STATISTICAL ANALYSIS OF WAITING TIME FOR BREAST CANCER SURGERY

By

Huimin Lu

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

STATISTICAL ANALYSIS OF WAITING TIME FOR BREAST CANCER SURGERY

BY

Huimin Lu

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

MASTER OF SCIENCE

Contents

Abstract iv

Acknowledgements v

List of Figures vi

List of Outputs vii

List of Tables viii

Chapter 1 Introduction 1

1.1 Medical Background 1

1.2 Objectives. 3

1.3 Available Datasets. 4

1.4 Confidentiality. 5

1.5 Statistical Methods and Analysis. 6

1.6 Structure of the Practicum. 7

Chapter 2 Basics of Survival Analysis 8

2.1 Basic Concepts and Functions in Survival Analysis 8

2.2 Estimation of the Survivor Function. 9

2.3 Estimation of the Median Survival Time. 10

2.4 Detection of the Difference on Survival Curves. 11

Chapter 3 The Cox Regression Model 14

3.1 Proportional Hazards Model. 14

3.2 Partial Likelihood Estimations. 16

3.2.1 Estimation Procedures without Ties. 17

3.2.2 Estimation Procedures with Ties. 18

3.3 Interpretation of SAS Outputs. 19

3.4 Model Selection. 21

3.5 Estimation of the Survivor Function. 22

3.6 Goodness of Fit Assessment of the PH Assumption. 24

Chapter 4 Data Preparation 27

4.1 Data Cleaning. 27

4.2 Definition of Variables. 29

Chapter 5 Results of the Waiting Time Analysis 37

5.1 Frequency Tables for Variables. 37

5.2 The Overall Pattern 43

5.3 Detection of the Difference on Waiting Times 50

Chapter 6 Results of the Survival Time Analysis 59

6.1 The Survival Curve. 59

6.2 Detection of the Difference on Survival Curves. 62

6.3 Identification of Significant Covariates. 68

6.4 Model Selection. 71

6.5 Interpretation of the Cox Regression Analysis. 73

6.6 Estimation of the Survivor Function. 77

6.7 Goodness of Fit Assessment of the PH Assumption. 79

Chapter 7 Summary and Future Studies 83

Appendices 85

A.1 Variables List 85

A.2 ICD9 (ICD10) Diagnosis Codes for Breast Cancer. 87

A.3 ICD9 (ICD10) Surgery Treatment Procedure Codes. 88

A.4 SAS Codes of Data Cleaning. 90

A.5 SAS Codes of Defining Variables. 104

A.6 SAS Codes of the Waiting Time Analysis. 105

A.7 SAS Codes of the Survival Time Analysis. 106

Bibliography. 112

Abstract

Breast cancer is the most common cancer diagnosed among Canadian women. Even though cancer care services have been improving, people still continue complaining about waiting times to access surgery. People are always curious about how long they can survive after diagnosed with breast cancer.

In this practicum, we use non-parametric (the Kaplan-Meier) method and semi-parametric (Cox regression model) method to do waiting time (e.g., from diagnosis to first surgery) and survival time (e.g., from diagnosis to death or emigration) analysis respectively. The data are from Cancer Registry at CancerCare Manitoba and the Manitoba Health's Population Registration File at Manitoba Health.

For the waiting time analysis, we investigate the waiting time curve and test the difference on waiting times by diagnosis age group, cancer stage, region, and between urban and rural. For the survival time analysis, we test the difference on survival by cancer stage, region, and between urban and rural, income within urban and within rural. We then identify significant covariates (e.g., waiting times, diagnosis age and cancer stage) that affect survival times. Finally, we select the best statistical model by incorporating significant covariates into the model. Results from this practicum indicate that waiting times are significant different by cancer stage and region. The survival times are significant different by diagnosis age, cancer stage and income within urban.

Acknowledgements

I would like to acknowledge the guidance of Dr. Xikui Wang, my academic advisor, from the University of Manitoba throughout this research. His valuable comments and suggestions have helped me to successfully complete this research study.

I also thank Dr. Smiley W. Cheng of department of statistics, and Dr. Jeffrey S. Pai of Warren Centre for Actuarial Studies and Research for reading my practicum.

I really appreciate CancerCare Manitoba and Manitoba Health for providing me with an opportunity of accessing the datasets. I could not complete this research without the datasets.

I am very thankful to my manager, Dr. Donna Tunner; my supervisor, Dr. Alain Demers, and my co-workers, Katherine Fradette and Zoann Nugent, from Epidemiology and Cancer Registry department at CancerCare Manitoba for sharing with me their knowledgeable experience and wonderful ideas, which are invaluable for the completion of this research.

The Manitoba Graduate Scholarships (MGS) also helped me to complete this study.

Finally, many thanks to my parents, my boyfriend, my younger sister and brother's support and love; they are always with me wherever I go in my life.

List of Figures

4.1 Map of RHAs in Manitoba. 30

5.1 Kaplan-Meier Estimates of the Waiting Time Functions 49

5.2 Waiting Time Curves for Age Group. 53

5.3 Waiting Time Curves for Cancer Stage. 54

5.4 Waiting Time Curves for Region. 56

5.5 Waiting Time Curves for Urban and Rural. 57

6.1 Kaplan-Meier Estimates of the Survivor Function. 61

6.2 Plot of Log-Survivor Versus Survival Times. 62

6.3 Survival Curves for Cancer Stage. 63

6.4 Survival Curves for Region. 64

6.5 Survival Curves for Urban and Rural. 65

6.6 Survival Curves for Income within Urban. 66

6.7: Survival Curves for Income within Rural. 67

6.8: Plot of Log-cumulative Hazard Versus Log-surv for Cancer Stage. 79

6.9: Plot of Weighted Schoenfeld Residuals Versus Survival Times. 82

List of Outputs

5.1 Kaplan-Meier Estimates of Waiting Times. 49

5.2 Test of Difference in Waiting Time for Age Group. 52

5.3 Test of Difference in Waiting Time Cancer Stage. 54

5.4 Test of Difference in Waiting Time for Region. 55

5.5 Test of Difference in Waiting Time for Urban and Rural. 57

6.1 Test of Difference in Waiting Time for Wait_zero. 60

6.2 Summary Statistics for Survival Times. 61

6.3 Test of Difference in Survival Time for Cancer Stage. 63

6.4 Test of Difference in Survival Time for Region. 64

6.5 Test of Difference in Survival Time for Urban and Rural. 65

6.6 Test of Difference in Survival Time for Income within Urban. 66

6.7: Test of Difference in Survival Time for Income within Rural. 67

6.8: Test of the Null Hypothesis: $\beta_1=0$ for Waiting Times. 70

6.9: Test of the Null Hypothesis: $\beta_1=\beta_2=\beta_3$ for Diagnosis Age. 71

6.10: Values of *–2LogL* for Different Models Fitted into the Data. 72

6.11: Results of the Cox Regression Analysis. 74

6.12: Comparison of the HR between Cancer Stages. 76

6.13: Portion of Estimation of Survivor Functions at Sample Means. 78

List of Tables

4.1 Stage and Corresponding TNM Summary Table. 34

5.1 Frequency Table for Surgery Type. 38

5.2 Frequency Table for Censor. 39

5.3 Frequency Table for Waiting Times in Weeks. 39

5.4 Frequency Table for Age Group. 40

5.5 Frequency Table for Cancer Stage. 41

5.6 Frequency Table for Income. 42

5.7 Frequency Table for Region. 43

5.8 Waiting Time Data (36 out of 2101) 46

6.1 Frequency Table for Wait_Zero. 60

Chapter 1

Introduction

1.1 Medical Background

In Canada, the most frequently diagnosed cancer among Canadian women is breast cancer, which is the second leading cause of cancer death behind lung cancer. There is an upward trend for incidence rates among women of age over 50, while (in the past decade) mortality rates are starting to decline.

In Manitoba, about 800 women are newly diagnosed with breast cancer and about 200 patients among newly and previously diagnosed with breast cancer die each year. Even though the incidence rate is really high compared with the other provinces, the mortality rate has been stabilized due to better screening programs and improved treatments.

In 1996, the province established a screening program called Screening Mammography. It is a most effective way to detect breast cancer at its early stage. A woman without symptoms of breast cancer between the age of 50 and 69 should have a screening mammography once every two years. In Manitoba, the screening program is called Manitoba Breast Screening Program (MBSP) that includes two procedures: a breast X-ray (mammogram) and information session on breast health. The earlier a cancer is detected, the more likely treatments are successful, and a higher chance for

the breast cancer patients to survive.

The treatment options for breast cancer patients include surgery, radiation therapy, chemotherapy and hormonal therapy. Among these treatment options, surgery is the most common procedure and can be carried out easily in local hospitals. The types of surgical procedures for breast cancer include breast-conserving surgery (BCS), mastectomy, axillary lymph node dissection (ALND), sentinel lymph node biopsy (SLNB), and breast reconstruction. Details for each procedure are as follows:

1. Segmental (lumpectomy) mastectomy, which is usually followed by radiation therapy, is a breast-conserving surgery and removes the lump and up-to one-quarter of the breast tissue.
2. Mastectomy includes simple, modified radical and radical mastectomy. Simple mastectomy removes the entire breast tissue excluding lymph nodes under the arm and muscles under the breast; modified radical mastectomy removes the entire breast tissue including some lymph nodes and small muscles; and radical mastectomy removes the entire breast tissue, lymph nodes and all muscles.
3. Axillary lymph node dissection removes all axillary lymph nodes.
4. Sentinel lymph node biopsy removes one to three sentinel nodes to test for cancer. If the nodes contain cancer, then all axillary nodes are removed by the axillary lymph node dissection procedure.

5. Breast reconstruction reconstructs a breast that has been removed by mastectomy procedure. The procedure is done at the time of mastectomy or after.

Which surgical procedure should be taken is determined by the stage of tumour and the patient's preference.

1.2 Objectives

The most intensive procedure of breast cancer is surgery. Even though treatments have been improved and the screening program has been established, people still complained about the waiting time to access surgery and wondered whether their survival times are affected by some factors (e.g., diagnosis age, cancer stage). The following lists the detail of our objectives.

1. What is the overall pattern of waiting time from diagnosis to first surgery and whether waiting times are different among cancer stage, diagnosis age group, region, and between urban and rural?
2. What is the overall survival curve and whether the survival times are affected by waiting times, cancer stage, diagnosis age, region, between urban and rural, and income within urban and within rural?
3. Select the optimal statistical model that best describes the survival times with significant covariates incorporated into the model.

Hopefully, our study will provide some helpful information to improve the healthcare system.

1.3 Available Datasets

In this study, there are two datasets available. They are the Manitoba Cancer Registry database housed at CancerCare Manitoba (CCMB), and the Manitoba Health's Population Registration File (MHPR) from Manitoba Health.

