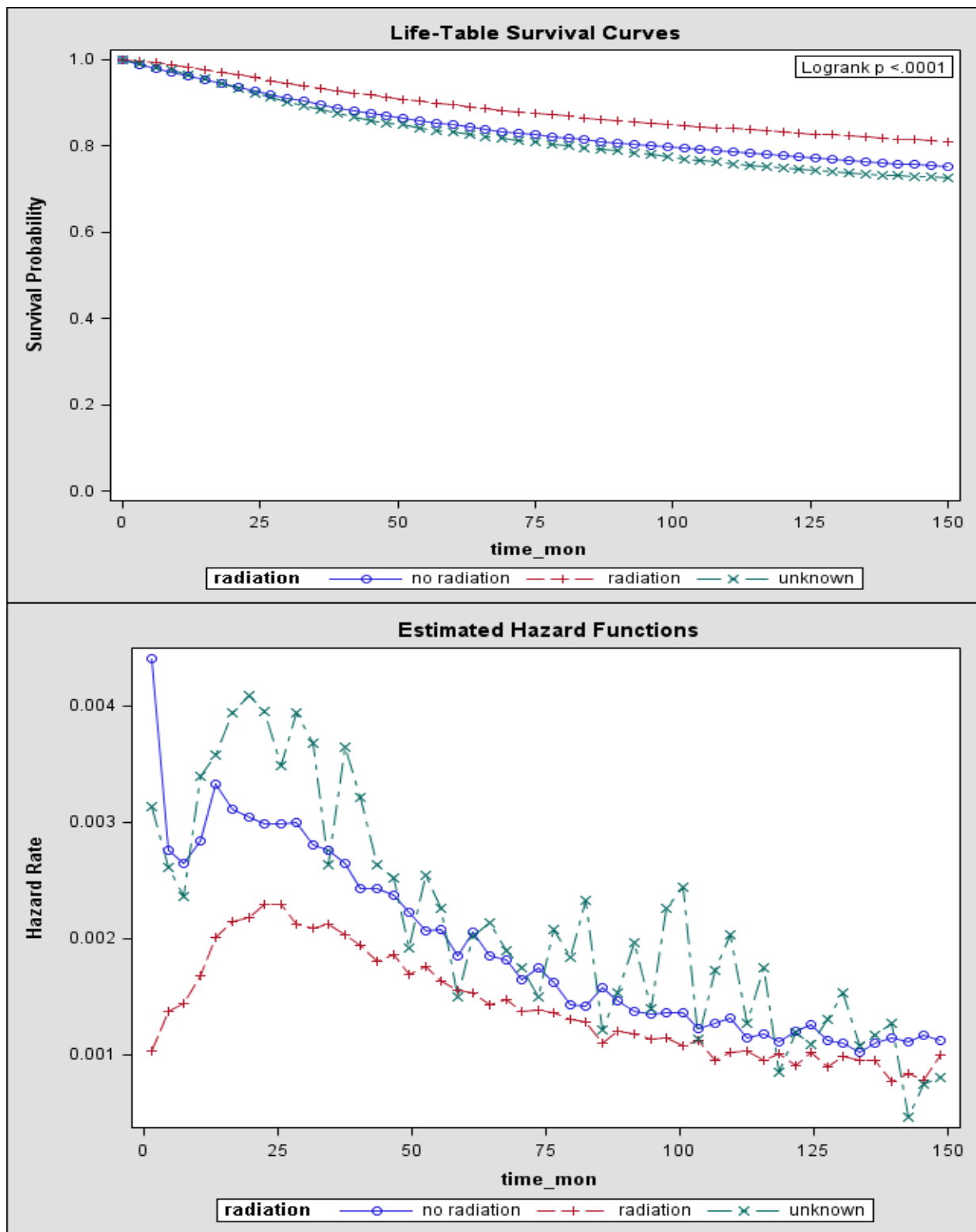
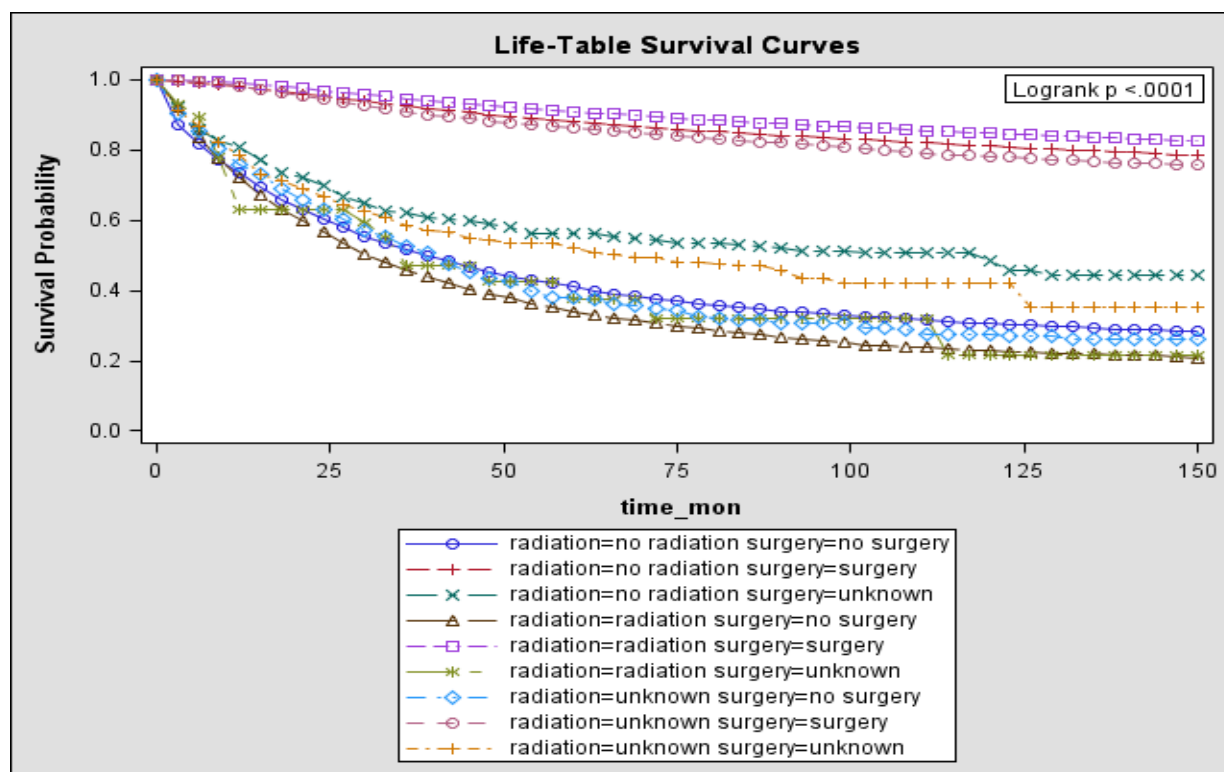


Figure 3.12 The effect of treatments on survival function



Figures 3.10 and 3.11 showed that both surgery and radiation treatments significantly improved the chance of surviving ( $P < 0.0001$ ). The hazard curve of the surgery group maintained a low and flat pattern. However, hazard rate of the non-surgery group was very high at the beginning and dropped fast, after 15 months, the hazard curve maintained a slow decreasing pattern (table 3.10). The radiotherapy group maintained the lowest hazard rate, the peak of the hazard curve appeared around 25 months and then the hazard rate slowly decreased (table 3.11). The patients received both surgery and radiation treatments had the best survival ability compared to those who only received radiotherapy or did not received any surgical or radiation treatment (figure 3.12,  $P < 0.0001$ ), suggesting that the treatments patients received are an important prognosis factor that can improve the survival chance of breast cancer patients.

### 3.3 Cox proportional hazard model and hazard ratio

In this study, all the variables: demographic variables (age race/ethnic), social variable (marital status), histopathologic variables (tumor cell differentiation grade, tumor size, lymph node or distant organ metastasis and disease stage) and clinical treatments can be used as covariates to explain the response variable. Cox proportional hazard regression is a powerful method to examine the effect of variables by controlling other covariates in a complex dataset (SAS Institute Inc. 2012). In this section, the Cox proportional hazard model was employed to check the effect of all risk and prognosis factors on disease specific and overall survival of invasive breast cancer patients and to calculate the hazard ratios of all the factors. First, a reference category in each factor was chosen and compared to the other levels in the same factor. The reference set of covariates contained non-Hispanic White, 20-39 year-old, married, tumor cell well differentiated, tumor size 2cm or less, localized extension, negative lymph node metastasis, positive tumor makers, surgery treatment and radiotherapy. Tables 17, 18 and 19 showed the obtained relative hazard ratio among the categories for all the factors.

We can found that African Americans had relatively higher hazard ratio compared to Whites and other race/ethnicity groups on both breast cancer specific survival and overall survival, which were 1.352 and 1.235 respectively ( $P < 0.0001$ , table 17). The 20-39 year-old age group had the highest mortality risk on breast cancer specific survival ( $P < 0.0001$ ). The 65 and above age group had the lowest survival chance on overall survival (3.965,  $P < 0.0001$ ). Compared to married people, both single and currently single patients had significantly higher death hazard on breast cancer specific and overall survival ( $P < 0.0001$ ). For example, for breast cancer specific survival, compared to married people, the hazard ratio of single patients was 1.208 ( $P < 0.0001$ ); for overall survival, the hazard ratio was 1.405 ( $P < 0.0001$ ).

Table 17. The effect of age race/ethnicity and marital status on hazard ratio

Factors	Breast cancer specific survival			Overall survival		
	Hazard Ratio	95% C.I.	P-value	Hazard Ratio	95% C.I.	P-value
<i>age</i>						
20-39 (reference)	1.000			1.000		
40-49	0.751	0.726 0.776	<.0001	0.982	0.948 1.107	0.3042
50-64	0.709	0.682 0.737	<.0001	1.456	1.392 1.523	<.0001
65 and above	0.786	0.748 0.826	<.0001	3.965	3.740 4.204	<.0001
<i>Race/ethnicity</i>						
White (reference)	1.000			1.000		
Asian	0.907	0.835 0.984	0.0186	0.814	0.764 0.847	<.0001
Black	1.352	1.306 1.400	<.0001	1.235	1.202 1.268	<.0001
Hispanic	1.056	0.996 1.119	0.0678	1.001	0.957 1.047	0.9715
Native	1.336	1.151 1.551	0.0001	1.232	1.100 1.380	0.0003
Other	0.337	0.159 0.714	0.0045	0.444	0.280 0.702	0.0005
<i>Marital</i>						
Married (reference)	1.000					
Ever married	1.241	1.212 1.270	<.0001	1.495	1.472 1.518	<.0001
Single	1.208	1.162 1.257	<.0001	1.405	1.367 1.445	<.0001
Unknown	1.061	1.000 1.216	0.0496	1.309	1.257 1.364	<.0001

Table 18. The effect of histological and pathological factors on hazard ratio

Factors	Breast cancer specific survival			Overall survival		
	Hazard Ratio	95% C.I.	P-value	Hazard Ratio	95% C.I.	P-value
<i>Cell differentiation grade</i>						
Well (reference)	1.000			1.000		
Moderately	1.756	1.680 1.836	<.0001	1.185	1.137 1.234	<.0001
Poorly differentiated	2.456	2.324 2.596	<.0001	1.513	1.403 1.632	<.0001
Undifferentiated	2.287	2.108 2.482	<.0001	1.489	1.326 1.671	<.0001
Unknown	1.608	1.469 1.760	<.0001	1.302	1.125 1.507	0.0004
<i>Tumor size</i>						
2cm and less (reference)	1.000			1.000		
Paget's disease	0.347	0.191 0.628	0.0005	2.046	1.988 2.254	0.0004
2-5cm	1.734	1.688 1.781	<.0001	2.116	1.412 1.449	<.0001
Over 5cm	1.745	1.666 1.828	<.0001	3.530	3.121 3.992	<.0001
Unknown	1.178	1.077 1.289	0.0003	5.404	4.252 6.867	<.0001
<i>Lymph node</i>						
Negative (reference)	1.000					
Positive	2.725	2.656 2.794	<.0001	1.751	1.715 1.788	<.0001
Unknown	2.132	2.037 2.232	<.0001	2.160	2.082 2.224	<.0001
<i>Tumor Extension</i>						
Localized (reference)	1.000					
Distant	4.294	4.082 4.517	<.0001	3.188	3.057 3.325	<.0001
Regional	1.883	1.821 1.948	<.0001	1.580	1.540 1.622	<.0001
Unclear	1.154	1.059 1.258	0.0011	1.005	0.937 1.078	0.8971
<i>Tumor Marker PRA</i>						

PRA positive (reference)	1.000						
Negative	1.257	1.228 1.287	<.0001	1.136	1.110 1.163	<.0001	
Borderline	1.304	1.197 1.421	<.0001	1.146	1.067 1.232	0.0002	
Unknown	1.018	0.963 1.075	0.5280	0.936	0.874 1.003	0.0624	
<i>Tumor Marker ERA</i>							
ERA positive (reference)							
Negative	1.363	1.330 1.397	<.0001	1.169	1.130 1.208	<.0001	
Borderline	1.298	1.180 1.427	<.0001	1.060	0.968 1.161	0.2082	
Unknown	1.216	1.149 1.285	<.0001	1.148	0.963 1.367	0.1227	

Table 19. The effect of treatments on hazard ratio

Factors	Breast cancer specific survival			Overall survival		
	Hazard Ratio	95% C.I.	P value	Hazard Ratio	95% C.I.	P value
<i>Surgery</i>						
Surgery (reference)	1.000			1.000		
No surgery	2.165	2.108 2.223	<.0001	1.370	1.283 1.463	<.0001
Unknown	2.229	1.977 2.513	<.0001	2.314	2.026 2.643	<.0001
<i>Radiation</i>						
Radiation (reference)	1.000			1.000		
No radiation	1.296	1.244 1.349	<.0001	1.508	1.471 1.546	<.0001
Unknown	1.039	0.981 1.101	0.1907	1.113	1.069 1.158	<.0001

Table 18 showed that for breast cancer specific survival, the well differentiated tumor had a lower death hazard, and poorly differentiated tumor (2.456,  $P < 0.0001$ ) and undifferentiated tumor (2.287,  $P < 0.0001$ ) had higher mortality risk. Tumor size, lymph node metastasis status and tumor extension had significant effect on breast cancer survival. The hazard ratio for comparing tumor size 2-5cm

and less than 2cm tumor was 1.734 ( $P<0.0001$ ) and the hazard ratio of tumor size over 5cm group compared to less than 2cm tumor group was 1.745 ( $P<0.0001$ ). Positive lymph node metastasis had relatively higher hazard ( $P<0.0001$ ), the ratios were 2.725 for breast cancer specific survival and 1.751 for overall survival, respectively. Distant tumor extension had the highest risk for both breast cancer specific and overall survival ( $P<0.0001$ ). The hazard ratios were 4.294 and 3.188, respectively. Patients with positive tumor marker (ERA or PRA) expression had better survival chance. Table 19 indicated that both surgery and radiotherapy were associated with lower mortality risk. For breast cancer specific survival, the hazard ratio of non-surgery group compared to patients who received surgery treatment was 2.165 ( $P<0.0001$ ) and the hazard ratio between non-radiotherapy group and radiotherapy group was 1.296 ( $P<0.0001$ ).

In summary, we found that after adjusted for the demographic factors, social factor and biomedical prognosis factors, the age of 20-39 year-old was a significant risk factor for breast cancer specific survival, and the 65 year-old and above age group had the worst overall survival chance. Among all race/ethnicity groups, the African Americans had the highest mortality risk (table 17) for breast cancer survival or overall survival. For both breast cancer specific survival and overall survival, larger tumor size, poorly differentiated tumor or undifferentiated tumor, lymph node and/or distant metastasis, and negative tumor markers associated with higher mortality risk. On the contrary, small tumor (less than 2cm), well differentiation tumor cell, no lymph node and/or distant organ metastasis, positive tumor markers and surgery or radiotherapy associated with less risk of death (tables 18 and 19).

### 3.4 Logistic regression analysis for 5-year survival status and the odds ratio

In this part, a logistic regression analysis was used to seek the effects of demographic factors, social factor, histopathologic factors and clinical treatment factors on 5-year overall survival of patients and to calculate the odds ratio. In this section, a SAS logistic procedure was used to do regression analysis (SAS institute inc. 2012).

Table 20. Relative odds ratio of demographic factors and social factor

	<i>Odds Ratio</i>	<i>95% confidence interval</i>		<i>P-value</i>
<i>age</i>				
20-39 (reference)	1.000			
40 to 49	1.234	1.176	1.295	<.0001
50 to 64	0.968	0.916	1.022	0.2380
65 and above	0.368	0.344	0.393	<.0001
<i>Race/ethnicity</i>				
Non- Hispanic White (reference)	1.000			
Asian	1.298	1.168	1.441	<.0001
African American	0.723	0.691	0.757	<.0001
Hispanic	1.031	0.956	1.111	0.4302
Native	0.747	0.616	0.906	0.0031
Other	4.949	2.046	10.556	0.0002
<i>Marital status</i>				
Married (reference)	1.000			
Ever married	0.600	0.584	0.617	<.0001
Single	0.667	0.636	0.700	<.0001
Unknown	0.699	0.651	0.750	<.0001

Table 20 showed the results of logistic analysis for the 5-year survival status variable. After adjusted by all the other risk and prognosis factors, the odds ratio of over 65-year-old group compared to 20-39 year-old group was 0.368 ( $P < 0.0001$ ); the odds ratio of 40 to 49 years old group compared to 20-39 year-old group was 1.234 ( $P < 0.0001$ ); the odds ratio of 50 to 64 years old group compared to 20-39 year-old group was 0.968 ( $P = 0.2380$ ). We concluded that the over 65 year-old age group had the lowest 5-year survival probability. African Americans had a lower 5-year survival chance, the odds ratio compared to Non-Hispanic Whites was 0.723 ( $P < 0.0001$ ). Compared with married people, single (Odds Ratio 0.667,  $P < 0.0001$ ) and currently single (Odds Ratio 0.600,  $P < 0.0001$ ) patients also had a lower 5-year survival chance.

Table 21 showed that well differentiated tumor group had better 5-year relative survival probability ( $P < 0.0001$ ). The odd ratios for poorly differentiated tumor group and undifferentiated tumor group were 0.573 and 0.585, respectively. The odds ratio for the comparison between positive lymph node group and negative group was 0.446 ( $P < 0.0001$ ). The odds ratios for regional and distant tumor extension were 0.464 and 0.144, respectively. All the results indicated that early disease stage (without lymph node metastasis or distant organ metastasis, smaller tumor size, localized tumor) had a better 5-year survival chance. The tumor marker ERA and PRA also affected survival: compared to positive ERA, the odds ratio of negative ERA was 0.646; compare to positive PRA, the odds ratio of negative PRA was 0.774.

Table 22 showed that the clinical treatments (Surgery and/or Radiotherapy) had positive effect on the 5-year survival of breast cancer patients. Compared to the patients who received surgery treatment, the odds ratio of the patients who did not receive any surgical treatment was 0.558 ( $P < 0.0001$ ) and also the radiotherapy had a positive effect on survival chance. Compared to radiotherapy group, the odds ratio of non-radiotherapy group was only 0.492 ( $P < 0.0001$ ).

Table 21. Relative odds ratio of histopathologic prognosis factors

	<i>Odds Ratio</i>	<i>95% confidence interval</i>		<i>P-value</i>
<i>Tumor cell differentiation grade</i>				
Well differentiated (reference)	1.000			
Moderately	0.846	0.786	0.912	<.0001
Poorly	0.573	0.499	0.658	<.0001
Undifferentiated	0.585	0.473	0.724	<.0001
Unknown	0.851	0.648	1.117	0.2441
<i>Tumor size</i>				
2cm and less (reference)	1.000			
Paget disease	2.161	1.331	3.508	0.0018
2 to 5cm	0.585	0.569	0.602	<.0001
over 5cm	0.467	0.443	0.493	<.0001
Unknown	0.868	0.788	0.955	0.0037
<i>Lymph node</i>				
Negative (reference)	1.000			
Positive	0.446	0.436	0.456	<.0001
Unknown	0.335	0.327	0.344	<.0001
<i>Tumor Marker ERA</i>				
ERA positive (reference)	1.000			
Borderline	0.649	0.570	0.738	<.0001
Negative	0.646	0.623	0.670	<.0001
Unknown	0.692	0.608	0.788	<.0001
<i>Tumor Marker PRA</i>				
PRA positive (reference)	1.000			

Borderline	0.755	0.677	0.841	<.0001
Negative	0.774	0.751	0.798	<.0001
Unknown	1.038	0.962	1.120	0.3359
<i>Tumor extension</i>				
Localized (reference)	1.000			
Distant	0.144	0.123	0.168	<.0001
Regional	0.464	0.426	0.504	<.0001
Unclear	0.753	0.602	0.941	0.0128

Table 22. Relative odds ratio of clinical treatments factors

	<i>Odds Ratio</i>	<i>95% confidence interval</i>		<i>P-value</i>
<i>Surgery</i>				
Surgery treatment (reference)	1.000			
No surgery	0.558	0.489	0.635	<.0001
Unknown	0.323	0.250	0.416	<.0001
<i>radiotherapy</i>				
Radiotherapy (reference)	1.000			
No radiation	0.492	0.468	0.517	<.0001
Unknown	0.951	0.884	1.024	0.1828

Table 23. The raw odds ratio and adjusted odds ratio for some risk and prognosis factors

Factors	Raw odds ratio and 95% C.I.	adjusted odds ratio and 95% C.I. <sup>a</sup>	adjusted odds ratio and 95% C.I. <sup>b</sup>
<i>age</i>			
20-39	0.728 (0.702 0.755)	0.750 (0.722 0.778)	0.951 (0.912 0.992)
40-49	1.097 (1.067 1.127)	1.101 (1.071 1.132)	1.222 (1.185 1.261)
Over 65	0.404 (0.396 0.412)	0.446 (0.437 0.454)	0.393 (0.384 0.402)
<i>Race/ethnicity</i>			
Hispanic	1.001 (0.971 1.032)	0.858 (0.832 0.885)	0.957 (0.917 0.997)
Black	0.570 (0.556 0.585)	0.545 (0.531 0.560)	0.709 (0.685 0.735)
Asian	1.489 (1.436 1.545)	1.231 (1.186 1.278)	1.109 (1.054 1.167)
Native	0.894 (0.797 1.003)	0.761 (0.676 0.856)	0.641 (0.546 0.752)
<i>Social factor</i>			
married	1.742 (1.704 1.782)	1.642 (1.605 1.681)	1.274 (1.241 1.309)
Ever married	0.742 (0.725 0.759)	0.920 (0.899 0.943)	0.811 (0.789 0.834)
<i>Cell grade</i>			
Well	1.477 (1.435 1.521)	1.555 (1.509 1.602)	1.187 (1.149 1.226)
poorly	0.516 (0.506 0.526)	0.454 (0.445 0.464)	0.661 (0.645 0.677)
undifferentiated	0.528 (0.503 0.555)	0.450 (0.428 0.474)	0.659 (0.622 0.698)
<i>Tumor size</i>			
2cm or less	1.320 (0.850 2.050)	1.108 (0.707 1.737)	0.488 (0.303 0.787)
2-5	0.523 (0.336 0.812)	0.413 (0.264 0.648)	0.279 (0.173 0.450)
Over 5cm	0.191 (0.123 0.297)	0.139 (0.089 0.218)	0.219 (0.136 0.353)
<i>Lymph node</i>			
negative	3.470 (3.404 3.537)	3.799 (3.725 3.876)	2.249 (2.199 2.299)

<i>Extension</i>			
local	5.341 (5.139 5.550)	4.697 (4.512 4.891)	1.332 (1.262 1.405)
Regional	1.265 (1.207 1.326)	1.205 (1.147 1.265)	0.639 (0.602 0.679)
distant	0.298 (0.284 0.313)	0.237 (0.225 0.250)	0.201 (0.189 0.213)
<i>Tumor marker</i>			
ERA positive	1.467 (1.430 1.504)	1.794 (1.747 1.843)	1.538 (1.492 1.585)
PRA negative	0.941 (0.860 1.029)	0.936 (0.854 1.027)	1.107 (0.919 1.125)
PRA positive	1.393 (1.273 1.524)	1.318 (1.201 1.446)	1.305 (1.179 1.145)
<i>Treatments</i>			
Surgery	4.440 (3.770 5.229)	4.060 (3.431 4.804)	2.117 (1.909 2.349)
Non surgery	0.447 (0.378 0.527)	0.434 (0.366 0.515)	0.391 (0.374 0.408)
Radiotherapy	1.805 (1.774 1.836)	1.594 (1.566 1.622)	1.592 (1.561 1.624)

Note: <sup>a</sup> adjusted by demographic factors and social factors; <sup>b</sup> adjusted by demographic factors, social factors and other prognosis factors.