The Manitoba Cancer Registry is a population-based cancer registry that contains all cancer cases in Manitoba. It collects patient demographics (i.e., patients' name, sex, birth date and region of residence at diagnosis), tumour characteristics (i.e., tumour type and cancer stage at diagnosis), vital status (alive or deceased) and some treatment information. The cohort is all Manitoba women diagnosed with invasive (ICD-9 174 and ICD-10 C50) or in situ (ICD-9 233.0 and ICD-10 D05) breast cancer from 1995 to 2003 (approximately 7000 women). ICD-9 and ICD-10 are the 9th and 10th version of International Classification of Diseases (ICD) coding system developed by the World Health Organization (WHO). ICD-9 had been used to classify diseases, health conditions and procedures up to December 31, 1999, and was replaced by ICD-10 on January 1, 2000. Appendices A.2 lists ICD9 and ICD10 diagnosis codes for female breast cancer. Some women do not have any treatment information and we also cannot tell whether they are still waiting for treatments or they are not qualified for surgeries because of their health conditions. Therefore, we only select women who (about 6000) have a surgery (or surgeries). In most cases,

women have more than one tumour. In order to simplify the analysis, the algorithm is used to select one tumour per woman. See section 4.1 for details.

The MHPR provides patient's last cancellation date and reasons (death or emigration from the province) of terminating coverage.

Finally, the Manitoba Health data is linked to the Cancer Registry data by scrambled unique personal identifiers.

1.4 Confidentiality

As an employee of CCMB, I have signed Personal Health Information Act (PHIA) compliance certificates in order to keep all health data confidential. This study has been approved by Research Resource Impact Committee (RRIC) at CCMB, Health Research Ethics Board (HREB) of the Faculty of Medicine at the University of Manitoba and Health Information Privacy Committee (HIPC) at Manitoba Health. Therefore, all data related to this study are kept confidentially.

The cohort is extracted from Cancer Registry data at CCMB, and scrambled unique personal identifiers are created before sending a request to Manitoba Health to get patients' information on termination of coverage. The Manitoba Health data is extracted at Manitoba Health according to the cohort extracted from the Manitoba Cancer Registry data, and then sent back to the Department of Epidemiology and Cancer Registry at CCMB with scrambled personal identifiers on a password protected disc. The analysis is then performed at the Department of Epidemiology and

Cancer Registry at CCMB. During the analysis, scrambled personal identifiers are used to track the patients. All the other personal identifiers, such as patients' name, address and personal health identification number (PHIN), are removed from the final database used for the analyses. The results summarized in this practicum are only based on groups of patients in this cohort.

1.5 Statistical Methods and Analysis

This is a retrospective study. The cohort consists of Manitoba women who were diagnosed with breast cancer from 1995 to 2003 and had a surgery (or surgeries). There are two parts of analysis - the waiting time analysis (from diagnosis to first surgery) and the survival time analysis (from diagnosis to death or emigration).

1. For the waiting time analysis, the Kaplan-Meier method is used to understand the overall pattern for waiting times to first surgery. The log-rank test (also known as the Mantel-Haenszel test) is used to determine whether waiting times are the same by cancer stage, diagnosis age group, region, and between urban and rural, at the 5% significant level (α). If the $P-value$ is less than 5% for any covariates, we conclude that waiting times are different by hat covariate; otherwise, there are no difference on waiting times.
2. For the survival time analysis, the Kaplan-Meier method is used to understand the overall survival curve and also the survival curves by

cancer stage, region, and between urban and rural, income within urban and within rural. The differences of these covariates on survival are tested by the log-rank test at the 5% significant level (α). Finally, the Cox regression model is used to find out how continuous variables (i.e., waiting times to first surgery, diagnosis age and cancer stage) affect survival times.

The analyses are performed with SAS 9.1. The Kaplan-Meier method can be invoked by specifying METHOD=KM in PROC LIFETEST statement and Cox regression model can be produced by PROC PHREG procedure.

1.6 Structure of the Practicum

In chapter 2, we briefly review the theoretical background regarding the Kaplan-Meier method. In chapter 3, we review the Cox regression model, and show how the numerical computation can be implemented in the statistical SAS software. In chapter 4, we introduce how the data are cleaned and variables are defined. In chapter 5, we interpret the results for the waiting time analysis. In chapter 6, we interpret the results for the survival time analysis. In chapter 7, we summarize our studies and discuss future studies for Breast Cancer. In Appendices, SAS codes are listed.

Chapter 2

Basics of Survival Analysis

2.1 Basic Concepts and Functions in Survival Analysis

In clinical studies, survival time can be defined as the time from diagnosis to an event, such as a surgery or death.

Let T denote the survival time that is any non-negative random variable. The survivor function $S(t)$ is the probability of an individual surviving longer than $t > 0$ and is given by

$$S(t) = P(T > t) \qquad (2.1.1)$$

$$= 1 - F(t).$$

Moreover, $S(0) = 1$ and $S(\infty) = 0$. The value of $S(t)$ decreases with increasing survival time. $F(t)$ is the cumulative distribution function and is the probability of an individual surviving less than or equal to t.

The hazard function $h_0(t)$ is used to express the risk of death at time t and is the probability of an individual experiencing an event in a small interval, $(t, t + \Delta t)$, conditional on having survived to time t:

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \right\}. \qquad (2.1.2)$$

We can rewrite the numerator of 2.1.2 as

$$\frac{P(t \le T < t + \Delta t)}{P(T \ge t)}$$

$$= \frac{F(t + \Delta t) - F(t)}{S(t)}.$$

Plugging back to 2.1.2, we have

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)}$$

$$= \frac{f(t)}{S(t)}$$

$$= -\frac{d}{dt} \{\ln(S(t)\}.$$

Then taking integral for both sides, we have

$$S(t) = \exp\{- H(t)\}. \qquad (2.1.3)$$

Equation (2.1.3) indicates that the survivor function is equal to the exponential of the negative of the cumulative hazards function.

2.2 Estimation of the Survivor Function

In clinical and epidemiological studies, the Kaplan-Meier method (also known as the product-limit (PL) estimator) is the most popular method for the preliminary survival analysis of data. This method can produce the overall estimated survivor function, the estimated survivor functions by different groups (e.g., age), and the median (or percentile) survival time estimator(s). It can also test the fit of some parametric regression models (e.g., exponential).

Suppose that there are n individuals with k distinct observed survival

times in the study. At each observed survival time, there is at least one event occurred and the events occur independently. The observed distinct survival times are ordered as $t_{(1)} < t_{(2)} \ldots < t_{(k)}$, where $k \leq n$. Let n_j, $j = 1,2,\ldots,k$, be the number of individuals who have not experienced the events right before $t_{(j)}$ and individuals who are censored at $t_{(j)}$ are included. Let d_j be the number of individuals who have experienced an event at $t_{(j)}$. Then, the Kaplan-Meier estimate of the survivor function is

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j}\right), j = 1,2,\ldots,k, t \in [t_{(j)}, t_{(j+1)}). \qquad (2.2.1)$$

For $0 \leq t < t_{(1)}$, $\hat{S}(t) = 1$ since $t_{(1)}$ is the first observed survival time; for $t_{(1)} \leq t < t_{(k)}$, the estimate of the survivor function is expressed as equation (2.1.1); for $t \geq t_{(k)}$, if the survival time is censored, then $\hat{S}(t)$ is undefined. Otherwise $\hat{S}(t) = 0$.

$\hat{S}(t)$ does not change between the consecutive observed survival times and it decreases with the observed survival times increasing, therefore, the plot of $\hat{S}(t)$ is a step function and is plotted using PLOTS=(S) in SAS.

2.3 Estimation of the Median Survival Time

After the survivor function has been estimated, it is easy to estimate the median survival time.

In clinical and epidemiological studies, the median is used to summarize the data. The median survival time $t(50)$ is defined as the time at which 50% of individuals have not experienced an event. The formula is given as

$$S\{t(50)\} = 0.5\,. \tag{2.3.1}$$

Because the estimate of the survivor function is a step function, the estimate of the median event time $\hat{t}(50)$ is defined to be the smallest observed survival time such that its corresponding estimated survivor function is less than 0.5. It is equivalent to the following equation

$$\hat{t}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) < 0.5\}, j = 1,2,...,k. \tag{2.3.2}$$

if $\hat{S}(t_{(j)}) = 0.5$ does not exist, or is equal to

$$\hat{t}(50) = \frac{(t_{(j)} + t_{(j+1)})}{2} \tag{2.3.3}$$

if $\hat{S}(t_{(j)}) = 0.5$ exists.

2.4 Detection of the Difference on Survival Curves

The question "whether the survivor functions are the same by different groups (e.g., age)" can be answered by the following hypothesis test procedure.

First of all, let's test whether the survivor functions are the same in 2 groups. The hypotheses are:

$$H_0 : S_1(t) = S_2(t)$$

$$H_a : S_1(t) \neq S_2(t)$$

The statistic is the log-rank test (also known as the Mantel-Haenszel test) and can be obtained using a STRATA statement in PROC LIFETEST procedure. Before introducing the log-rank test formula, we assume that there are k independent and distinct observed survival times in the combined group, which are ordered as

$t_{(1)} < t_{(2)} ... < t_{(k)}$. Let d_{1j} and d_{2j}, $j = 1, 2, ..., k$, be the number of individuals having experienced an event from group 1 and 2 at $t_{(j)}$ respectively. Let n_{1j} and n_{2j} be the number of individuals at risk at $t_{(j)}$ from group 1 and 2 respectively. The total number of events occurring at $t_{(j)}$ is $d_j = d_{1j} + d_{2j}$ and the total number of individuals at risk at $t_{(j)}$ is $n_j = n_{1j} + n_{2j}$.

The log-rank statistic for each group is the sum of the difference between the observed number of events and the corresponding expected number of events at all observed survival times, and is expressed as

$$U_L = \sum_{j=1}^{k} (d_{ij} - e_{ij}), \tag{2.4.1}$$

where $i = 1,2$ and $e_{ij} = {n_{1j} d_j}/{n_j}$.

The log-rank test statistic (Mantel and Haenszel (1959)) is

$$\frac{U^2_L}{V_L} \sim \chi_1^2, \tag{2.4.2}$$

where $V_L = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n^2_j (n_j - 1)}$.

The log-rank statistic follows a chi-square distribution with one degree of freedom when H_0 is true.

We can extend the log-rank test statistic to test the difference on survival in g groups. Then the test statistic would follow a chi-square distribution with $g-1$ degrees of freedom when H_0 is true. The test statistic is

$$U_L' V_L^{-1} U_L, \tag{2.4.3}$$

where U_L is a vector with g elements and U_L' is its transpose. Each element of

U_L can be expressed as

$$\sum_{j=1}^{k}(d_{gj}-e_{gj}); \tag{2.4.4}$$

Let V_L be a g by g *variance-covariance matrix*, with element

$$V_{Lmr}=\sum_{j=1}^{k}\frac{n_{mj}d_j(n_j-d_j)}{n_j(n_j-1)}(\delta_{mr}-\frac{n_{rj}}{n_j}), \tag{2.4.5}$$

where $m,r=1,2,...,g$. When $m=r$, we have $\delta_{mr}=1$ and V_{Lmr} is the variance and the values are along the diagonal in the *variance-covariance matrix*. When $m\neq r$, we have $\delta_{mr}=0$ and V_{Lmr} is the covariance and the values are off the diagonal.