From table 23 we found that after adjustment, the oldest age group patients still had a lower odds ratio (0.393), indicating a lower 5-year survival probability ( $P < 0.0001$ ). African American also had the lowest odds ratio among all race/ethnicity groups (0.709,  $P < 0.0001$ ). This indicated that African American had lower 5-year relative survival chance. The subjects with well differentiated tumors, positive tumor markers, small tumor size or localized tumor had relatively higher odds ratios and better survival chance. Surgery and radiation treatments were all able to improve the 5-year survival chance. However, poorly differentiated or undifferentiated tumors, tumor size over 5cm and regional or distant extension had negative effects on breast cancer 5-year survival. Patients who did not receive any surgical treatments had lower 5-year survival chance (odds ratio 0.391).

#### **4 RESEACH DEFICIENCY AND FURTHER PROSPECT**

In this study, some limitations must be considered. First of all, in this study, the majority of the participants were non-Hispanic White patients, and other race/ethnicity groups were less than 10%, resulting in a significant effect on the p-value when we do survival analysis. On the other hand, in breast cancer research, some other important risk factors should be considered, such as the patient's personal history and patient's social-economic status (SES). Postmenopausal age is considered as an important risk factor in breast cancer and highly associated with disease prognosis (Mandelblatt and Andrews 1991). The preventable death rate was considered significantly higher in lower SES population because of less disease-related information they knew and less early detection and examination they received (Farley and Flannery, 1989). However, the SEER study program does not include such important information. We expect in future study, the patients' SES information, such as gross household income, education level and the patients' personal history to be included. Study containing some more factors will give us more accurate estimation of the prognosis and survival of breast cancer.

#### **5 CONCLUSIONS**

Despite some limitations, the results from this study provide evidence that many factors, such as demographic factors, social factors, and biomedical prognosis factors have significant effects on prognosis and survival of the female invasive breast cancer patients. The mortality increases in the older age groups. For example, postmenopausal females (over 50 year-old) have a higher risk of death. African American breast cancer patients have a higher mortality risk. Nodal involvement and distant metastasis, larger tumor size, negative tumor marker expression and tumor cell poorly differentiated or undifferentiated are all significant high-risk biomedical factors. Besides, the lymph node and distant metastasis and larger tumor size are the symptoms of later disease stage. The patients in earlier disease stage and/or with smaller tumor size have lower risk of death. We can not control some risk factors or high-risk prog-

nosis factors of breast cancer, such as the tumor cell differentiation grade and tumor markers. However, through early detection and valid treatments, such as surgery and radiotherapy, the risk of death can be effectively reduced. It was considered that the long-term survival of breast cancer patients was highly dependent on the stage of disease (National Cancer Institute 1990). Both Surgery and radiotherapy contribute to the survival of breast cancer patients. Based on the evidence shown in this study, we conclude that early detection and treatment can prolong the lifespan of breast cancer patients and also improve their life quality. Furthermore, it is important for women to detect the breast cancer at an early stage, before it metastasizes to lymph nodes and distant organs.

## REFERENCES

American Cancer Society, (2011), "Breast Cancer Facts & Figures 2011-2012", Atlanta: American Cancer Society, Inc. Available at:

<http://www.cancer.org/Research/CancerFactsFigures/BreastCancerFactsFigures/breast-cancer-facts-and-figures-2011-2012>

Albain, K., Allred, D., Clark, G. (1994), "Breast cancer outcome and predictors of outcome: are there age differentials?" *Journal of the National Cancer Institute Monograph*, 16, 35-42.

Anderson, W., Jatoi, I., Devesa, S. (2005), "Distinct breast cancer incidence and prognostic patterns in the NCI's SEER program: suggesting a positive link between etiology and outcome," *Breast Cancer Research and Treatment*, 90, 127-137.

Bassett, M., Krieger, N., (1986), "Social class and Black-White Difference in Breast Cancer Survival," *American Journal of Public Health*, 76, 1400-1403.

Bethesda, (1990) "Annual cancer statistics review, including cancer trends: 1950-1985", National Cancer Institute.

CDC, (2000) "Cancer Mortality Surveillance in United States, 1990-2000", Available at:

<http://www.cdc.gov/mmwr/preview/mmwrhtml/ss5303a1.htm#tab1>

Cox, D.R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society-B*, 34, 187-220.

Diab, S., Elledge, R., Clark, G. (2000), "Tumor characteristics and clinical outcome of elderly women with breast cancer," *Journal of the National Cancer Institute*, 92(7), 550-556.

Early Breast Cancer Trialists' Collaborative Group (2005), "Effects of radiotherapy and differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomized trials," *The lancet*, 366, 2087-2106.

Elston, C.W., Ellis, I.O., (1991) "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, 19, 403-410.

Farley, T., Flannery, J., (1989) "Late-stage diagnosis of breast cancer in women of lower socioeconomic status: public health implications", *American Journal of Public Health*, 79, 1508-1512.

Howlander, N., Noone, M. , Krapcho, M. , Neyman, N. et al. (2011) " SEER Cancer Statistics Review, 1975-2008", National Cancer Institute. Bethesda, MD, Available at:  
[http://seer.cancer.gov/csr/1975\\_2008/](http://seer.cancer.gov/csr/1975_2008/)

Jatoi, I., Tsimelzon, A. et al. (2005) "Hazard rates of recurrences following diagnosis of primary breast cancer," *Breast Cancer Research and Treatment*, 89, 173-178.

Jeanne Mandelblatt, et al (1991) "Determinants of late stage diagnosis of Breast cancer and Cervical cancer: the impact of age race social class and hospital type," *American Journal of Public Health*, 81, 646-649.

Klein, J.P., Moeschberger, M.L., (2003), "Survival Analysis Techniques for Censored and Truncated Data." (2nd ed), Springer-Verlag New York, Inc., 244-245.

Li, Cl., et al. (2005) "Clinical characteristics of different histologic types of breast cancer," *British Journal of Cancer*, 93, 1046-1052.

Mandelblatt, J., Andrews, H. et al. (1991) "Determinants of late stage diagnosis of breast cancer and cervical cancer: The impact of age, race, social class and hospital type", *American Journal of Public Health*, 81, 646-649.

SAS Institute Inc. (2012) "SAS/STAT(R) 9.2 User's Guide, Second Edition", NC: SAS Institute Inc.

Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1973-2008), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, Available at:  
<http://www.seer.cancer.gov>

## APPENDICES

SAS code:

```

option nodate nocenter;
libname seerbc 'D:\whw\data\seerbc';
libname format 'D:\whw\data\seerbc\format';
/*creast sas dataset*/
/*cases from San francisco, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico,Seattle, Utah, Atlanta
1973-2008*/
data seerbc.breast1;
infile 'F:\thesis\thesis\thesis\SEER_1973_2008_TEXTDATA\incidence\yr1973_2008.seer9\breast.txt' lrecl=279;
input @ 1 CASENUM 8. /* Patient ID */
      @ 9 REG 10. /* SEER registry */
      @ 19 MAR_STAT 1. /* Marital status at diagnosis */
      @ 20 RACE 2. /* Race/ethnicity */
      @ 22 ORIGIN 1. /* Spanish surname or origin */
      @ 23 NHIA 1. /* NHIA Derived Hisp Origin */
      @ 24 SEX 1. /* Sex */
      @ 25 AGE_DX 3. /* Age at diagnosis */
      @ 28 YR_BIRTH 4. /* Year of birth */
      @ 32 PLC_BRTH 3. /* Place of birth */
      @ 35 SEQ_NUM 2. /* Sequence number */
      @ 37 DATE_mo 2. /* Month of diagnosis */
      @ 39 DATE_yr 4. /* Year of diagnosis */
      @ 43 SITEO2V $char4. /* Primary site ICD-O-2 (1973+) */
      @ 47 LATERAL 1. /* Laterality */
      @ 48 HISTO2V 4. /* Histologic Type ICD-O-2 */
      @ 52 BEHO2V 1. /* Behavior Code ICD-O-2 */
      @ 53 HISTO3V 4. /* Histologic Type ICD-O-3 */
      @ 57 BEHO3V 1. /* Behavior code ICD-O-3 */
      @ 58 GRADE 1. /* Grade */
      @ 59 DX_CONF 1. /* Diagnostic confirmation */
      @ 60 REPT_SRC 1. /* Type of reporting source */
      @ 61 EOD10_SZ 3. /* EOD 10 - size (1988+) */
      @ 64 EOD10_EX 2. /* EOD 10 - extension */
      @ 66 EOD10_PE 2. /* EOD 10 - path extension */
      @ 68 EOD10_ND 1. /* EOD 10 - lymph node */
      @ 69 EOD10_PN 2. /* EOD 10 - positive lymph nodes examined */
      @ 71 EOD10_NE 2. /* EOD 10 - number of lymph nodes examined */
      @ 73 EOD13 $char13. /* EOD--old 13 digit */
      @ 86 EOD2 $char2. /* EOD--old 2 digit */
      @ 88 EOD4 $char4. /* EOD--old 4 digit */
      @ 92 EODCODE $char1. /* Coding system for EOD */
      @ 93 TUMOR_1V $char1. /* Tumor marker 1 */
      @ 94 TUMOR_2V $char1. /* Tumor marker 2 */
      @ 95 TUMOR_3V $char1. /* Tumor marker 3 */
      @ 96 CS_SIZE 3. /* CS Tumor size */
      @ 99 CS_EXT 3. /* CS Extension */
      @ 102 CS_NODE 3. /* CS Lymph Nodes */
      @ 105 CS_METS 2. /* CS Mets at DX */
      @ 107 CS_SSF1 3. /* CS Site-Specific Factor 1 */
      @ 110 CS_SSF2 3. /* CS Site-Specific Factor 2 */

```

@ 113 CS\_SSF3 3. /\* CS Site-Specific Factor 3 \*/  
 @ 116 CS\_SSF4 3. /\* CS Site-Specific Factor 4 \*/  
 @ 119 CS\_SSF5 3. /\* CS Site-Specific Factor 5 \*/  
 @ 122 CS\_SSF6 3. /\* CS Site-Specific Factor 6 \*/  
 @ 125 CS\_SSF25 3. /\* CS Site-Specific Factor 25 \*/  
 @ 128 D\_AJCC\_T 2. /\* Derived AJCC T \*/  
 @ 130 D\_AJCC\_N 2. /\* Derived AJCC N \*/  
 @ 132 D\_AJCC\_M 2. /\* Derived AJCC M \*/  
 @ 134 D\_AJCC\_S 2. /\* Derived AJCC Stage Group \*/  
 @ 136 D\_SSG77 1. /\* Derived SS1977 \*/  
 @ 137 D\_SSG00 1. /\* Derived SS2000 \*/  
 @ 138 D\_AJCC\_F 1. /\* Derived AJCC - flag \*/  
 @ 139 D\_SSG77F 1. /\* Derived SS1977 - flag \*/  
 @ 140 D\_SSG00F 1. /\* Derived SS2000 - flag \*/  
 @ 141 CSV\_ORG 6. /\* CS Version Input Original \*/  
 @ 147 CSV\_DER 6. /\* CS Version Derived \*/  
 @ 153 CSV\_CUR 6. /\* CS Version Input Current \*/  
 @ 159 SURGPRIM 2. /\* RX Summ--surg prim site \*/  
 @ 161 SCOPE 1. /\* RX Summ--scope reg LN sur \*/  
 @ 162 SURGOTH 1. /\* RX Summ--surg oth reg/dis \*/  
 @ 163 SURGNODE 2. /\* Number of lymph nodes \*/  
 @ 165 RECONST 1. /\* Reconstruction \*/  
 @ 166 NO\_SURG 1. /\* Reason no cancer-directed surgery \*/  
 @ 167 RADIATN 1. /\* Radiation \*/  
 @ 168 RAD\_BRN 1. /\* Radiation to Brain and/or CNS \*/  
 @ 169 RAD\_SURG 1. /\* Radiation sequence with surgery \*/  
 @ 170 SS\_SURG 2. /\* Site specific surgery (1983-1997) \*/  
 @ 172 SRPRIMO2 2. /\* Surgery to Primary Site \*/  
 @ 174 SCOPE02 1. /\* Scope of lymph node surgery \*/  
 @ 175 SRGOTH02 1. /\* Surgery to other sites \*/  
 @ 176 REC\_NO 2. /\* Record number \*/  
 @ 178 O\_SITAGE 1. /\* Age-site edit override \*/  
 @ 179 O\_SEQCON 1. /\* Sequence number-dx conf override \*/  
 @ 180 O\_SEQLAT 1. /\* Site-type-lat-seq override \*/  
 @ 181 O\_SURCON 1. /\* Surgery-diagnostic conf override \*/  
 @ 182 O\_SITYP 1. /\* Site-type edit override \*/  
 @ 183 H\_BENIGN 1. /\* Histology edit override \*/  
 @ 184 O\_RPTSRC 1. /\* Report source sequence override \*/  
 @ 185 O\_DFSITE 1. /\* Seq-ill-defined site override \*/  
 @ 186 O\_LEUKDX 1. /\* Leuk-Lymph dx confirmation override \*/  
 @ 187 O\_SITBEH 1. /\* Site-behavior override \*/  
 @ 188 O\_EODDT 1. /\* Site-EOD-dx date override \*/  
 @ 189 O\_SITEOD 1. /\* Site-laterality-EOD override \*/  
 @ 190 O\_SITMOR 1. /\* Site-laterality-morph override \*/  
 @ 191 TYPEFUP 1. /\* Type of followup expected \*/  
 @ 192 AGE\_REC 2. /\* Age recode \*/  
 @ 194 SITE\_REC 5. /\* Site recode \*/  
 @ 199 SITE2\_RE 5. /\* Site rec with Kaposi and mesothelioma \*/  
 @ 204 ICDOTO9V \$char4. /\* Recode ICD-O-2 to 9 \*/  
 @ 208 ICDOT10V \$char4. /\* Recode ICD-O-2 to 10 \*/  
 @ 212 ICCCSITE \$char3. /\* ICCC site recode \*/  
 @ 215 ICCCSM \$char3. /\* SEER modified ICCC site recode \*/  
 @ 218 ICCC3 \$char3. /\* ICCC ICD-O-3 Recode \*/

```

@ 221 ICC3EXT $char3. /* ICC3 ICD-O-3 Extended Recode */
@ 224 BEHANAL 1. /* Behavior recode for analysis */
@ 225 ICD_CODE 1. /* ICD-O Coding Scheme */
@ 226 HISTREC 2. /* Broad Histology recode */
@ 228 BRAINREC 2. /* Brain recode */
@ 230 CS202SCH 3. /* CS Schema v0202*/
@ 233 RAC_RECA 1. /* Race recode A */
@ 234 RAC_RECY 1. /* Race recode Y */
@ 235 NHIAREC 1. /* Origin Recode NHIA */
@ 236 HST_STGA 1. /* SEER historic stage A */
@ 237 AJCC_STG 2. /* AJCC stage 3rd edition (1988+) */
@ 239 AJ_3SEER 2. /* SEER modified AJCC stage 3rd ed (1988+) */
@ 241 SSG77 1. /* SEER Summary Stage 1977 (1995-2000) */
@ 242 SSG2000 1. /* SEER Summary Stage 2000 2000 (2001-2003) */
@ 243 NUMPRIMS 2. /* Number of primaries */
@ 245 FIRSTPRM 1. /* First malignant primary indicator */
@ 246 STCOUNTY 5. /* State-county recode */
@ 251 SURV_TM 4. /* Survival time recode */
@ 255 ICD_5DIG 5. /* Cause of death to SEER site recode */
@ 260 CODKM 5. /* COD to site rec KM */
@ 265 STAT_REC 1. /* Vital status recode (study cutoff used) */
@ 266 IHS 1. /* IHS link */
@ 267 hss_2000 1. /* Historic SSG 2000 Stage */
@ 268 AYA_rec 2. /* AYA recode */
@ 270 Lymphrec 2. /* Lymphoma recode */
@ 272 dth_cl 1. /* SEER cause of death classification */
@ 273 o_dth_cl 1. /* SEER other cause of death classification */
@ 274 exteval 1. /* CS EXT/Size Eval */
@ 275 nodeeval 1. /* CS Nodes Eval */
@ 276 metseval 1. /* CS Mets Eval */
;
run;
/*cases from Alaska, San Jose and Monterey, Los Angeles, Rural Georgia 1992-2008*/
data seerbc.breast2;
infile 'F:\thesis\thesis\thesis\SEER_1973_2008_TEXTDATA\incidence\yr1992_2008.sj_la_rg_ak\breast.txt'
lrecl=279;
input @ 1 CASENUM 8. /* Patient ID */
@ 9 REG 10. /* SEER registry */
@ 19 MAR_STAT 1. /* Marital status at diagnosis */
@ 20 RACE 2. /* Race/ethnicity */
@ 22 ORIGIN 1. /* Spanish surname or origin */
@ 23 NHIA 1. /* NHIA Derived Hisp Origin */
@ 24 SEX 1. /* Sex */
@ 25 AGE_DX 3. /* Age at diagnosis */
@ 28 YR_BRTH 4. /* Year of birth */
@ 32 PLC_BRTH 3. /* Place of birth */
@ 35 SEQ_NUM 2. /* Sequence number */
@ 37 DATE_mo 2. /* Month of diagnosis */
@ 39 DATE_yr 4. /* Year of diagnosis */
@ 43 SITEO2V $char4. /* Primary site ICD-O-2 (1973+) */
@ 47 LATERAL 1. /* Laterality */
@ 48 HISTO2V 4. /* Histologic Type ICD-O-2 */
@ 52 BEHO2V 1. /* Behavior Code ICD-O-2 */

```

@ 53 HISTO3V 4. /\* Histologic Type ICD-O-3 \*/  
 @ 57 BEHO3V 1. /\* Behavior code ICD-O-3 \*/  
 @ 58 GRADE 1. /\* Grade \*/  
 @ 59 DX\_CONF 1. /\* Diagnostic confirmation \*/  
 @ 60 REPT\_SRC 1. /\* Type of reporting source \*/  
 @ 61 EOD10\_SZ 3. /\* EOD 10 - size (1988+) \*/  
 @ 64 EOD10\_EX 2. /\* EOD 10 - extension \*/  
 @ 66 EOD10\_PE 2. /\* EOD 10 - path extension \*/  
 @ 68 EOD10\_ND 1. /\* EOD 10 - lymph node \*/  
 @ 69 EOD10\_PN 2. /\* EOD 10 - positive lymph nodes examined \*/  
 @ 71 EOD10\_NE 2. /\* EOD 10 - number of lymph nodes examined \*/  
 @ 73 EOD13 \$char13. /\* EOD--old 13 digit \*/  
 @ 86 EOD2 \$char2. /\* EOD--old 2 digit \*/  
 @ 88 EOD4 \$char4. /\* EOD--old 4 digit \*/  
 @ 92 EODCODE \$char1. /\* Coding system for EOD \*/  
 @ 93 TUMOR\_1V \$char1. /\* Tumor marker 1 \*/  
 @ 94 TUMOR\_2V \$char1. /\* Tumor marker 2 \*/  
 @ 95 TUMOR\_3V \$char1. /\* Tumor marker 3 \*/  
 @ 96 CS\_SIZE 3. /\* CS Tumor size \*/  
 @ 99 CS\_EXT 3. /\* CS Extension \*/  
 @ 102 CS\_NODE 3. /\* CS Lymph Nodes \*/  
 @ 105 CS\_METS 2. /\* CS Mets at DX \*/  
 @ 107 CS\_SSF1 3. /\* CS Site-Specific Factor 1 \*/  
 @ 110 CS\_SSF2 3. /\* CS Site-Specific Factor 2 \*/  
 @ 113 CS\_SSF3 3. /\* CS Site-Specific Factor 3 \*/  
 @ 116 CS\_SSF4 3. /\* CS Site-Specific Factor 4 \*/  
 @ 119 CS\_SSF5 3. /\* CS Site-Specific Factor 5 \*/  
 @ 122 CS\_SSF6 3. /\* CS Site-Specific Factor 6 \*/  
 @ 125 CS\_SSF25 3. /\* CS Site-Specific Factor 25 \*/  
 @ 128 D\_AJCC\_T 2. /\* Derived AJCC T \*/  
 @ 130 D\_AJCC\_N 2. /\* Derived AJCC N \*/  
 @ 132 D\_AJCC\_M 2. /\* Derived AJCC M \*/  
 @ 134 D\_AJCC\_S 2. /\* Derived AJCC Stage Group \*/  
 @ 136 D\_SSG77 1. /\* Derived SS1977 \*/  
 @ 137 D\_SSG00 1. /\* Derived SS2000 \*/  
 @ 138 D\_AJCC\_F 1. /\* Derived AJCC - flag \*/  
 @ 139 D\_SSG77F 1. /\* Derived SS1977 - flag \*/  
 @ 140 D\_SSG00F 1. /\* Derived SS2000 - flag \*/  
 @ 141 CSV\_ORG 6. /\* CS Version Input Original \*/  
 @ 147 CSV\_DER 6. /\* CS Version Derived \*/  
 @ 153 CSV\_CUR 6. /\* CS Version Input Current \*/  
 @ 159 SURGPRIM 2. /\* RX Summ--surg prim site \*/  
 @ 161 SCOPE 1. /\* RX Summ--scope reg LN sur \*/  
 @ 162 SURGOTH 1. /\* RX Summ--surg oth reg/dis \*/  
 @ 163 SURGNODE 2. /\* Number of lymph nodes \*/  
 @ 165 RECONST 1. /\* Reconstruction \*/  
 @ 166 NO\_SURG 1. /\* Reason no cancer-directed surgery \*/  
 @ 167 RADIATN 1. /\* Radiation \*/  
 @ 168 RAD\_BRN 1. /\* Radiation to Brain and/or CNS \*/  
 @ 169 RAD\_SURG 1. /\* Radiation sequence with surgery \*/  
 @ 170 SS\_SURG 2. /\* Site specific surgery (1983-1997) \*/  
 @ 172 SRPRIMO2 2. /\* Surgery to Primary Site \*/  
 @ 174 SCOPE02 1. /\* Scope of lymph node surgery \*/