We draw our conclusion based on the $P-value$ in SAS output. If the $P-value<5\%$, there is strong evidence to reject the null hypothesis and conclude that different groups have different survivor times. Otherwise, there is not enough evidence to reject the null hypothesis and the conclusion is that there seems to have no difference on survival times by groups.

Chapter 3

The Cox Regression Model

In medical and epidemiological studies, the form of the distribution of the survival time usually is unknown. Therefore, parametric methods have to be replaced in order to identify significant prognostic factors. In this chapter, we review the Cox regression model and its related statistical inference.

Cox first proposed the Cox regression model, also known as the proportional hazards model in 1972. Since then, it has become the widely used model for model fitting and for identifying significant prognostic factors in medical and epidemiological studies. Since the model does not require any specific form of the survival time distribution, it is also known as the semi-parametric regression model.

3. 1 Proportional Hazards Model

The general form of the proportional hazards model introduced by Cox (1972) can be written as

$$h(t \mid X) = h_0(t)\exp(\beta' X), \tag{3.1.1}$$

where t is the event time, $X = (x_1, x_2, ..., x_p)'$ is a $p \times 1$ column vector of covariates whose values are recorded at the time of origin, $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ is a $p \times 1$ column vector of regression coefficients, $h_0(t)$ is called the baseline hazard function when all

covariates of the hazard function have values of zero.

From (3.1.1), the hazard function for the i^{th} individual can be expressed as

$$h_i(t) = h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}). \quad (3.1.2)$$

This model is named as a proportional hazards model because the ratio of the hazard function (or the hazard ratio) for any two individuals does not change with survival time t. In order to understand why the Cox regression model is also named as the proportional hazards model, suppose there are two individuals i and j, and their hazard functions are expressed as (3.1.2). Then the hazard ratio becomes

$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{1i} - x_{1j}) + \beta_2(x_{2i} - x_{2j}) + ... + \beta_p(x_{pi} - x_{pj})\}. \quad (3.1.3)$$

It is obvious that (3.1.3) is independent of time, which means any two hazard functions graphed in the same plot should be parallel to each other throughout the study time.

Subsequently, from (3.1.2)

$$\frac{h_i(t)}{h_0(t)} = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}). \quad (3.1.4)$$

and

$$\ln\frac{h_i(t)}{h_o(t)} = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} = \sum_{k=1}^{p} \beta_k x_{ki}. \quad (3.1.5)$$

Equation (3.1.5) is a linear function of the products of covariates and their corresponding coefficients.

Among many available statistical programs, PROC PHREG in SAS is the most powerful one for handling ties in observed data, and tied data are often observed in medical studies.

3. 2 Partial Likelihood Estimations

The maximum partial likelihood method proposed by Cox (1972) can be used to estimate the coefficients $\beta_1, \beta_2, ..., \beta_p$.

The likelihood function for the proportional hazards model of (3.1.2) can be defined as two terms:

The first term includes both $h_0(t)$, the baseline hazard function, and β, the vector of coefficients.

The second term only includes β.

Only the first term is considered as the ordinary likelihood function, while the second term is ignored. Even though there is a missing term in the partial likelihood function, the estimates obtained by the partial likelihood still have the two important properties as the usual maximum likelihood estimates. The sampling distribution of the parameter estimator is approximately normal and the estimators are unbiased. This is the first property. The other one is that it considers the ranks of the survival times instead of the numerical values during the estimation procedure.

Cox (1972) proposed two different partial likelihood functions. The first one is based on the survival times observed in a continuous scale without ties and the second one is in a situation where the survival times are observed at discrete times with ties. Later, Breslow (1974) and Efron (1977) modified Cox's partial likelihood function when the observed survival times are observed continuously with ties. So far, the above estimation methods can be applied when the survival times are measured on

either a continuous scale or a discrete scale. Different estimation procedures of survival times with or without ties are reviewed in the following two sections, respectively.

3. 2. 1 Estimation Procedures without Ties

Suppose that there are k distinct observed survival times that are recorded from n observed individuals, and there are n-k right-censored observations. At each observed survival time, there is only one event occurred. Therefore, the k distinct observed survival times can be ordered as $t_{(1)} < t_{(2)} < ... < t_{(k)}$ and their corresponding covariates at time $t_{(j)}$ are $x_{(1)}, x_{(2)}, ..., x_{(k)}$. Let $R(t_{(j)})$, $j = 1,2,...,k$, be the set of individuals who are at risk at time $t_{(j)}$, which means that these individuals have not experienced an event at time $t_{(j)}$.

Cox (1972) defined the partial likelihood function based on (3.1.2) in the following form

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta' x_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\beta' x_l)}, j = 1, 2, ..., k, \qquad (3.2.1)$$

or equivalently

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right]^{\delta_i} i = 1,2,...n, \qquad (3.2.2)$$

where δ_i is an event indicator whose values is equal to zero if t_i is right-censored and one if t_i is uncensored. Then the corresponding log-partial likelihood function is

$$\ln L(\beta) = \sum_{i=1}^{n} \delta_i \{\beta' x_i - \ln \sum_{l \in R(t_i)} \exp(\beta' x_l)\}, \qquad (3.2.3)$$

The maximum partial likelihood estimators, $\hat{\beta}$, in the proportional hazards model is a solution to the following simultaneous equations by applying the Newton-Raphson iterative procedure (Collett (1994)):

$$\frac{\partial(L(\beta))}{\partial \beta} = 0. \qquad (3.2.4)$$

3. 2. 2 Estimation Procedures with Ties

Suppose that there are k distinct observed survival times from n observed individuals in the study. Let $R(t_{(j)})$, $j = 1, 2, ..., k$, be the risk set at $t_{(j)}$ and $m_{(j)}$ be the number of events occurring at time $t_{(j)}$. Then there are $m_{(j)}!$ possible ways to order their survival times when the observed survival times are at a discrete scale. The partial likelihood of each possibility can be written as (3.2.1), hence, the sum, Z_j, is the union of the partial likelihood of all possibilities with p covariates at time $t_{(j)}$. Finally, the partial likelihood with ties is the product of each Z_j.

Cox (1972) proposed the partial likelihood function with ties, when the observed event times are at a discrete scale, as the following formula

$$L_d(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta' z_j)}{\sum_{l \in R(t_{(j)}; m_{(j)})} \exp(\beta' z_l)}. \qquad (3.2.5)$$

The partial likelihood function with ties when the observed event times are at a continuous scale, defined by Breslow (1974), is

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta' z_j)}{\left[\sum_{l\in R(t_{(j)})}\exp(\beta' x_l)\right]^{m_{(j)}}}, \qquad (3.2.6)$$

where Z_j is a $p\times 1$ vector and each element is expressed as the sum of h^{th}, $h=1,2,...p$, covariates for all individuals who experienced an event at time $t_{(j)}$.

Another approximate partial likelihood function with ties when the observed survival times at a continuous scale is given by Efron (1977)

$$L(\beta)=\prod_{j=1}^{k}\frac{\exp(\beta' z_j)}{\prod_{d=1}^{m_{(j)}}\left[\sum_{l\in R(t_{(j)})}\exp(\beta' x_l)-(d-1)m_{(i)}^{-1}\sum_{l\in M^{*}_{(j)}}\exp(\beta' x_l)\right]} \qquad (3.2.7)$$

where $d=1, 2, ..., m_{(j)}$ and $M_{(j)}^{*}$ is the set of all individuals who experienced the event at time $t_{(j)}$.

The computations are difficult by hand. The approximation is accomplished by specifying TIES=BRESLOW or EFRON or EXACT after MODEL step in the PROC PHREG procedure in SAS. In this study, we choose TIES=BRESLOW to handle ties.

3. 3 Interpretation of SAS Outputs

Once the observed data are fitted into the proportional hazards model by SAS, the SAS output provides the parameter estimates with the following additional useful information: the standard error, Wald chi-square test, the $P-value$, and hazard ratio and its confidence interval for any parameter, β_j, where $j=1, 2, ..., p$.

The standard error is used to obtain the $100(1-\alpha)\%$ confidence interval for

any parameter, β_j,

$$(\hat{\beta}_j - z_{\alpha/2} se(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2} se(\hat{\beta}_j)), \tag{3.3.1}$$

where $Z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ percentile of the standard normal distribution. If zero falls within the confidence interval, then β_j should not be included in the model; otherwise, β_j should be included in the model. In order to get the confidence interval for parameters in the SAS output, RISKLIMITS option must be specified under the MODEL step.

The Wald chi-square test is to test the hull hypothesis that $\beta_j = 0$, that is, whether the corresponding covariate, x_j, has an effect on the hazard or not. The Wald statistic follows a Chi-square distribution with one degree of freedom and its value can be obtained by the equation

$$\left[\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right]^2 \tag{3.3.2}$$

The $P-value$ is interpreted for testing $\beta_j = 0$ that is whether a covariate has an effect on the PH model when all other covariates $x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_p$, with their corresponding coefficients, $\beta_1, \beta_2, ..., \beta_{j-1}, \beta_{j+1}, ..., \cdots \beta_p$ are already in the model. The null hypothesis will not be rejected if the $P-value$ is greater than $\alpha = 5\%$, so that covariate x_j should not be kept in the model in the presence of other covariates $x_1, x_2 ..., x_{j-1}, x_{j+1} ..., x_p$.

Consequently, the hazard ratio can be obtained by plugging the parameter estimators back into (3.1.4). Furthermore, if we take the exponential of the lower boundary, $\hat{\beta}_{jL}$, and of the upper boundary, $\hat{\beta}_{jU}$, of $\hat{\beta}_j$ obtained in (3.3.1), a $100(1-\alpha)\%$ confidence interval for the hazard ratio is obtained,

$$(\exp(\hat{\beta}_{jL}), \exp(\hat{\beta}_{jU})). \tag{3.3.3}$$

We get the confidence interval for hazard ratio in the SAS output by specifying the RISKLIMITS option under the MODEL step.

3. 4 Model Selection

In this section, we briefly review how to choose the most appropriate model by the likelihood-ratio method for nested models. If the first model contains a subset covariates of the second model, it is said to be a nested model. For example, there are p covariates fitted in Model (1) and there are $p+q$ covariates (including these p covariates in Model (1)) fitted in Model (2), then Model (1) is called a nested model within Model (2). We can evaluate the fit of Model (1) by the likelihood-ratio approach

$$-2\ln\hat{L} = -2\ln\hat{L}(1) - (-2\ln\hat{L}(2)) = -2\ln\{\frac{\hat{L}(1)}{\hat{L}(2)})\}, \tag{3.4.1}$$

where $\hat{L}(1)$ and $\hat{L}(2)$ are the maximized partial likelihood function of Model (1) and Model (2) respectively. This is a log-partial likelihood statistic for testing the null hypothesis that $\beta_{p+1}, \beta_{p+2}, ..., \beta_{p+q}$ are all zero. If the statistic is large enough, we would reject the null hypothesis and conclude that the extra q covariates are significant in the model and Model (2) is preferred. Otherwise, Model (1) is preferred.