@ 175 SRGOTH02 1. /\* Surgery to other sites \*/  
 @ 176 REC\_NO 2. /\* Record number \*/  
 @ 178 O\_SITAGE 1. /\* Age-site edit override \*/  
 @ 179 O\_SEQCON 1. /\* Sequence number-dx conf override \*/  
 @ 180 O\_SEQLAT 1. /\* Site-type-lat-seq override \*/  
 @ 181 O\_SURCON 1. /\* Surgery-diagnostic conf override \*/  
 @ 182 O\_SITTY 1. /\* Site-type edit override \*/  
 @ 183 H\_BENIGN 1. /\* Histology edit override \*/  
 @ 184 O\_RPTSRC 1. /\* Report source sequence override \*/  
 @ 185 O\_DFSITE 1. /\* Seq-ill-defined site override \*/  
 @ 186 O\_LEUKDX 1. /\* Leuk-Lymph dx confirmation override \*/  
 @ 187 O\_SITBEH 1. /\* Site-behavior override \*/  
 @ 188 O\_EODDT 1. /\* Site-EOD-dx date override \*/  
 @ 189 O\_SITEOD 1. /\* Site-laterality-EOD override \*/  
 @ 190 O\_SITMOR 1. /\* Site-laterality-morph override \*/  
 @ 191 TYPEFUP 1. /\* Type of followup expected \*/  
 @ 192 AGE\_REC 2. /\* Age recode \*/  
 @ 194 SITE\_REC 5. /\* Site recode \*/  
 @ 199 SITE2\_RE 5. /\* Site rec with Kaposis and mesothelioma \*/  
 @ 204 ICDOTO9V \$char4. /\* Recode ICD-O-2 to 9 \*/  
 @ 208 ICDOT10V \$char4. /\* Recode ICD-O-2 to 10 \*/  
 @ 212 ICCCSITE \$char3. /\* ICCC site recode \*/  
 @ 215 ICCCSM \$char3. /\* SEER modified ICCC site recode \*/  
 @ 218 ICC3 \$char3. /\* ICCC ICD-O-3 Recode \*/  
 @ 221 ICC3EXT \$char3. /\* ICCC ICD-O-3 Extended Recode \*/  
 @ 224 BEHANAL 1. /\* Behavior recode for analysis \*/  
 @ 225 ICD\_CODE 1. /\* ICD-O Coding Scheme \*/  
 @ 226 HISTREC 2. /\* Broad Histology recode \*/  
 @ 228 BRAINREC 2. /\* Brain recode \*/  
 @ 230 CS202SCH 3. /\* CS Schema v0202 \*/  
 @ 233 RAC\_RECA 1. /\* Race recode A \*/  
 @ 234 RAC\_RECY 1. /\* Race recode Y \*/  
 @ 235 NHIAREC 1. /\* Origin Recode NHIA \*/  
 @ 236 HST\_STGA 1. /\* SEER historic stage A \*/  
 @ 237 AJCC\_STG 2. /\* AJCC stage 3rd edition (1988+) \*/  
 @ 239 AJ\_3SEER 2. /\* SEER modified AJCC stage 3rd ed (1988+) \*/  
 @ 241 SSG77 1. /\* SEER Summary Stage 1977 (1995-2000) \*/  
 @ 242 SSG2000 1. /\* SEER Summary Stage 2000 (2001-2003) \*/  
 @ 243 NUMPRIMS 2. /\* Number of primaries \*/  
 @ 245 FIRSTPRM 1. /\* First malignant primary indicator \*/  
 @ 246 STCOUNTY 5. /\* State-county recode \*/  
 @ 251 SURV\_TM 4. /\* Survival time recode \*/  
 @ 255 ICD\_5DIG 5. /\* Cause of death to SEER site recode \*/  
 @ 260 CODKM 5. /\* COD to site rec KM \*/  
 @ 265 STAT\_REC 1. /\* Vital status recode (study cutoff used) \*/  
 @ 266 IHS 1. /\* IHS link \*/  
 @ 267 hss\_2000 1. /\* Historic SSG 2000 Stage \*/  
 @ 268 AYA\_rec 2. /\* AYA recode \*/  
 @ 270 Lymphrec 2. /\* Lymphoma recode \*/  
 @ 272 dth\_cl 1. /\* SEER cause of death classification \*/  
 @ 273 o\_dth\_cl 1. /\* SEER other cause of death classification \*/  
 @ 274 exteval 1. /\* CS EXT/Size Eval \*/  
 @ 275 nodeeval 1. /\* CS Nodes Eval \*/

```

@ 276 metseval 1. /* CS Mets Eval */
;
run;
/*cases from Great California, Kentucky, Louisiana and New Jersey from 2000*/
data seerbc.breast3;
infile 'F:\thesis\thesis\thesis\SEER_1973_2008_TEXTDATA\incidence\yr2000_2008.ca_ky_lo_nj\breast.txt'
lrecl=279;
input @ 1 CASENUM 8. /* Patient ID */
@ 9 REG 10. /* SEER registry */
@ 19 MAR_STAT 1. /* Marital status at diagnosis */
@ 20 RACE 2. /* Race/ethnicity */
@ 22 ORIGIN 1. /* Spanish surname or origin */
@ 23 NHIA 1. /* NHIA Derived Hisp Origin */
@ 24 SEX 1. /* Sex */
@ 25 AGE_DX 3. /* Age at diagnosis */
@ 28 YR_BRTH 4. /* Year of birth */
@ 32 PLC_BRTH 3. /* Place of birth */
@ 35 SEQ_NUM 2. /* Sequence number */
@ 37 DATE_mo 2. /* Month of diagnosis */
@ 39 DATE_yr 4. /* Year of diagnosis */
@ 43 SITEO2V $char4. /* Primary site ICD-O-2 (1973+) */
@ 47 LATERAL 1. /* Laterality */
@ 48 HISTO2V 4. /* Histologic Type ICD-O-2 */
@ 52 BEHO2V 1. /* Behavior Code ICD-O-2 */
@ 53 HISTO3V 4. /* Histologic Type ICD-O-3 */
@ 57 BEHO3V 1. /* Behavior code ICD-O-3 */
@ 58 GRADE 1. /* Grade */
@ 59 DX_CONF 1. /* Diagnostic confirmation */
@ 60 REPT_SRC 1. /* Type of reporting source */
@ 61 EOD10_SZ 3. /* EOD 10 - size (1988+) */
@ 64 EOD10_EX 2. /* EOD 10 - extension */
@ 66 EOD10_PE 2. /* EOD 10 - path extension */
@ 68 EOD10_ND 1. /* EOD 10 - lymph node */
@ 69 EOD10_PN 2. /* EOD 10 - positive lymph nodes examined */
@ 71 EOD10_NE 2. /* EOD 10 - number of lymph nodes examined */
@ 73 EOD13 $char13. /* EOD--old 13 digit */
@ 86 EOD2 $char2. /* EOD--old 2 digit */
@ 88 EOD4 $char4. /* EOD--old 4 digit */
@ 92 EODCODE $char1. /* Coding system for EOD */
@ 93 TUMOR_1V $char1. /* Tumor marker 1 */
@ 94 TUMOR_2V $char1. /* Tumor marker 2 */
@ 95 TUMOR_3V $char1. /* Tumor marker 3 */
@ 96 CS_SIZE 3. /* CS Tumor size */
@ 99 CS_EXT 3. /* CS Extension */
@ 102 CS_NODE 3. /* CS Lymph Nodes */
@ 105 CS_METS 2. /* CS Mets at DX */
@ 107 CS_SSF1 3. /* CS Site-Specific Factor 1 */
@ 110 CS_SSF2 3. /* CS Site-Specific Factor 2 */
@ 113 CS_SSF3 3. /* CS Site-Specific Factor 3 */
@ 116 CS_SSF4 3. /* CS Site-Specific Factor 4 */
@ 119 CS_SSF5 3. /* CS Site-Specific Factor 5 */
@ 122 CS_SSF6 3. /* CS Site-Specific Factor 6 */
@ 125 CS_SSF25 3. /* CS Site-Specific Factor 25 */

```

@ 128 D\_AJCC\_T 2. /\* Derived AJCC T \*/  
 @ 130 D\_AJCC\_N 2. /\* Derived AJCC N \*/  
 @ 132 D\_AJCC\_M 2. /\* Derived AJCC M \*/  
 @ 134 D\_AJCC\_S 2. /\* Derived AJCC Stage Group \*/  
 @ 136 D\_SSG77 1. /\* Derived SS1977 \*/  
 @ 137 D\_SSG00 1. /\* Derived SS2000 \*/  
 @ 138 D\_AJCC\_F 1. /\* Derived AJCC - flag \*/  
 @ 139 D\_SSG77F 1. /\* Derived SS1977 - flag \*/  
 @ 140 D\_SSG00F 1. /\* Derived SS2000 - flag \*/  
 @ 141 CSV\_ORG 6. /\* CS Version Input Original \*/  
 @ 147 CSV\_DER 6. /\* CS Version Derived \*/  
 @ 153 CSV\_CUR 6. /\* CS Version Input Current \*/  
 @ 159 SURGPRIM 2. /\* RX Summ--surg prim site \*/  
 @ 161 SCOPE 1. /\* RX Summ--scope reg LN sur \*/  
 @ 162 SURGOTH 1. /\* RX Summ--surg oth reg/dis \*/  
 @ 163 SURGNODE 2. /\* Number of lymph nodes \*/  
 @ 165 RECONST 1. /\* Reconstruction \*/  
 @ 166 NO\_SURG 1. /\* Reason no cancer-directed surgery \*/  
 @ 167 RADIATN 1. /\* Radiation \*/  
 @ 168 RAD\_BRN 1. /\* Radiation to Brain and/or CNS \*/  
 @ 169 RAD\_SURG 1. /\* Radiation sequence with surgery \*/  
 @ 170 SS\_SURG 2. /\* Site specific surgery (1983-1997) \*/  
 @ 172 SRPRIMO2 2. /\* Surgery to Primary Site \*/  
 @ 174 SCOPE02 1. /\* Scope of lymph node surgery \*/  
 @ 175 SRGOTH02 1. /\* Surgery to other sites \*/  
 @ 176 REC\_NO 2. /\* Record number \*/  
 @ 178 O\_SITAGE 1. /\* Age-site edit override \*/  
 @ 179 O\_SEQCON 1. /\* Sequence number-dx conf override \*/  
 @ 180 O\_SEQLAT 1. /\* Site-type-lat-seq override \*/  
 @ 181 O\_SURCON 1. /\* Surgery-diagnostic conf override \*/  
 @ 182 O\_SITTYP 1. /\* Site-type edit override \*/  
 @ 183 H\_BENIGN 1. /\* Histology edit override \*/  
 @ 184 O\_RPTSRC 1. /\* Report source sequence override \*/  
 @ 185 O\_DFSITE 1. /\* Seq-ill-defined site override \*/  
 @ 186 O\_LEUKDX 1. /\* Leuk-Lymph dx confirmation override \*/  
 @ 187 O\_SITBEH 1. /\* Site-behavior override \*/  
 @ 188 O\_EODDT 1. /\* Site-EOD-dx date override \*/  
 @ 189 O\_SITEOD 1. /\* Site-laterality-EOD override \*/  
 @ 190 O\_SITMOR 1. /\* Site-laterality-morph override \*/  
 @ 191 TYPEFUP 1. /\* Type of followup expected \*/  
 @ 192 AGE\_REC 2. /\* Age recode \*/  
 @ 194 SITE\_REC 5. /\* Site recode \*/  
 @ 199 SITE2\_RE 5. /\* Site rec with Kaposis and mesothelioma \*/  
 @ 204 ICDOTO9V \$char4. /\* Recode ICD-O-2 to 9 \*/  
 @ 208 ICDOT10V \$char4. /\* Recode ICD-O-2 to 10 \*/  
 @ 212 ICCCSITE \$char3. /\* ICCC site recode \*/  
 @ 215 ICCCSM \$char3. /\* SEER modified ICCC site recode \*/  
 @ 218 ICC3 \$char3. /\* ICCC ICD-O-3 Recode \*/  
 @ 221 ICC3EXT \$char3. /\* ICCC ICD-O-3 Extended Recode \*/  
 @ 224 BEHANAL 1. /\* Behavior recode for analysis \*/  
 @ 225 ICD\_CODE 1. /\* ICD-O Coding Scheme \*/  
 @ 226 HISTREC 2. /\* Broad Histology recode \*/  
 @ 228 BRAINREC 2. /\* Brain recode \*/

```

@ 230 CS202SCH 3. /* CS Schema v0202*/
@ 233 RAC_RECA 1. /* Race recode A */
@ 234 RAC_RECY 1. /* Race recode Y */
@ 235 NHIAREC 1. /* Origin Recode NHIA */
@ 236 HST_STGA 1. /* SEER historic stage A */
@ 237 AJCC_STG 2. /* AJCC stage 3rd edition (1988+) */
@ 239 AJ_3SEER 2. /* SEER modified AJCC stage 3rd ed (1988+) */
@ 241 SSG77 1. /* SEER Summary Stage 1977 (1995-2000) */
@ 242 SSG2000 1. /* SEER Summary Stage 2000 2000 (2001-2003) */
@ 243 NUMPRIMS 2. /* Number of primaries */
@ 245 FIRSTPRM 1. /* First malignant primary indicator */
@ 246 STCOUNTY 5. /* State-county recode */
@ 251 SURV_TM 4. /* Survival time recode */
@ 255 ICD_5DIG 5. /* Cause of death to SEER site recode */
@ 260 CODKM 5. /* COD to site rec KM */
@ 265 STAT_REC 1. /* Vital status recode (study cutoff used) */
@ 266 IHS 1. /* IHS link */
@ 267 hss_2000 1. /* Historic SSG 2000 Stage */
@ 268 AYA_rec 2. /* AYA recode */
@ 270 Lymphrec 2. /* Lymphoma recode */
@ 272 dth_cl 1. /* SEER cause of death classification */
@ 273 o_dth_cl 1. /* SEER other cause of death classification */
@ 274 exteval 1. /* CS EXT/Size Eval */
@ 275 nodeeval 1. /* CS Nodes Eval */
@ 276 metseval 1. /* CS Mets Eval */
;
run;
proc sort data=seerbc.breast1; by CASENUM;run;
proc sort data=seerbc.breast2; by CASENUM;run;
proc sort data=seerbc.breast3; by CASENUM;run;
data seerbc.breast;
merge seerbc.breast1 seerbc.breast2 seerbc.breast3;
by CASENUM;
run;
data seerbc.total;
merge seerbc.breast seerbc.time_month;
by CASENUM;
run;
DATA seerbc.one; set seerbc.total;
keep CASENUM REG MAR_STAT SEX AGE_DX YR_BRTH SEQ_NUM DATE_mo DATE_yr SITEO2V LATERAL HISTO3V
BEHO3V GRADE DX_CONF REPT_SRC EOD10_SZ EOD10_EX EOD10_ND EOD10_PN EOD10_NE TUMOR_1V TU-
MOR_2V SURGPRIM NO_SURG RADIATN RAD_SURG SS_SURG TYPEFUP SITE_REC BEHANAL HISTREC AGE_REC
RAC_RECY NHIAREC HST_STGA AJCC_STG AJ_3SEER SSG77 SSG2000 NUMPRIMS FIRSTPRM ICD_5DIG STAT_REC
dth_cl o_dth_cl time_mon;
if sex=1 then delete;
if BEHO3V=3;
if 1990<=DATE_yr<=2002;
if GRADE in (5,6,7,8) then delete;
if HST_STGA=0 then delete;
if ssg77=0 | ssg2000=0 then delete;
if AJCC_STG=0 then delete;
if REPT_SRC in (6,7,8) then delete;
if AGE_DX=999 then delete;

```

```

if AGE_DX<20 then delete;
if TYPEFUP in (1,3) then delete;
if rac_recy=9 then delete;
if HISTO3V>=8800 then delete;
run;
/*categories*/
data seerbc.cat;set seerbc.one;
racegroup=0;
if RAC_RECY=1 & NHIAREC=0 then racegroup=1;
if RAC_RECY=2 & NHIAREC=0 then racegroup=2;
if NHIAREC=1 then racegroup=3;
if RAC_RECY=4 & NHIAREC=0 then racegroup=4;
if RAC_RECY=3 & NHIAREC=0 then racegroup=5;
if RAC_RECY=7 & NHIAREC=0 then racegroup=6;
if RAC_RECY=9 & NHIAREC=0 then racegroup=6;
agegroup=0;
if AGE_DX<040 then agegroup=1;
if 040<=AGE_DX<050 then agegroup=2;
if 050<=AGE_DX<065 then agegroup=3;
if 065<=AGE_DX then agegroup=4;
marital=0;
if MAR_STAT=2 then marital=1;
if MAR_STAT in (3,4,5) then marital=2;
if MAR_STAT=1 then marital=3;
if MAR_STAT=9 then marital=4;
cellgrade=0;
if GRADE=1 then cellgrade=1;
if GRADE=2 then cellgrade=2;
if GRADE=3 then cellgrade=3;
if GRADE=4 then cellgrade=4;
if GRADE=9 then cellgrade=5;
tumorsize=0;
if EOD10_SZ<=020 then tumorsize=1;
if 020<EOD10_SZ<=050 then tumorsize=2;
if 990=>EOD10_SZ>050 | EOD10_SZ=998 then tumorsize=3;
if EOD10_SZ=997 then tumorsize=4;
if EOD10_SZ=999 then tumorsize=5;
extension=0;
if EOD10_EX in (05,10) then extension=1;
else if 10<=EOD10_EX<20 then extension=1;
else if 20<=EOD10_EX<70 then extension=2;
else if 70<=EOD10_EX<99 then extension=3;
else if EOD10_EX=99 then extension=4;
lymphnode=0;
if EOD10_PN=00 & EOD10_ND=0 then lymphnode=1;
if EOD10_ND in (1,2,3,4,5,6,7,8) | 01<=EOD10_PN<=89 then lymphnode=2;
if EOD10_ND in (1,2,3,4,5,6,7,8) | EOD10_PN in (90, 95, 97) then lymphnode=2;
if EOD10_ND=0 & EOD10_PN in (98,99) then lymphnode=3;
if EOD10_ND=9 & EOD10_PN in (0,98,99) then lymphnode=3;
surgery=0;
if SURGPRIM=00 | SS_SURG in (00,01,02,03,04,05,06,07) then surgery=1;
if SURGPRIM=99 | SS_SURG=09 then surgery=3;
if SURGPRIM=. & SS_SURG=98 then surgery=2;

```

```

if 10<=SURGPRIM<99 then surgery=2;
if 10<=SS_SURG<=90 then surgery=2;
radiation=0;
if RADIATN in (0,7) then radiation=1;
if RADIATN in (1,2,3,4,5,6) then radiation=2;
if RADIATN in (8,9) then radiation=3;
ERA=0;
if TUMOR_1V=1 then ERA=1;
if TUMOR_1V=2 then ERA=2;
if TUMOR_1V=3 then ERA=3;
if TUMOR_1V in (0,8,9) then ERA=4;
PRA=0;
if TUMOR_2V=1 then PRA=1;
if TUMOR_2V=2 then PRA=2;
if TUMOR_2V=3 then PRA=3;
if TUMOR_2V in (0,8,9) then PRA=4;
cause=0;
if ICD_5DIG=26000 then cause=1;
else if ICD_5DIG in (41000, 99999) then cause=4;
else if 20000<ICD_5DIG<26000 | 26000<ICD_5DIG<38000 then cause=2;
else if 38000<=ICD_5DIG<=50300 then cause=3;
else if ICD_5DIG=0 then cause=5;
run;
proc format library=format;
value dth_cl
0='alive or dead of other cause'
1='dead'
9='unknown';
value cellgrade
1='well differentiated'
2='moderately differentiated'
3='poorly differentiated'
4='undifferentiated'
5='unknown';
value agegroup
1='20-39'
2='40 to 49'
3='50 to 64'
4='65 and above';
value racegroup
1='non-hispanic white'
2='black'
3='hispanic'
4='asian'
5='native'
6='other';
value marital
1='married'
3='single'
2='ever married'
4='unknown';
value tumorsize
1='2cm or less'

```

```

2='2 to 5cm'
3='over 5cm'
4='Paget'
5='unknown';
value extension
1='localized'
2='regional'
3='distant'
4='unclear';
value lymphnode
1='negative'
2='positive'
3='unknown';
value surgery
1='no surgery'
2='surgery'
3='unknown';
value radiation
1='no radiation'
2='radiation'
3='unknown';
value ERA
1='positive'
2='negative'
3='borderline'
4='unknown';
value PRA
1='positive'
2='negative'
3='borderline'
4='unknown';
value cause
1='Breast cancer '
2='Other cancer'
3='Non_cancer death'
4='unclear'
5='alive';
run;
libname format 'D:\whw\data\seerbc\format';
options fmtsearch=(format);
/*frequency tables*/
proc freq data=seerbc.categ;
table racegroup agegroup marital cellgrade tumorsize extension lymphnode surgery radiation era pra cause
racegroup*marital agegroup*dth_cl agegroup*cause racegroup*cause racegroup*dth_cl cellgrade*extension sur-
gery*radiation/chisq;
format racegroup racegroup. agegroup agegroup. marital marital. cellgrade cellgrade. tumorsize tumorsize.
extension extension. lymphnode lymphnode. surgery surgery. radiation radiation. era era. pra pra.
cause cause. dth_cl dth_cl. ;
run;
/*mean and median age*/
proc means data=seerbc.categ mean median;
var age_dx;
title 'median age at diagnosis';

```

```

run;
/*age groups and survival status two-way table*/
proc template;
define crosstabs Base.Freq.CrossTabFreqs;
define header myheader;
text 'age/survival or death status Two-Way Table';
title 'age groups and survival status two-way table';
end;
end;
run;
ods listing close;
ods html file='body.html';
proc freq data=seerbc.categ;
tables agegroup*cause/chisq;
format agegroup agegroup. cause cause.;
run;
ods html close;
ods listing;

/*race and survival or death status*/
proc template;
define crosstabs Base.Freq.CrossTabFreqs;
define header myheader;
text 'Race/ethnicity and survival or death status two-way table';
title 'Race/ethnicity and survival or death status two-way table';
end;
end;
run;
ods listing close;
ods html file='body.html';
proc freq data=seerbc.categ;
tables racegroup*cause/chisq;
format cause cause. racegroup racegroup.;
run;
ods html close;
ods listing;

/*race and marital status*/
proc template;
define crosstabs Base.Freq.CrossTabFreqs;
define header myheader;
text 'race/marital status Two-Way Table';
title 'Race_Ethnic and Marital Status';
end;
end;
run;
ods listing close;
ods html file='body.html';
proc freq data=seerbc.categ;
tables racegroup*marital/chisq;
format racegroup racegroup. marital marital.;
run;
ods html close;

```

```

ods listing;

/*tumor cell diffrenciation grade and tumor extension*/
proc template;
define crosstabs Base.Freq.CrossTabFreqs;
define header myheader;
text 'cell diffrenciation grade and tumor extension Two-Way Table';
title 'cell diffrenciation grade and tumor extension';
end;
end;
run;
ods listing close;
ods html file='body.html';
proc freq data=seerbc.categ;
tables cellgrade*extension/chisq;
format cellgrade cellgrade. extension extension.;
run;
ods html close;
ods listing;

/*treatment*/
proc template;
define crosstabs Base.Freq.CrossTabFreqs;
define header myheader;
text 'Surgery and Radiotherapy Two-Way Table';
title 'Summary of Treatments';
end;
end;
run;
ods listing close;
ods html file='body.html';
proc freq data=seerbc.categ;
tables surgery*radiation/chisq;
format surgery surgery. radiation radiation.;
run;
ods html close;
ods listing;
/*add censoring indicator for breast cancer specific survival analysis*/
data seerbc.life;
set seerbc.categ;
censor=0;
if cause in (2,3,4,5) then censor=1;
run;
/*create dummy variable*/
data seerbc.life1;
set seerbc.life;
age20_39=0;
if agegroup=1 then age20_39=1;
age40_49=0;
if agegroup=2 then age40_49=1;
age50_64=0;
if agegroup=3 then age50_64=1;
over65=0;

```

```
if agegroup=4 then over65=1;
White=0;
if racegroup=1 then White=1;
Black=0;
if racegroup=2 then Black=1;
Hispanic=0;
if racegroup=3 then Hispanic=1;
Asian=0;
if racegroup=4 then Asian=1;
Native=0;
if racegroup=5 then Native=1;
Other=0;
if racegroup=6 then Other=1;
married=0;
if marital=1 then married=1;
ever_married=0;
if marital=2 then ever_married=1;
single=0;
if marital=3 then single=1;
mar_unknown=0;
if marital=4 then mar_unknown=1;
well=0;
if cellgrade=1 then well=1;
morderately=0;
if cellgrade=2 then morderately=1;
poorly=0;
if cellgrade=3 then poorly=1;
undiff=0;
if cellgrade=4 then undiff=1;
grade_unknown=0;
if cellgrade=5 then grade_unknown=1;
less2=0;
if tumorsize=1 then less2=1;
size2to5=0;
if tumorsize=2 then size2to5=1;
size5=0;
if tumorsize=3 then size5=1;
paget=0;
if tumorsize=4 then paget=1;
size_unknown=0;
if tumorsize=5 then size_unknown=1;
local=0;
if extension=1 then local=1;
regional=0;
if extension=2 then regional=1;
distant=0;
if extension=3 then distant=1;
ext_unknown=0;
if extension=4 then ext_unknown=1;
Innegative=0;
if lymphnode=1 then Innegative=1;
Inpositive=0;
if lymphnode=2 then Inpositive=1;
```

```

ln_unknown=0;
if lymphnode=3 then ln_unknown=1;
no_surgery=0;
if surgery=1 then no_surgery=1;
surg=0;
if surgery=2 then surg=1;
surg_unknown=0;
if surgery=3 then surg_unknown=1;
no_rad=0;
if radiation=1 then no_rad=1;
rad=0;
if radiation=2 then rad=1;
rad_unknown=0;
if radiation=3 then rad_unknown=1;
e_positive=0;
if ERA=1 then e_positive=1;
e_negative=0;
if ERA=2 then e_negative=1;
e_border=0;
if ERA=3 then e_border=1;
e_unknown=0;
if ERA=4 then e_unknown=1;
p_positive=0;
if PRA=1 then p_positive=1;
p_negative=0;
if PRA=2 then p_negative=1;
p_border=0;
if PRA=3 then p_border=1;
p_unknown=0;
if PRA=4 then p_unknown=1;
run;
/*add censoring indicator for overall survival analysis*/
data seerbc.life2;
set seerbc.life1;
censor1=0;
if cause=5 then censor1=1;
run;
/*effect of factors on survival and hazard function*/
/*age*/
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life1 alpha=0.05 ALPHAQT=0.05 method=life intervals=(0 to 150 by 3) plots=(
s(pvalue),ls, lls, h, p);
title 'effect of age on breast cancer specific survival or hazard function';
time time_mon*censor(1);
test White Black Hispanic Asian Native Other married ever_married single mar_unknown well morderately poorly
undiff grade_unknown less2 size2to5 size5 paget size_unknown local regional distant ext_unknown lnpositive
lnnegative ln_unknown no_surgery surg surg_unknown no_rad rad rad_unknown e_positive e_negative e_border
e_unknown p_positive p_negative p_border p_unknown;
strata agegroup/test= WILCOXON LOGRANK adjust=sidak diff=control('65 and above');