3.5 Estimation of the Survivor Function

Once the most appropriate model is identified and parameters are estimated, the survivor function could be estimated. The survivor function at time t with p covariates in the proportional hazards model is

$$s(t) = [s_0(t)]^{\exp(\sum_{j=1}^{p} \beta_j x_j)} \tag{3.5.1}$$

where $s(t)$ is the probability of an individual whose survival time is longer than t and $s_0(t)$ is the baseline survivor function at time t with all covariates equal to zero. In order to estimate the survivor function, we have to estimate the baseline survivor function first. Breslow (1974) and Kalbfleisch and Prentice (1980) proposed two different approaches.

Suppose that there are k distinct observed survival times from n observed individuals and they can be ordered as $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$. Let $R(t_{(j)}), j = 1,2,...,k$, denote the risk set and $m_{(j)}$ the number of uncensored observations at time $t_{(j)}$. By assuming that the baseline hazard function is constant between the consecutive failure times, Breslow (1974) proposed the estimated cumulative baseline hazard function as

$$\hat{H}_0(t) = \sum_{t_{(j)} \le t} \frac{m_{(j)}}{\sum_{l \in R(t_{(j)})} \exp(\beta' x_l)} \tag{3.5.2}$$

Consequently, the baseline survivor function is estimated as

$$\hat{s}_0(t) = \exp[-\hat{H}_0(t)] = \prod_{t_{(j)} \le t} \{\exp[\frac{m_{(j)}}{\sum_{l \in R(t_{(j)})} \exp(\beta' x_l)}]\}. \tag{3.5.3}$$

By substituting (3.5.3) and the estimators of β parameters back into (3.5.1), the survivor function is estimated.

The other estimated baseline survivor function proposed by Kalbfleisch and Prentice (1980) is

$$\hat{s}_0(t) = \prod_{j=0}^{k} \hat{\xi}_j \,, t_{(j)} < t \le t_{(j+1)}, \quad j = 1,2,...,k, \tag{3.5.4}$$

where $\hat{\xi}_j$ can be obtained by the following equation when no tied data are observed:

$$\hat{\xi}_j = [1 - \frac{\exp(\hat{\beta}' x_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)}]^{\exp(-\hat{\beta}' x_{(j)})}. \tag{3.5.5}$$

When tied data are observed, the estimated baseline survivor function (3.5.4) is considered as a step function, and $\hat{\xi}_j$ is the estimated survival probability for an individual from time $t_{(j)}$ to $t_{(j+1)}$ that can be obtained by solving the following k equations simultaneously,

$$\sum_{l \in M_{(j)}} \cdot \frac{\exp(\hat{\beta}' x_l)}{1 - \hat{\xi}_j^{\exp(\hat{\beta}' x_l)}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l), \quad j = 1,2,...,k, \tag{3.5.6}$$

where $M_{(j)}^{*}$ is the set of individuals who fail at time $t_{(j)}$.

The above numerical calculation is very complicated, but can be accomplished by the BASELINE statement in PROC PHREG. SAS takes the observed average of every covariate, $\bar{x}_1, \bar{x}_2, ..., \bar{x}_p$, to interpret the estimated survivor function. Therefore, the estimated survivor function for the i^{th} individual now becomes

$$\hat{s}_i(t, \bar{x}) = [\hat{s}_0(t, \bar{x})]^{\exp(\hat{\beta}' x_i)}. \tag{3.5.7}$$

3.6 Goodness of Fit Assessment of the PH Assumption

After fitting the proportional hazards (PH) model to the observed data, the adequacy of the model needs to be validated. In this section, we review several methods for checking the adequacy of the PH assumption.

At the beginning of this section, we explained that the reason that the Cox regression model is also known as proportional hazards model is that the covariates are independent of time. The first method is to check the proportional hazards assumption by introducing time-dependent covariates in the model. Therefore, we can incorporate the interaction between covariates and time into the model and the interaction terms are the product of the i^{th} covariate, x_i, and time t. Next, we can test the significance of the coefficients for the interaction terms by the Wald test introduced in Section 3.3. The proportional hazards assumption is violated if the coefficients of the interaction terms are significant.

The second method is to stratify the survival data according to a covariate with m levels, and then apply (3.5.8) to estimate the survivor functions under each stratum. Finally, plot $\log(-\log(\hat{s}_j(t;\bar{x}_j)))$, $j = 1,2,...,m$, versus t. If the assumption is adequate, the m curves should be parallel. Otherwise, the assumption is violated.

The third method is based on the residuals. There are three types of residuals: modified Cox-Snell residuals, Schoenfeld residuals and deviance residuals. Now, suppose there are n individuals in the study, then the residuals formula at time t_i for the i^{th} individual and covariate x_i, $i = 1,2,...,n$, can be expressed as follows.

1. The Modified Cox-Snell residuals is given by

$$r_i = \hat{H}_i(t_i; x_i) = -\ln \hat{s}(t_i; x_i), \tag{3.6.1}$$

where $\hat{H}_i(t_i)$ and $\hat{s}_i(t_i)$ are the estimated cumulative hazard and survivor functions at the uncensored time t_i.

2. The Schoenfeld residuals (Schornfeld, 1982) is given by

$$r_{ji} = \delta_i \left[x_{ji} - \frac{\sum_{l \in R(t_{(i)})} x_{jl} \exp(\hat{\beta}' x_l)}{\sum_{l \in R(t_{(i)})} \exp(\hat{\beta}' x_l)} \right], \tag{3.6.2}$$

where $j = 1,2,...,p$, x_{ji} is the value of the j^{th} covariate for the i^{th} individual, $R(t_{(i)})$ is the group of individuals at risk and δ_i takes the value of zero if $t_{(i)}$ is right-censored and one if $t_{(i)}$ is uncensored. There is a more effective method based on the weighted Schoenfeld residuals, which was proposed by Grambsch and Therneau (1994) as

$$r_{ji}^{*} = r \operatorname{var}(\hat{\beta}) r_{ji} \tag{3.6.3}$$

where r is the number of events, $r_{ji} = (r_{1i}, r_{2i}, ..., r_{pi})$ is the vector of Schoenfeld residuals for the i^{th} individual and $\operatorname{var}(\hat{\beta})$ is the estimated covariance matrix of $\hat{\beta}$.

3. The deviance residuals proposed by Therneau *et al.* (1990) is given by

$$r_{Di} = sign(r_{Mi}) \sqrt{-2[r_{Mi} + \delta_i \ln(\delta_i - r_{Mi})]}, \tag{3.6.4}$$

where $r_{Mi} = \delta_i - r_i$ is the martingale residual proposed by Fleming and Harrington (1991) for the i^{th} individual and δ_i is equal to 1 if $t_{(i)}$ is uncensored and 0 otherwise. *Sing*() is set to $+1$ if its argument is positive, 0 if it is zero and -1 if it is negative.

If the proportional hazards assumption is adequate, the plot of Cox-Snell

residuals versus its Kaplan-Meier estimated survivor function $(\hat{s}(t))$ should be on a 45^0 straight line, the plot of the weighted Schoenfeld residuals versus a covariate and the plot of deviance residuals versus the survivor time should be symmetrically distributed about zero and should not show any particular pattern.

Chapter 4

Data Preparation

4.1 Data Cleaning

The Manitoba Cancer Registry dataset is updated monthly. Variables ICD9 and ICD10 diagnosis codes (See Appendices A.2), sex, diagnosis year and postal code at diagnosis from the December 2006 dataset are used to define the cohort (7321 women), and then to update death date by linking with the August 2007 dataset. 145 women who have the same death date as the diagnosis date are deleted from the dataset. The reason is that it is reasonable to infer that these women did not receive any treatment since they were diagnosed with breast cancer upon death.

In the dataset, most of the women have multiple tumours. In order to simplify our analyses, we use the same algorithm defined by the Epidemiology and Cancer Registry department at CCMB to select one tumour per woman. Details of the algorithm are listed as follows:

Step 1. If the diagnosis dates of tumours are more than six months after the first diagnosis date, then the later tumours are deleted from the data.

Step 2. Check the data with the pathological summary stage. If both pathological summary stages are known and different, then tumours with the lower stage are deleted from the data.

Step 3. Check the data with the pathological nodal status. If both pathological nodal statuses are known and different, then tumours with lower pathological nodal status are deleted from the data.

Step 4. Check the data with the number of positive nodes. Tumours with lower or missing number of positive nodes are deleted from the data.

Step 5. Check the data with the pathological tumour stage. If the pathological tumour stages are known and different, tumours with the lower pathological tumour stage are deleted from the data.

Step 6. Check the data with the size of tumour. Tumours with smaller size are deleted from the data.

Step 7. Randomly select one tumour left from the above steps.

In order to keep patients' information confidential, the scrambled unique personal identifier is created for each patient before sending a request to Manitoba Health. We also delete women who do not have a Manitoba Health Personal Identification Number (MHPIN) because Manitoba Health can not have information for these women if they do not have a MHPIN. After the dataset that contains patients' cancellation dates and reasons of termination of coverage were sent back to Epidemiology and Cancer Registry department at CCMB from Manitoba Health, we merge the cohort data from Manitoba Cancer Registry with the Manitoba Health data by scrambled unique personal identifier.

Next, we merge the treatment file with the data created from the previous step and attach the treatment procedure codes and dates. It is common that women

have more than one treatment procedures throughout the study period (1995-2003). Based on our objectives, we first select all procedures related to surgeries (See Appendices A.3), and then select the first surgery in order to calculate the waiting times from diagnosis to the first surgery. Finally, there are 6820 women in the dataset. (See Appendices A.4 for SAS codes.)

4.2 Definition of Variables

Diagnosis age, postal code at diagnosis and cancer stage are the variables from Manitoba Cancer Registry dataset and are used in both the waiting time and survival time analyses. (See Appendices 8.5 for SAS codes.) The following gives details about how we define variables:

1. Diagnosis age is a categorical variable with three groups: "0-49", "50-69" and "70+" for the waiting time analysis, and is a continuous variable for the survival time analysis.
2. Postal code at diagnosis is used to assign a Regional Health Authority (RHA) to each patient. The RHAs are specified names of geographic areas set up by the province. The responsibilities of RHAs are providing delivery and administration of health services. The Manitoba RHAs include: Winnipeg, Brandon, South Eastman, Assiniboine, Central, Parkland, North Eastman, Interlake, Burntwood, Norman and Churchill. It is a category variable for both analyses. Figure 4.1 is a map of RHAs

in Manitoba. We define a new categorical variable called 'region' with four groups according to the location of RHAs. They are East (Central, Interlake, North Eastman and South Eastman), North (Burntwood, Churchill, and Norman), West (Assiniboine, Brandon, Parkland) and Winnipeg.