```

```

format agegroup agegroup.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerb.c.life2 alpha=0.05 ALPHAQT=0.05 method=life intervals=(0 to 150 by 3) plots=(
s(pvalue),ls, lls, h, p) notable;
title 'effect of age on overall survival or hazard function';
time time_mon*censor1(1);
test White Black Hispanic Asian Native Other married ever_married single mar_unknown
well moderately poorly undiff grade_unknown less2 size2to5 size5 size_unknown local regional distant
ext_unknown
Inpositive Innegative ln_unknown no_surgery surg surg_unknown no_rad rad rad_unknown e_positive e_negative
e_border
e_unknown p_positive p_negative p_border p_unknown;
strata agegroup/test= WILCOXON LOGRANK adjust=sidak diff=control('65 and above');
format agegroup agegroup.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*Race_Ethnic*/
title 'effect of Race/ethnicity on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerb.c.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) ;
time time_mon*censor(1);
test agegroup marital cellgrade tumorsize extension lymphnode surgery radiation era pra;
strata racegroup/test= WILCOXON LOGRANK ;
format racegroup racegroup.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*Marital Status*/
title 'effect of Marital Status on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;

```

```

ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=marital
outtest=marital_statitic ;
time time_mon*censor(1);
test agegroup racegroup cellgrade tumorsize extension lymphnode surgery radiation era pra;
strata marital/test= WILCOXON LOGRANK ;
format marital marital.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*cellgrade*/
title 'effect of tumor cell differentiation grade on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=cellgrade
outtest=cellgrade_statitic ;
time time_mon*censor(1);
test agegroup racegroup marital tumorsize extension lymphnode surgery radiation era pra;
strata cellgrade/test= WILCOXON LOGRANK ;
format cellgrade cellgrade.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*tumorsize*/
title 'effect of tumorsize at diagnosis on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=tumorsize
outtest=tumorsize_statitic ;
time time_mon*censor(1);
test agegroup racegroup marital cellgrade extension lymphnode surgery radiation era pra;
strata tumorsize/test= WILCOXON LOGRANK ;
format tumorsize tumorsize.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*extension*/

```

```

title 'effect of tumor extension on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq =_wilcox1
LogUniChiSq =_logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=extension
outtest=extension_statitic ;
time time_mon*censor(1);
test agegroup racegroup marital cellgrade tumorsize lymphnode surgery radiation era pra;
strata extension/test= WILCOXON LOGRANK ;
format extension extension.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*lymphnode*/
title 'effect of lymph node metastais on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq =_wilcox1
LogUniChiSq =_logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p)
outsurv=lymphnode outtest=lymphnode_statitic ;
time time_mon*censor(1);
test agegroup racegroup cellgrade marital tumorsize extension surgery radiation era pra;
strata lymphnode/test= WILCOXON LOGRANK ;
format lymphnode lymphnode.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*ERA*/
title 'effect of tumor marker ERA on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq =_wilcox1
LogUniChiSq =_logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=ERA
outtest=ERA_statitic ;
time time_mon*censor(1);
test agegroup racegroup marital cellgrade tumorsize extension lymphnode surgery radiation pra;
strata ERA/test= WILCOXON LOGRANK ;
format ERA ERA.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;

```

```

ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*PRA*/
title 'effect of tumor marker PRA on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 180 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=PRA
outtest=PRA_statitic ;
time time_mon*censor(1);
test agegroup racegroup cellgrade marital tumorsize extension lymphnode surgery radiation era;
strata PRA/test= WILCOXON LOGRANK ;
format PRA PRA.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*surgery*/
title 'effect of surgery treatment on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 150 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=surgery
outtest=surgery_statitic ;
time time_mon*censor(1);
test agegroup racegroup marital cellgrade tumorsize extension lymphnode radiation ERA PRA;
strata surgery/test= WILCOXON LOGRANK ;
format surgery surgery.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*radiation*/
title 'effect of radiotherapy on breast cancer specific survival or hazard function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 150 by 3) plots=( s(pvalue),ls, lls, h, p)
outsurv=radiotherapy outtest=radiotherapy_statitic ;
time time_mon*censor(1);
test agegroup racegroup marital cellgrade tumorsize extension lymphnode surgery era pra;
strata radiation/test= WILCOXON LOGRANK ;

```

```

format radiation radiation.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*combined treatments*/
title 'effect of treatment on breast cancer specific survival function';
ODS HTML;
ODS GRAPHICS ON;
ODS TRACE ON;
ODS OUTPUT WilUniChiSq = _wilcox1
LogUniChiSq = _logrank1;
proc lifetest data=seerbc.life method=life intervals=(0 to 150 by 3) plots=( s(pvalue),ls, lls, h, p) outsurv=treatment
outtest=treatment_statistic ;
time time_mon*censor(1);
test agegroup racegroup marital cellgrade tumorsize extension lymphnode era pra;
strata radiation surgery/test= WILCOXON LOGRANK ;
format radiation radiation. surgery surgery.;
run;
ODS OUTPUT CLOSE;
ODS TRACE OFF;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

/*Cox proportional hazard regression*/
/*Cox proportional hazard regression for bc specific survival */
ods graphics on;
proc phreg data=seerbc.life plots=(survival cumhaz);
class racegroup(ref='non-hispanic white') agegroup(ref='20-39') marital(ref='married') cellgrade(ref='well
differentiated')
tumorsize(ref='2cm or less') lymphnode(ref='negative') surgery(ref='surgery') radiation(ref='radiation') extension
(ref='localized') ERA (ref='positive') PRA (ref='positive') /param=ref;
model time_mon*censor(1)=racegroup agegroup marital cellgrade tumorsize lymphnode extension ERA PRA sur-
gery radiation grade_size age_ext age_grade race_treat surg_rad grade_ext grade_ln size_ln surg_ext rad_grade
race_mar surg_size ERA_PRA marker_surg surg_grade marker_rad surg_ln_size/rl selection=stepwise ties=efron;
age_ext=agegroup*extension;
age_grade=agegroup*cellgrade;
surg_grade=surgery*cellgrade;
surg_size=surgery*tumorsize;
race_mar=racegroup*marital;
rad_grade=radiation*cellgrade;
surg_ext=surgery*extension;
size_ln=tumorsize*lymphnode;
grade_ln=cellgrade*lymphnode;
grade_ext=cellgrade*extension;
surg_rad=surgery*radiation;
race_treat=racegroup*surgery*radiation;
grade_size=cellgrade*tumorsize;
surg_ln_size=surgery*lymphnode*tumorsize;
ERA_PRA=ERA*PRA;
marker_surg=ERA*PRA*surgery;

```

```

marker_rad=ERA*PRA*radiation;
baseline survival=_all_ CUMHAZ=_all_;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA.;
run;
ods graphics off;

ods graphics on;
proc phreg data=seerbc.life plots=(survival cumhaz);
class racegroup(ref='non-hispanic white') agegroup(ref='20-39') marital(ref='married') cellgrade(ref='well
differentiated')
tumorsize(ref='2cm or less') lymphnode(ref='negative') surgery(ref='surgery') radiation(ref='radiation') extension
(ref='localized') ERA (ref='positive') PRA (ref='positive') /param=ref;
model time_mon*censor(1)=racegroup agegroup marital surg_size cellgrade tumorsize lymphnode extension ERA
PRA surgery radiation race_treat grade_ext size_In rad_grade age_ext race_mar marker_surg marker_rad/rl
ties=efron;
age_ext=agegroup*extension;
race_mar=racegroup*marital;
rad_grade=radiation*cellgrade;
size_In=tumorsize*lymphnode;
grade_ext=cellgrade*extension;
race_treat=racegroup*surgery*radiation;
marker_surg=ERA*PRA*surgery;
marker_rad=ERA*PRA*radiation;
surg_size=surgery*tumorsize;
baseline survival=_all_ CUMHAZ=_all_;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA.;
run;
ods graphics off;

/*Cox proportional hazard regression for overall survival */
ods graphics on;
proc phreg data=seerbc.life2 plots=(survival cumhaz);
class racegroup(ref='non-hispanic white') agegroup(ref='20-39') marital(ref='married') cellgrade(ref='well
differentiated')tumorsize(ref='2cm or less') lymphnode(ref='negative') surgery(ref='surgery')
radiation(ref='radiation') extension (ref='localized') ERA (ref='positive') PRA (ref='positive') /param=ref;
model time_mon*censor1(1)=racegroup agegroup marital cellgrade tumorsize lymphnode extension ERA PRA
surgery radiation grade_size race_age surg_rad grade_ext grade_In size_In surg_ext rad_grade race_mar surg_size
ERA_PRA marker_surg surg_grade marker_rad surg_In_size/rl selection=stepwise ties=efron;
surg_grade=surgery*cellgrade;
surg_size=surgery*tumorsize;
race_mar=racegroup*marital;
rad_grade=radiation*cellgrade;
surg_ext=surgery*extension;
size_In=tumorsize*lymphnode;
grade_In=cellgrade*lymphnode;
grade_ext=cellgrade*extension;
surg_rad=surgery*radiation;
race_age=racegroup*agegroup;
grade_size=cellgrade*tumorsize;

```

```

surg_in_size=surgery*lymphnode*tumorsize;
ERA_PRA=ERA*PRA;
marker_surg=ERA*PRA*surgery;
marker_rad=ERA*PRA*radiation;
baseline survival=_all_ CUMHAZ=_all_;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA.;
run;
ods graphics off;

ods graphics on;
proc phreg data=seerbc.life2 plots=(survival cumhaz);
class racegroup(ref='non-hispanic white') agegroup(ref='20-39') marital(ref='married') cellgrade(ref='well
differentiated')
tumorsize(ref='2cm or less') lymphnode(ref='negative') surgery(ref='surgery') radiation(ref='radiation') extension
(ref='localized') ERA (ref='positive') PRA (ref='positive') /param=ref;
model time_mon*censor1(1)=racegroup agegroup marital cellgrade tumorsize lymphnode extension ERA PRA
surgery radiation race_age grade_ext grade_in size_in rad_grade race_mar surg_size ERA_PRA marker_surg
surg_grade marker_rad surg_in_size/rl ties=efron;
surg_grade=surgery*cellgrade;
surg_size=surgery*tumorsize;
race_mar=racegroup*marital;
rad_grade=radiation*cellgrade;
size_in=tumorsize*lymphnode;
grade_in=cellgrade*lymphnode;
grade_ext=cellgrade*extension;
race_age=racegroup*agegroup;
surg_in_size=surgery*lymphnode*tumorsize;
ERA_PRA=ERA*PRA;
marker_surg=ERA*PRA*surgery;
marker_rad=ERA*PRA*radiation;
baseline survival=_all_ CUMHAZ=_all_;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA.;
run;
ods graphics off;

/*censoring indicator for 5-year relative survival*/
data seerbc.logit;
set seerbc.life2;
five=0;
if time_mon>=60 then five=1;
surv=1;
if five=0 & cause in (1,2,3,4) then surv=0;
grade_size=cellgrade*tumorsize;
race_treat=radiation*surgery*race;
age_ext=agegroup*extension;
surg_rad=surgery*radiation;
grade_ext=cellgrade*extension;
grade_in=cellgrade*lymphnode;
size_in=tumorsize* lymphnode;