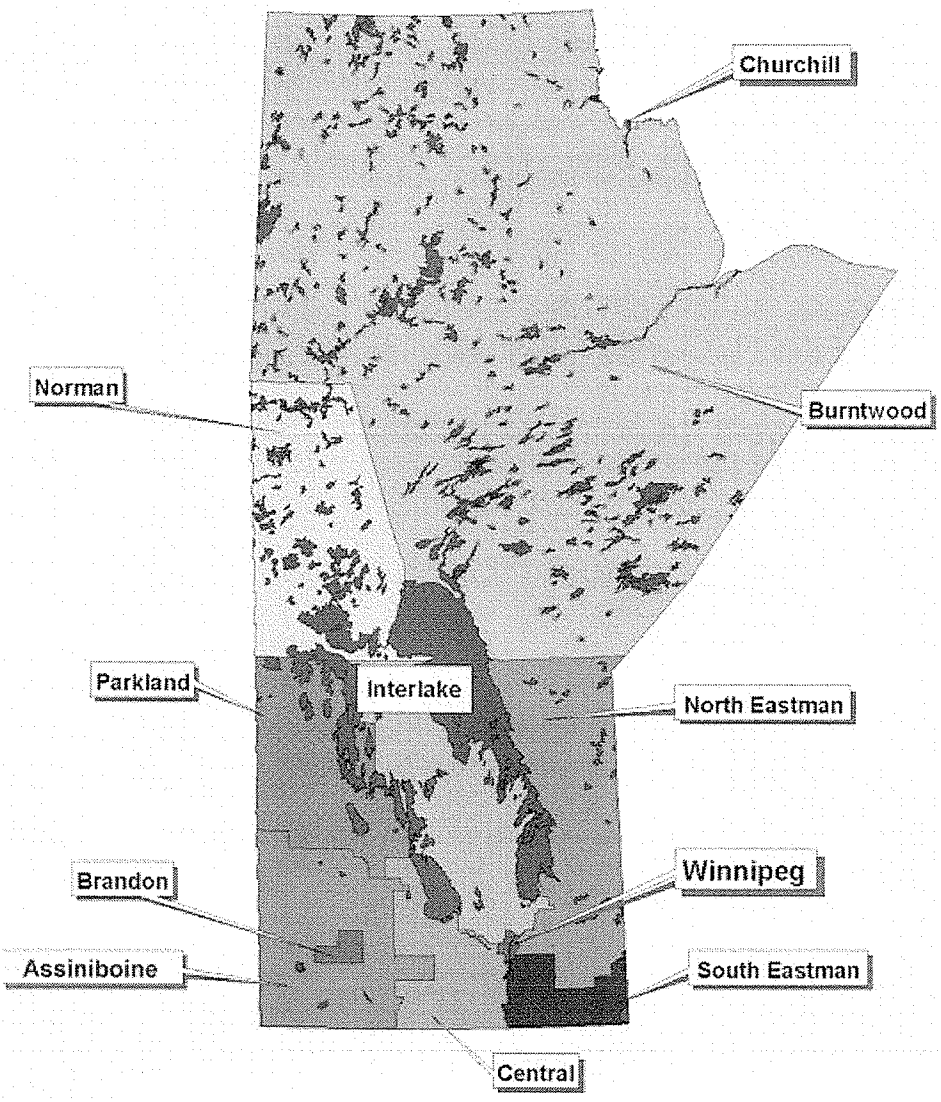


Figure 4.1: Map of RHAs in Manitoba

(Source: http://www.umanitoba.ca/centres/mchp/concept/concept.frame.shtml)

3. Postal code at diagnosis is also used to assign income level called Income Quintile to each patient. Income Quintile ranks the income from the poorest to the wealthiest based on average household income of residents by Statistics Canada Census data. It is ordered as: u_1, u_2, u_3, u_4, u_5 for urban populations and r_1, r_2, r_3, r_4, r_5 for rural populations, where subscript number *1* represents the poorest and *5* the wealthiest. For details, please read the Income Quintile available at Manitoba Centre for Health Policy (MCHP) website. A variable 'urban' taking a value of one for women with urban income quintile and a value of zero for women with rural income quintile is defined. We use 'urban' as a categorical variable for both analyses.

4. Cancer stage is manually determined by a certified tumour registrar according to pathology reports and patients' charts. Identifying the stage helps physicians to make decision about which treatment to take for breast cancer patients. The process of determining stage is called staging and staging describes the extent of a cancer at diagnosis according to the TNM classification system, where T stands for tumour, N for node and M for metastasis. The tumour size and whether the cancer has spread to lymph nodes and other parts of the body determine the stage of the cancer. Therefore, each patient is assigned with one of the following T, N and M categories before determining the stage:

a) Tumour (T):

- TX: Tumour cannot be assessed.
- T0: No evidence of tumour.
- Tis: Carcinoma in situ or lobular carcinoma in situ or Paget disease
- T1: Tumour is 2 cm or less.
- T2: Tumour is from 2 cm to 5 cm.
- T3: Tumour is greater than 5 cm.
- T4: Any size tumour spread to the chest wall or skin.

b) Node (N):

- NX: Nodes cannot be assessed.
- N0: Lymph nodes are cancer-free.
- N1: Cancer has spread to axillary lymph nodes on the same side with breast cancer.
- N2: Cancer has spread to ipsilateral (same side of body as breast cancer) lymph nodes fixed to one another or to other structures under the arm.
- N3: Cancer has spread to the ipsilateral mammary lymph nodes or the ipsilateral (same side of body as breast cancer) supraclavicular lymph nodes

c) Metastasis (M):

- MX: Metastasis cannot be assessed.
- M0: Cancer is not found in other parts of the body.

- M1: Cancer is found in other parts of the body.

Staging is the combination of each of T, N and M categories, which identifies the size and location of the cancer in a patient's body. Table 4.1 lists the TNM classifications within each stage of breast cancer. Stage is a categorical variable with values: stage 0, stage 1, stage 2, stage 3, and stage 4 for the waiting time analysis. For the survival time analysis, four dummy variables stage1, stage2, stage3 and stage4 are defined and each one has a value of one if it is in that stage; otherwise it is set to zero.

Stage	T Value	N Value	M Value
Stage 0	Tis	N0	M0
Stage I	T1	N0	M0
Stage IIA	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIIA	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Stage IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

Table 4.1: Stage and Corresponding TNM Summary Table

5. For the waiting time analysis, waiting times to the first surgery is a continuous variable measured in weeks. December 31, 2003 is chosen as the censoring date because of the rules for selecting the cohort. If the

surgery date exists and is before the censoring date, the waiting time is calculated as the difference of the first surgery date and diagnosis date, and the variable 'censor' is set to one. Otherwise, the waiting times equal to the difference of the censor date and diagnosis date, and 'censor' is set to zero. There are 4719 women with the waiting time equal to zero because these women had a biopsy removing all the lumps, and then the pathologist decides whether the tumour is positive or negative. If the tumour is positive, a surgery will be performed on the same date. In order to analyze the overall waiting time pattern, we delete these women from the data for the waiting time analyses, which left us with 2101 women.

6. For the survival time analysis, the waiting time is a continuous variable measured in weeks and is the difference between the first surgery date and the diagnosis date. The survival time is a continuous variable in years. June 30, 2007 is chosen as the censoring date because it is the latest surgery date for the cohort selected. If the death date is before the censoring date, the survival time is calculated as the difference of the death date and the diagnosis date, and the variable 'censor' is set to one. If the death date is after the censoring date, the survival time is calculated as the difference of the censoring date and the diagnosis date, and the variable 'censor' is set to zero. If there is no death date but the date of termination of coverage is found, and the reason of termination is

equal to 'death' and its date is before the censoring date, the survival time is expressed as the difference of the cancellation date and the diagnosis date, and 'censor' takes a value of one. If there is no death date, and the reason of termination is not equal to 'death' and its date is before the censoring date, the survival time is expressed as the difference of the cancellation date and the diagnosis date, and 'censor' takes a value of zero. Otherwise, the survival time is equal to the difference between the censoring date and the diagnosis date, and 'censor' takes a value of zero.

Chapter 5

Results of the Waiting Time Analysis

The results of the waiting time analysis are presented in this chapter. All analyses are performed with SAS 9.1. (See Appendices A.6 for SAS codes.)

5.1 Frequency Tables for Variables

In this section, we list frequency tables for all variables needed for the waiting time analysis. The purpose is to show how the data are distributed in each categorical variable defined in section 4.2.

Table 5.1 shows the distribution of surgery options experienced by breast cancer women in the study. It shows that axillary node dissection, segmental mastectomy and mastectomy are the most common surgical procedures being involved, and reconstruction is rarely being performed.

Axillary node dissection	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	500	23.80	500	23.80
1	1601	76.20	2101	100.00

Mastectomy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	934	44.46	934	44.46
1	1167	55.54	2101	100.00

Segmental mastectomy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1004	47.79	1004	47.79
1	1097	52.21	2101	100.00

Sentinel lymph node biopsy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1965	93.53	1965	93.53
1	136	6.47	2101	100.00

Reconstruction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2099	99.90	2099	99.90
1	2	0.10	2101	100.00

Table 5.1: Frequency Table for Surgery Type

Table 5.2 shows censoring information. There are 2040 women having a surgery by the censor date (December 31, 2003) and 61 women are still waiting for a surgery.

Censor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0(No)	61	2.90	61	2.90
1 (Yes)	2040	97.10	2101	100.00

Table 5.2: Frequency Table for Censor

Table 5.3 shows how waiting times to the first surgery are distributed in five categories. 99.52% of women waited no longer than four weeks to receive their first surgery, which implies that the cancer services are really good in Manitoba.

Waiting Time	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< 2 weeks	1936	92.15	1936	92.15
2 - 4 weeks	112	5.33	2048	97.48
4 - 8 weeks	39	1.86	2087	99.33
8 -12 weeks	4	0.19	2091	99.52
>12 weeks	10	0.48	2101	100.00

Table 5.3: Frequency Table for Waiting Times in Weeks

Table 5.4 shows the distribution of women among age group. 50% of women were between 50 and 69 years old when they were diagnosed with breast cancer. This provides evidence to support that the Screening Mammography mentioned in section 1.1 is a good program to detect breast cancer.

Age Group	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-49	491	23.37	491	23.37
50-69	1059	50.41	1550	73.77
70+	551	26.23	2101	100.00

Table 5.4: Frequency Table for Age Group

Table 5.5 shows the distribution of women by cancer stage. 71% of women were diagnosed in early stage of breast cancer, including stage 0, stage 1 and stage 2 because of the screening program; 7% of women were diagnosed in later stage 3; only 2% of women were diagnosed in the advanced stage 4.

Stage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	426	20.28	426	20.28
Stage 0	203	9.66	629	29.94
Stage 1	588	27.99	1217	57.92
Stage 2	693	32.98	1910	90.91
Stage 3	143	6.81	2053	97.72
Stage 4	48	2.28	2101	100.00

Table 5.5: Frequency Table for Cancer Stage

Table 5.6 shows the distribution of women by income. There were 60% women from urban, which is the sum of U_1 to U_5.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	15	0.71	15	0.71
R1	153	7.28	168	8.00
R2	165	7.85	333	15.85
R3	203	9.66	536	25.51
R4	158	7.52	694	33.03
R5	152	7.23	846	40.27
U1	234	11.14	1080	51.40
U2	229	10.90	1309	62.30
U3	279	13.28	1588	75.58
U4	218	10.38	1806	85.96
U5	295	14.04	2101	100.00

Table 5.6: Frequency Table for Income

Table 5.7 shows the distribution of women by region. There were 54% of women diagnosed in Winnipeg and only 3% women diagnosed in the North.

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	1	0.05	1	0.05
East	485	23.08	486	23.13
North	64	3.05	550	26.18
West	408	19.42	958	45.60
Winnipeg	1143	54.40	2101	100.00

Table 5.7: Frequency Table for Region

5.2 The Overall Pattern

There were 199 distinct observed waiting times in the dataset. They were ordered as $t_{(1)} < t_{(2)} < t_{(3)} < .... < t_{(199)}$. Table 5.8 lists the first 36 out of 2101 data. Output 5.1 from the MTHOD=KM in PROC LIFETEST statement is the partial SAS output that displays the estimate of the Kaplan-Meier survivor function at each observed survival time. Figure 5.1 from the PLOT=(S) in PROC LIFETEST statement is the plot of KM Waiting Time function.

The following example gives details about how to obtain the same result with Output 5.1 for the estimate of survivor function at the observed survival time $t = 0.429$. The steps are to apply equation 2.1.1 and the data in Table 5.8. Since

$t = 0.429$ is the third observed survival time, $t_{(3)} = 0.429$. From Table 5.8, we can get all values for d_j and n_j. $d_1 = 20$, $d_2 = 4$ and $d_3 = 8$ are the number of women having a surgery at $t_{(j)}$ with $censor = 1$. $n_1 = 2101$, $n_2 = 2081$ and $n_3 = 2065$ are the number of women who have not received a surgery before $t_{(j)}$ and n_j includes women having a surgery at $t_{(j)}$. The calculation is given by

$$\hat{S}(0.429) = \prod_{j=1}^{3}(\frac{n_j - d_j}{n_j})$$

$$= (\frac{n_1 - d_1}{n_1})(\frac{n_2 - d_2}{n_2})(\frac{n_3 - d_3}{n_3})$$

$$= (\frac{2101 - 20}{2101})(\frac{2081 - 4}{2081})(\frac{2077 - 8}{2077})$$

$$= 0.9848$$

Obs	Waiting Time	Censor	Age Group	Stage	Income	Region
1	0.143	1	70+	Stage 0	U1	Winnipeg
2	0.143	1	70+	Stage 0	R1	West
3	0.143	1	50-69	Stage 2	U4	Winnipeg
4	0.143	1	70+	Stage 2	R4	East
5	0.143	1	70+	Stage 2	R2	East
6	0.143	0	50-69	Stage 3	R3	West
7	0.143	0	70+	Stage 0	R2	East
8	0.143	1	50-69	Stage 1	R2	East
9	0.143	1	00-49	Stage 1	R1	West
10	0.143	1	50-69	Stage 1	R2	East
11	0.143	1	50-69	Stage 0	U5	Winnipeg
12	0.143	1	50-69	Stage 1	R1	West
13	0.143	1	70+	Stage 1	R4	East
14	0.143	1	00-49	Stage 1	U3	Winnipeg
15	0.143	1	00-49	Stage 1	R1	West
16	0.143	1	50-69	Stage 2	R2	West
17	0.143	1	50-69	Stage 2	U1	Winnipeg
18	0.143	1	70+	Stage 1	U2	Winnipeg
19	0.143	1	50-69	Stage 2	U1	Winnipeg
20	0.143	1	70+	Stage 1	R5	East
21	0.143	1	70+	Stage 2	R1	West
22	0.143	1	00-49	Stage 2	U1	Winnipeg
23	0.286	1	50-69	Stage 0	U5	Winnipeg

Obs	Waiting Time	Censor	Age Group	Stage	Income	Region
24	0.286	0	50-69	Stage 3	U2	Winnipeg
25	0.286	0	50-69	Stage 2	U5	Winnipeg
26	0.286	1	00-49	Stage 1	U3	West
27	0.286	1	50-69	Stage 2	U4	Winnipeg
28	0.286	1	00-49	Stage 2	R5	East
29	0.429	1	70+	Stage 1	R4	East
30	0.429	1	00-49	Stage 1	U3	Winnipeg
31	0.429	1	50-69	Stage 2	U1	Winnipeg
32	0.429	1	70+	Stage 2	U1	West
33	0.429	1	70+		R1	East
34	0.429	1	50-69	Stage 2	U3	Winnipeg
35	0.429	1	00-49	Stage 1	U1	Winnipeg
36	0.429	1	50-69	Stage 2	R2	West

Table 5.8: Waiting Time Data (36 out of 2101)

From Output 5.1, $\hat{S}(4.429) = 0.4967$ is the first survivor function less than 0.5. Therefore, according to equation 2.2.2, the estimate of the median survival time is $\hat{t} = 4.429$, which is the same as the estimate of the *50 percentile* shown in Quintile Estimates table. In another word, 50% of women wait no longer than 4.429 weeks to receive their first surgery.

Product-Limit Survival Estimates						
Waiting Time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	2101
0.143		.	.	.	1	2100
0.143		.	.	.	2	2099
0.143		.	.	.	3	2098
0.143		.	.	.	4	2097
0.143		.	.	.	5	2096
0.143		.	.	.	6	2095
0.143		.	.	.	7	2094
0.143		.	.	.	8	2093
0.143		.	.	.	9	2092
0.143		.	.	.	10	2091
0.143		.	.	.	11	2090
0.143		.	.	.	12	2089
0.143		.	.	.	13	2088
0.143		.	.	.	14	2087
0.143		.	.	.	15	2086
0.143		.	.	.	16	2085
0.143		.	.	.	17	2084
0.143		.	.	.	18	2083
0.143		.	.	.	19	2082
0.143		0.9905	0.00952	0.00212	20	2081

Product-Limit Survival Estimates						
Waiting Time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.143	*	.	.	.	20	2080
0.143	*	.	.	.	20	2079
0.286		.	.	.	21	2078
0.286		.	.	.	22	2077
0.286		.	.	.	23	2076
0.286		0.9886	0.0114	0.00232	24	2075
0.286	*	.	.	.	24	2074
0.286	*	.	.	.	24	2073
0.429		.	.	.	25	2072
0.429		.	.	.	26	2071
0.429		.	.	.	27	2070
0.429		.	.	.	28	2069
0.429		.	.	.	29	2068
0.429		.	.	.	30	2067
0.429		.	.	.	31	2066
0.429		0.9848	0.0152	0.00267	32	2065
.		.	.	.	.	.
4.429		0.4967	0.5033	0.0110	1044	1019
.		.	.	.	.	.
184.429		0	1.0000	0	2040	0

Note: The marked survival times are censored observations.

Quartile Estimates			
Percent	Point Estimate	95% Confidence Interval	
		Lower	Upper
75	6.857	6.571	7.143
50	4.429	4.286	4.571
25	3.000	2.857	3.000

Mean	Standard Error
6.958	0.250

Output 5.1: Kaplan-Meier Estimates of Waiting Times

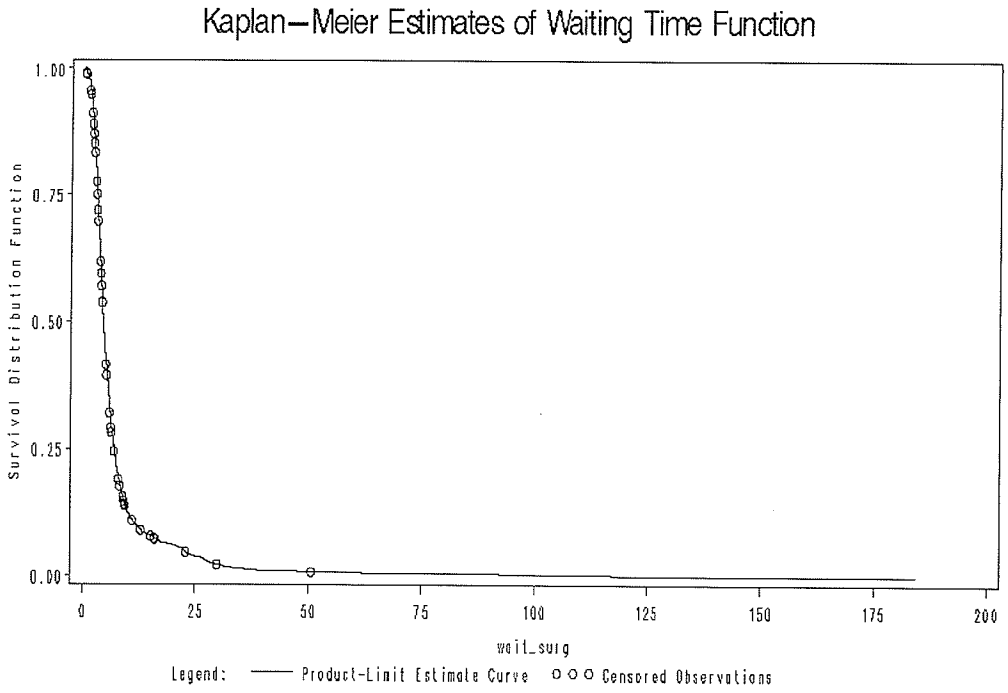


Figure 5.1: Kaplan-Meier Estimates of the Waiting Time Functions

5.3 Detection of the Difference on Waiting Times

In this section, we discuss the results of detecting the difference on survival curves by age group, cancer stage, region, and between urban and rural defined in section 4.2. In section 4.3, we discussed that we could make our conclusion based on the $P-value$ of the log-rank chi-square test from SAS output. The first part of this section shows how to input numerical values into equation 2.3.3 to calculate the log-rank test statistic by using SAS Output 5.2 for age group. In the rest of this section, the conclusions are determined by the $P-value$ from SAS output.

There are three categories for age group. Therefore, the log-rank test statistic follows a chi-square distribution with 2 degree of freedom.

The hypotheses are:

$H_0 : \hat{S}_1(t) = \hat{S}_2(t) = \hat{S}_3(t)$

H_a :*At least one of them is not equal.*

The test statistic is $\chi^2 = U_L' V_L^{-1} U_L$, where U_L is a 3×1 column vector and V_L is a 3×3 *variance-covariance matrix.* The values of each element of U_L and V_L are shown in the Rank Statistics table and in the Covariance Matrix for the Log-Rank Statistics table in Output 5.2. Finally, we input all values into equation 2.3.3 and calculated the log-rank statistic

$$U_L' V_L^{-1} U_L = \begin{pmatrix} 3.912 \\ -29.563 \\ 25.651 \end{pmatrix}' \begin{pmatrix} 349.177 & -236.729 & -112.448 \\ -236.729 & 491.592 & -254.864 \\ -112.448 & -254.864 & 367.312 \end{pmatrix}^{-1} \begin{pmatrix} 3.912 \\ -29.563 \\ 25.651 \end{pmatrix}$$
$$= 2.2311$$

The $P-value = 0.3277$ ($\chi^2 = 2.2311, d.f. = 2$), which is relatively large. Therefore, the log-rank test did not provide sufficient evidence of suggesting a difference among three age groups.

The last table of Output 5.2 gives the estimates of the median waiting times and mean for each age group respectively. The medians of each age group do not vary significantly. Figure 5.2 does not show waiting time curves differ from each other significantly.