```

```

surg_ext =surgery*extension;
rad_grade=radiation*grade;
race_mar=race*marital;
surg_size=surgery*tumorsize;
ERA_PRA=ERA*PRA;
marker_surg=surgery*ERA*PRA;
surg_grade=surgery*cellgrade;
marker_rad=radiation*ERA*PRA;
surg_In_size=surgery*lymphnode*tumorsize;
run;
proc format library=format;
value five
0='less than 5 years'
1='5 years and above';
value surv
0='death'
1='5 year survival';
run;

/*full logistic model*/
proc logistic data=seerbc.logit;
class racegroup(ref='non-hispanic white') agegroup(ref='20-39') marital(ref='married') cellgrade(ref='well
differentiated') tumorsize(ref='2cm or less ') lymphnode(ref='negative') extension (ref='localized') sur-
gery(ref='surgery') radiation(ref='radiation') ERA(ref='positive') PRA(ref='positive')/param=ref ;
model fiveyr_surv(event='survival or censor')=racegroup agegroup marital cellgrade tumorsize lymphnode exten-
sion ERA PRA surgery radiation grade_size race_treat age_ext surg_rad grade_ext grade_In size_In surg_ext
rad_grade race_mar surg_size ERA_PRA marker_surg surg_grade marker_rad surg_In_size /selection=stepwise
risklimits;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA. survival survival.
bcscensor bcscensor. fiveyr fiveyr. fiveyr_surv fiveyr_surv.;
run;

proc logistic data=seerbc.logit;
class racegroup(ref='non-hispanic white') agegroup(ref='20-39') marital(ref='married') cellgrade(ref='well
differentiated')tumorsize(ref='2cm or less ') lymphnode(ref='negative') extension (ref='localized') sur-
gery(ref='surgery') radiation(ref='radiation')ERA(ref='positive') PRA(ref='positive')/param=ref ;
model fiveyr_surv(event='survival or censor')=racegroup agegroup marital cellgrade tumorsize lymphnode exten-
sion ERA PRA surgery radiation grade_size race_treat age_ext grade_ext surg_ext rad_grade surg_grade
race_mar marker_surg surg_grade marker_rad/risklimits;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA. survival survival.
bcscensor bcscensor. fiveyr fiveyr. fiveyr_surv fiveyr_surv.;
run;

/*dummy variable model selection*/
data one;set seerbc.logit;
/*interactions*/
/*age*ext*/
Local_20_39=local*age20_39; local_40_49=local*age40_49; local_50_64=local*age50_64; local_65=local*over65;

```

```

reg_20_39=regional*age20_39;reg_40_49=regional*age40_49;reg_50_64=regional*age50_64;reg_65=regional*over65;
dist_20_39=distant*age20_39;dist_40_49=distant*age40_49;dist_50_64=distant*age50_64;dist_65=distant*over65;
unext_20_39=ext_unknown*age20_39;unext_40_49=ext_unknown*age40_49;unext_50_64=ext_unknown*age50_64;unext_65=ext_unknown*over65;
/*cellgrade*extension*/
well_local=well*local;mod_local=moderately*local;poor_local=poorly*local;undiff_local=undiff*local;unknown_local=grade_unknown*local;
well_region=well*regional;mod_region=moderately*regional;poor_region=poorly*regional;undiff_region=undiff*regional;unknown_region=grade_unknown*regional;
well_dist=well*distant;mod_dist=moderately*distant;poor_dist=poorly*distant;undiff_dist=undiff*distant;unknown_dist=grade_unknown*distant;
well_unknown=well*ext_unknown;mod_unknown=moderately*ext_unknown;poor_unknown=poorly*ext_unknown;undiff_unknown=undiff*ext_unknown;grad_ext_unknown=grade_unknown*ext_unknown;
/*size*lymphnode*/
less2_Inpo=less2*Inpositive;size2to5_Inpo=size2to5*Inpositive;size5_Inpo=size5*Inpositive;paget_Inpo=paget*Inpositive;unknownsz_Inpo=size_unknown*Inpositive;
less2_Inne=less2*Innegative;size2to5_Inne=size2to5*Innegative;size5_Inne=size5*Innegative;paget_Inne=paget*Innegative;unknownsz_Inne=size_unknown*Innegative;
less2_Inno=less2*In_unknown;size2to5_Inno=size2to5*In_unknown;size5_Inno=size5*In_unknown;paget_Inno=paget*In_unknown;unknownsz_Inno=size_unknown*In_unknown;
/*cellgrade*size*/
well_less2=well*less2;mod_less2=moderately*less2;poor_less2=poorly*less2;undiff_less2=undiff*less2;unknown_less2=grade_unknown*less2;
well_sz2_5=well*size2to5;mod_sz2_5=moderately*size2to5;poor_sz2_5=poorly*size2to5;undiff_sz2_5=undiff*size2to5;unknown_sz2_5=grade_unknown*size2to5;
well_sz5=well*size5;mod_sz5=moderately*size5;poor_sz5=poorly*size5;undiff_sz5=undiff*size5;unknown_sz5=grade_unknown*size5;
well_unsz=well*size_unknown;mod_unsz=moderately*size_unknown;poor_unsz=poorly*size_unknown;undiff_unsz=undiff*size_unknown;unknown_unsz=grade_unknown*size_unknown;
/*cellgrade*lymphnode*/
well_Inpo=well*Inpositive;mod_Inpo=moderately*Inpositive;poor_Inpo=poorly*Inpositive;undiff_Inpo=undiff*Inpositive;unknown_Inpo=grade_unknown*Inpositive;
well_Inno=well*In_unknown;mod_Inno=moderately*In_unknown;poor_Inno=poorly*In_unknown;undiff_Inno=undiff*In_unknown;unknown_Inno=grade_unknown*In_unknown;
well_Inne=well*Innegative;mod_Inne=moderately*Innegative;poor_Inne=poorly*Innegative;undiff_Inne=undiff*Innegative;unknown_Inne=grade_unknown*Innegative;
/*surgery*radiotherapy*/
sur_rad=surg*rad;nosur_rad=no_surgery*rad;unsur_rad=surg_unknown*rad;
sur_norad=surg*no_rad;nosur_norad=no_surgery*no_rad;unsur_norad=surg_unknown*no_rad;
sur_unrad=surg*rad_unknown;nosur_unrad=no_surgery*rad_unknown;unsur_unrad=surg_unknown*rad_unknown;
/*cellgrade*radiation*/
well_rad=well*rad;modrad=moderately*rad;poorrad=poorly*rad;undiffrad=undiff*rad;unknownrad=grade_unknown*rad;
well_norad=well*no_rad;mod_norad=moderately*no_rad;poor_norad=poorly*no_rad;undiff_norad=undiff*no_rad;unknown_norad=grade_unknown*no_rad;
well_unrad=well*rad_unknown;mod_unrad=moderately*rad_unknown;poor_unrad=poorly*rad_unknown;undiff_unrad=undiff*rad_unknown;unknown_unrad=grade_unknown*rad_unknown;
/*tumorsize*surgery*/
less2_surg=less2*surg;size2to5_surg=size2to5*surg;size5_surg=size5*surg;paget_surg=paget*surg;unknownsz_surg=size_unknown*surg;
less2_nosurg=less2*no_surgery;size2to5_nosurg=size2to5*no_surgery;size5_nosurg=size5*no_surgery;paget_nosurg=paget*no_surgery;unknownsz_nosurg=size_unknown*no_surgery;

```

```

less2_unsurg=less2*surg_unknown;size2to5_unsurg=size2to5*surg_unknown;size5_unsurg=size5*surg_unknown;
paget_unsurg=paget*surg_unknown;unknownsz_unsurg=size_unknown*surg_unknown;
/*surgery*extension*/
sur_local=surg*local;nosur_local=no_surgery*local;unsur_local=surg_unknown*local;
sur_region=surg*regional;nosur_region=no_surgery*regional;unsur_region=surg_unknown*regional;
sur_dist=surg*distant;nosur_dist=no_surgery*distant;unsur_dist=surg_unknown*distant;
sur_unkn=surg*ext_unknown;nosur_unkn=no_surgery*ext_unknown;unsur_unkn=surg_unknown*ext_unknown;
/*radiation*extension*/
lo-
cal_rad=local*rad;region_rad=regional*rad;dist_rad=distant*rad;unext_rad=ext_unknown*rad;local_norad=local*
no_rad;region_norad=regional*no_rad;dist_norad=distant*no_rad;unext_norad=ext_unknown*no_rad;local_unra
d=local*rad_unknown;region_unrad=regional*rad_unknown;dist_unrad=distant*rad_unknown;unext_unrad=ext_
unknown*rad_unknown;
/*ERA*PRA*/
ep_pp=e_positive*p_positive;ep_pn=e_positive*p_negative;ep_pb=e_positive*p_border;ep_pu=e_positive*p_un
known;
en_pp=e_negative*p_positive;en_pn=e_negative*p_negative;en_pb=e_negative*p_border;en_pu=e_negative*p_
unknown;
eb_pp=e_border*p_positive;eb_pn=e_border*p_negative;eb_pb=e_border*p_border;eb_pu=e_border*p_unkno
wn;
eu_pp=e_unknown*p_positive;eu_pn=e_unknown*p_negative;eu_pb=e_unknown*p_border;eu_pu=e_unknown*
p_unknown;
run;
/*dummy var model */
proc logistic data=one;
model fiveyr_surv(event='survival or censor')=age20_39 age40_49 age50_64 over65 White Black Hispanic Asian
Native Other married ever_married single mar_unknown well moderately poorly undiff grade_unknown less2
size2to5 size5 paget size_unknown local regional distant ext_unknown lnegative lnpositive ln_unknown
no_surgery surg surg_unknown no_rad rad rad_unknown e_positive e_negative e_border e_unknown p_positive
p_negative p_border p_unknown local_20_39 local_40_49 local_50_64 local_65 reg_20_39 reg_40_49 reg_50_64
reg_65 dist_20_39 dist_40_49 dist_50_64 dist_65 unext_20_39 unext_40_49 unext_50_64 unext_65 well_local
mod_local poor_local undiff_local unknown_local well_region mod_region poor_region undiff_region un-
known_region well_dist mod_dist poor_dist undiff_dist unknown_dist well_unknown mod_unknown
poor_unknown undiff_unknown grad_ext_unknown less2_lnp size2to5_lnp size5_lnp paget_lnp
unknownsz_lnp less2_lne size2to5_lne size5_lne paget_lne unknownsz_lne less2_lno size2to5_lno
size5_lno paget_lno unknownsz_lno well_less2 mod_less2 poor_less2 undiff_less2 unknown_less2 well_sz2_5
mod_sz2_5 poor_sz2_5 undiff_sz2_5 unknown_sz2_5 well_sz5 mod_sz5 poor_sz5 undiff_sz5 unknown_sz5
well_unsz mod_unsz poor_unsz undiff_unsz unknown_unsz well_lnp mod_lnp poor_lnp undiff_lnp un-
known_lnp well_lno mod_lno poor_lno undiff_lno unknown_lno well_lne mod_lne poor_lne
undiff_lne unknown_lne sur_rad nosur_rad unsur_rad sur_norad nosur_norad unsur_norad sur_unrad
nosur_unrad unsur_unrad well_rad modrad poorrad undiff_rad unknownrad well_norad mod_norad poor_norad
undiff_norad unknown_norad well_unrad mod_unrad poor_unrad undiff_unrad unknown_unrad less2_surg
size2to5_surg size5_surg paget_surg unknownsz_surg less2_nosurg size2to5_nosurg size5_nosurg paget_nosurg
unknownsz_nosurg less2_unsurg size2to5_unsurg size5_unsurg paget_unsurg unknownsz_unsurg sur_local
nosur_local unsur_local sur_region nosur_region unsur_region sur_dist nosur_dist unsur_dist sur_unkn
nosur_unkn unsur_unkn local_rad region_rad dist_rad unext_rad local_norad region_norad dist_norad
unext_norad local_unrad region_unrad dist_unrad unext_unrad ep_pp ep_pn ep_pb ep_pu en_pp en_pn en_pb
en_pu eb_pp eb_pn eb_pb eb_pu eu_pp eu_pn eu_pb eu_pu/risklimits SELECTION=stepwise ;
format
racegroup racegroup. agegroup agegroup. Marital marital. cellgrade cellgrade. tumorsize tumorsize. lymphnode
lymphnode. surgery surgery. radiation radiation. extension extension. ERA ERA. PRA PRA. survival survival.
fiveyr_surv fiveyr_surv.;
run;

```