Rank Statistics		
Age Group	Log-Rank	Wilcoxon
00-49	3.912	56258
50-69	-29.563	-97352
70+	25.651	41094

Covariance Matrix for the Log-Rank Statistics			
Age Group	00-49	50-69	70+
00-49	349.177	-236.729	-112.448
50-69	-236.729	491.592	-254.864
70+	-112.448	-254.864	367.312

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	2.2311	2	0.3277
Wilcoxon	13.1220	2	0.0014
-2Log(LR)	0.0385	2	0.9809

	00-49	50-69	70+
50 Percentile	4.413	4.714	4.286
Mean	6.936	6.886	7.167

Output 5.2: Test of Difference in Waiting Time for Age group

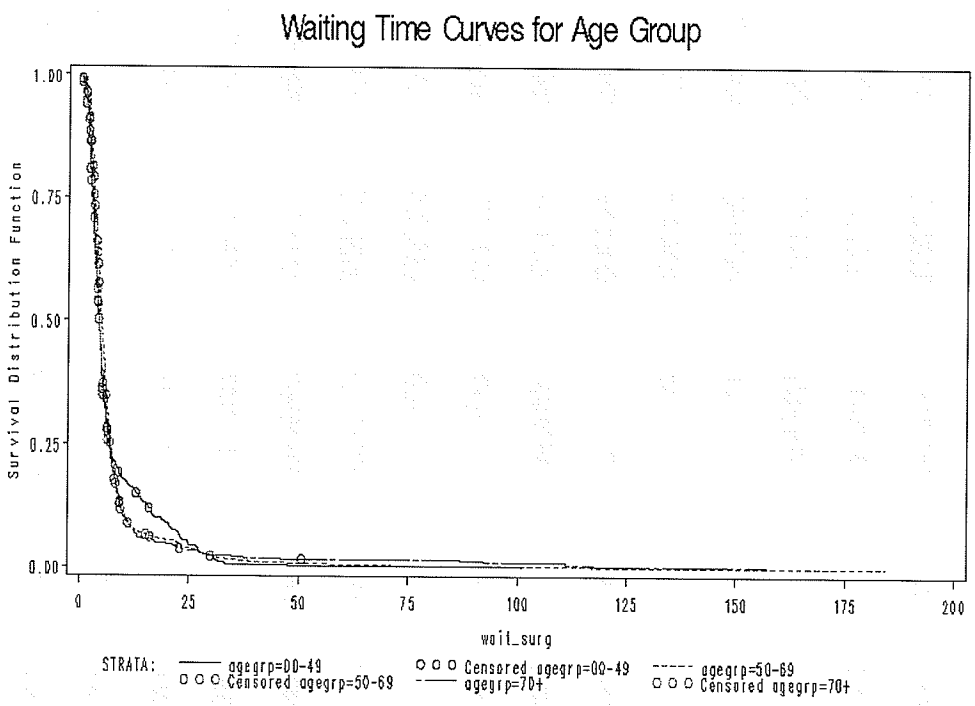


Figure 5.2: Waiting Time Curves for Age Group

Output 5.3 is the SAS output of testing a difference on the waiting times by different cancer stages. The $P-value<0.0001$ ($\chi^2=73.7135, d.f.=4$) is very small. Hence, the waiting times are significantly different among the cancer stages. Figure 5.3 shows women in stage 4 followed by stage 0 waited longer than women in other stages. We can get the same results by comparing the estimates of the median waiting times in Output 5.3. Since stage 0 is an early stage, a surgery is not really needed. The reasons for women in stage 4 waited longer are as follows:

1) Women in stage 4 usually have larger tumours detected. Therefore, tumours are needed to be shrunk by chemotherapy before having a surgery.

2) Women detected with stage 4 may have other comorbidities disease, such as heart attack. The health condition of these women prevents a surgery in a short time.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	73.7135	4	<.0001
Wilcoxon	45.3111	4	<.0001
-2Log(LR)	113.2843	4	<.0001

	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
50 Percentile	5.000	4.286	3.857	3.857	5.071
Mean	5.696	4.769	4.43	7.257	15.742

Output 5.3: Test of Difference in Waiting Time for Cancer Stage

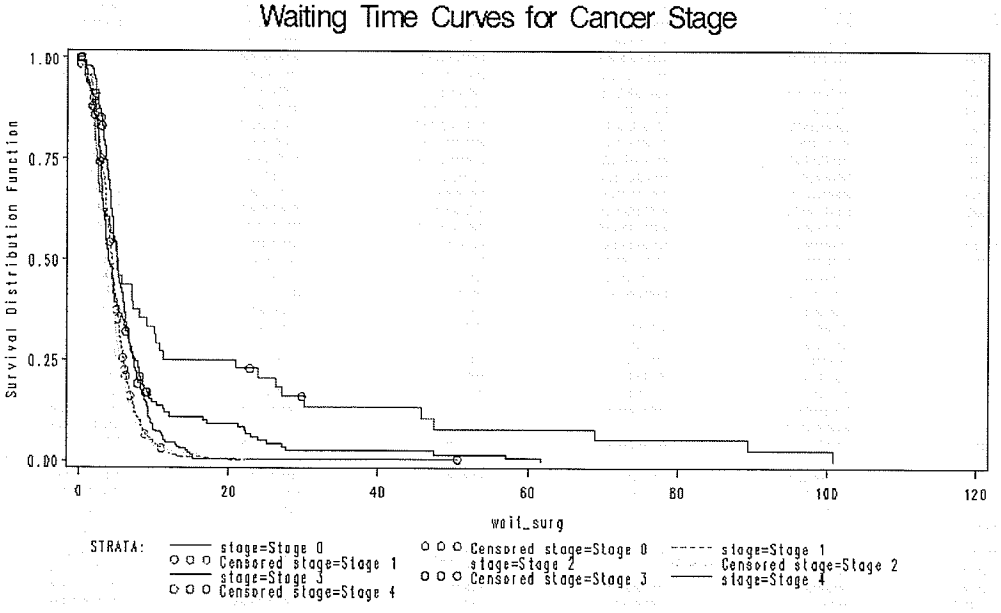


Figure 5.3: Waiting Time Curves for Cancer Stage

Output 5.4 is the results of testing the difference of waiting times in different regions. The $P-value=0.0005$ ($\chi^2=17.9116, d.f.=3$) is very small. Therefore, there was strong evidence to conclude that waiting times were different in the four regions. The estimates of the median waiting times and Figure 5.4 show that women waited longer in the North. The reason is that it lacks of hospitals and surgery can not be operated in the North. In order to have surgery, women have to travel down to Winnipeg or Brandon, while there is no flight or train transportation every day. Therefore women in the North waited longer. Even though we have more surgeons in Winnipeg, it is still short of surgeons comparing with the number of breast cancer woman waiting for surgery. Figure 5.4 shows the same results.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	17.9116	3	0.0005
Wilcoxon	25.5350	3	<.0001
-2Log(LR)	4.2478	3	0.2359

	East	North	West	Winnipeg
50 Percentile	4.429	5.786	4.143	4.571
Mean	7.065	8.311	6.556	6.952

Output 5.4: Test of Difference in Waiting Time for Region

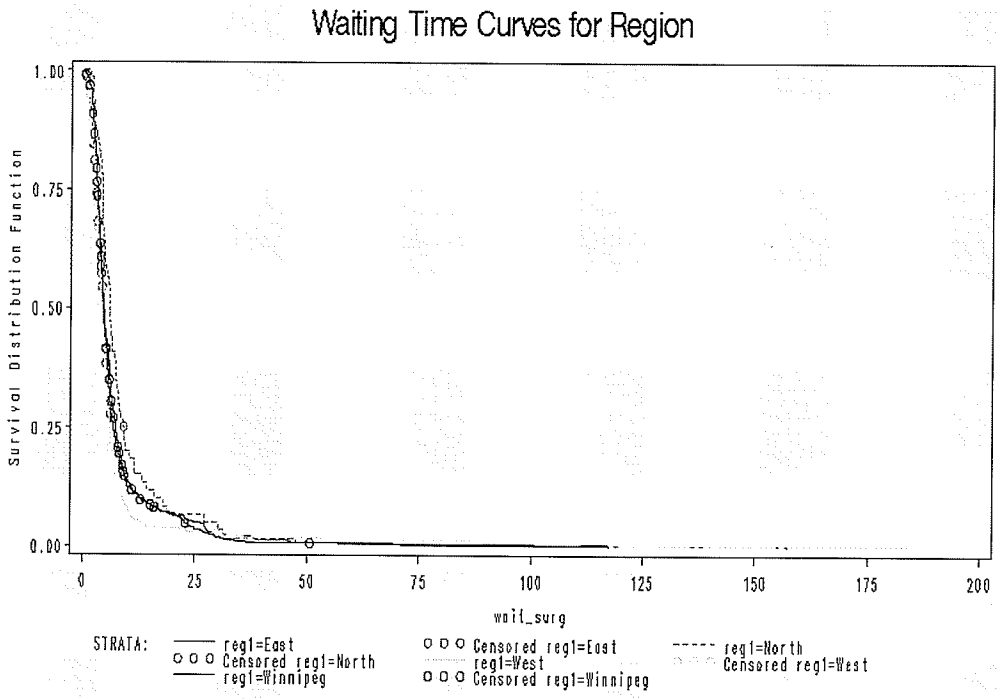


Figure 5.4: Waiting Time Curves for Region

From Output 5.5, we conclude that there is no difference on waiting times between two Urban and Rural since the $P-value = 0.4018$ ($\chi^2 = 0.7030, d.f. = 1$) is very large. The estimates of the median waiting times for urban and rural are close to each other. The waiting time curves are almost identical to each other throughout the study period as shown in Figure 5.5.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.7030	1	0.4018
Wilcoxon	4.0830	1	0.0433
-2Log(LR)	0.4914	1	0.4833

	Urban	Rural
50 Percentile	4.571	4.286
Mean	6.815	7.031

Output 5.5: Test of Difference in Waiting Time for Urban and Rural

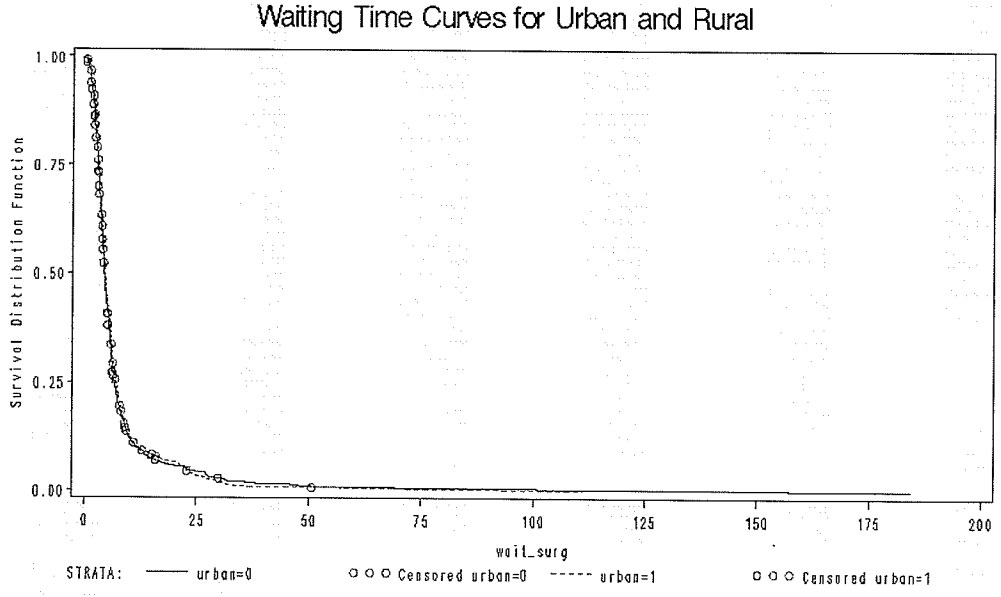


Figure 5.5: Waiting Time Curves for Urban and Rural

According to our analyses, we make the following conclusions about the waiting times to the first surgery for women whose waiting times are not equal to zero:

a) The estimates of the median and mean waiting times to the first surgery are 4.429 weeks and 6.958 weeks respectively.

b) There is no difference on waiting times to access surgery for different age groups, and Urban and Rural.

c) The wait times are significantly different by cancer stage and region. Women in stage 4 waited longer than women in other cancer stages. Women in the North waited for the longest time followed by women in Winnipeg.

Chapter 6

Results of the Survival Time Analysis

In this chapter, we discuss the results of the survival time analysis. All analyses are performed with SAS 9.1. (See Appendices A.7 for SAS codes.)

6.1 The Survival Curve

We divide women into two groups based on their waiting times. One group consists of women whose waiting times are equal to zero. The other group consists of non-zero waiting times. Since there were more than one third of women who had waiting times of zero in the frequency Table 6.1, we want to do a preliminary test by looking at whether the survival curves are different by the variable wait_zero which is set to one if waiting times are zero and zero otherwise. If the survival curves were different, we will analyze the data by comparing the two groups of women. Otherwise we will do the overall analysis for the survival times. The log-rank chi-square test in Output 6.1 did not provide sufficient evidence to detect a difference on survival between the two groups because the $P-value = 0.7431$ ($\chi^2 = 0.1074, d.f. = 1$) is very large. Hence, we will analyze the overall survival curve without dividing women into two groups.

wait_zero	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0 (NO)	2101	30.81	2101	30.81
1(Yes)	4719	69.19	6820	100.00

Table 6.1: Frequency Table for Wait_zero

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.1074	1	0.7431
Wilcoxon	1.0083	1	0.3153
-2Log(LR)	0.0104	1	0.918

Output 6.1: Test of Difference in Survival Time for Wait_zero

Output 6.2 is the summary of quartile estimates. But it does not give the estimate of the median survival time because the data are extremely skewed to the right. The mean survival time is 9.52 years for all women. Figure 6.1 shows that the survival probabilities are decreasing linearly. Before fitting the Cox regression model into the data, we should test whether there is violation against the PH assumption by plotting the negative of the log-survivor functions against observed survival times. Figure 6.2 shows a straight line starting at 0, which indicates that exponential distribution might be an appropriate model to describe the data. Because exponential distribution is a PH model, hence we conclude there is no evidence against the PH assumption.

Quartile Estimates			
Percent	Point Estimate	95% Confidence Interval	
		Lower	Upper
75	.	.	.
50	.	.	.
25	6.9843	6.5927	7.4196

Mean	Standard Error
9.5163	0.0501

Output 6.2: Summary Statistics for Survival Times

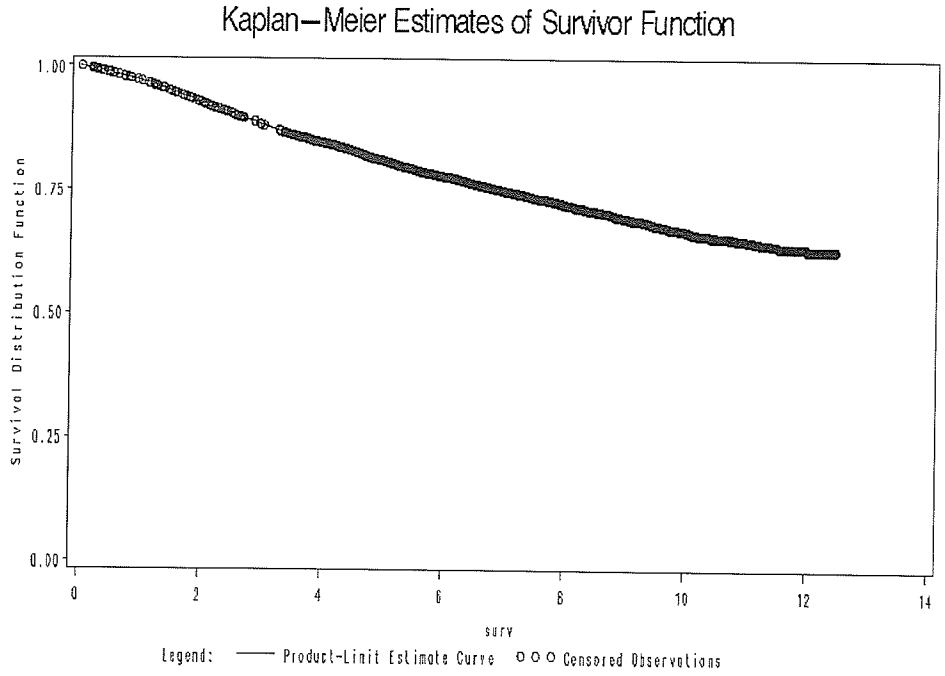


Figure 6.1: Kaplan-Meier Estimates of the Survivor Function

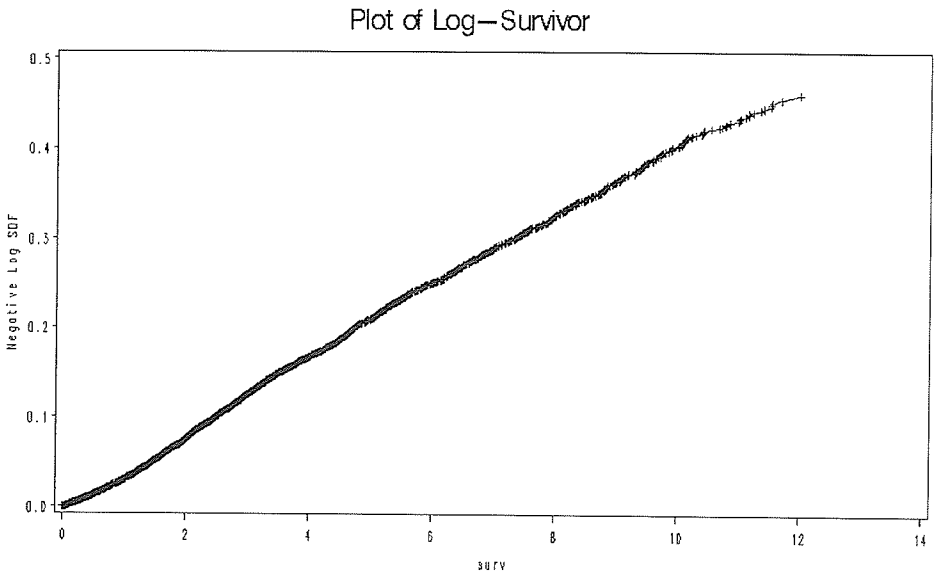


Figure 6.2: Plot of Log-Survivor Versus Survival Times

6.2 Detection of the Difference on Survival Curves

In this section, we discuss the results of testing the survival curves by categorical covariates such as cancer stage, region, urban, rural and income within urban and income within rural respectively.

The $P-value < 0.0001$ ($\chi^2 = 1580.9577, d.f. = 4$) of the log-rank test from Output 6.3 is very small, which indicates a difference on survival by cancer stage. Figure 6.3 clearly shows that the survival probabilities decrease with increasing cancer stage. Therefore, we will keep cancer stage in the Cox regression model.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1580.9577	4	<.0001
Wilcoxon	1655.4630	4	<.0001
-2Log(LR)	756.5250	4	<.0001

Output 6.3: Test of Difference in Survival Time for Cancer Stage

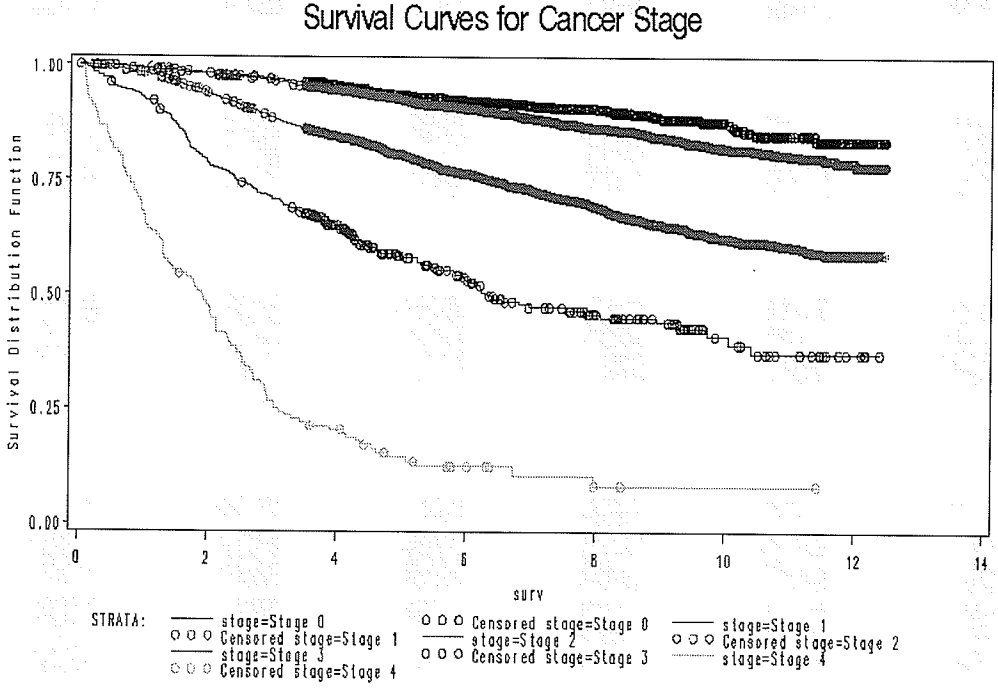


Figure 6.3: Survival Curves for Cancer Stage

Output 6.4 shows that there is some evidence of detecting the difference on survival for region since the $P-value=6.07\%$ is close to the significant level of 5%. Figure 6.4 shows that women from the North lived shorter compared to those from other regions. In the North, there are not enough health care facilities and there is no daily transportation. Therefore women cannot receive daily and good health care services like those from other regions.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	7.3812	3	0.0607
Wilcoxon	7.3192	3	0.0624
-2Log(LR)	7.1242	3	0.0680

Output 6.4: Test of Difference in Survival Time for Region

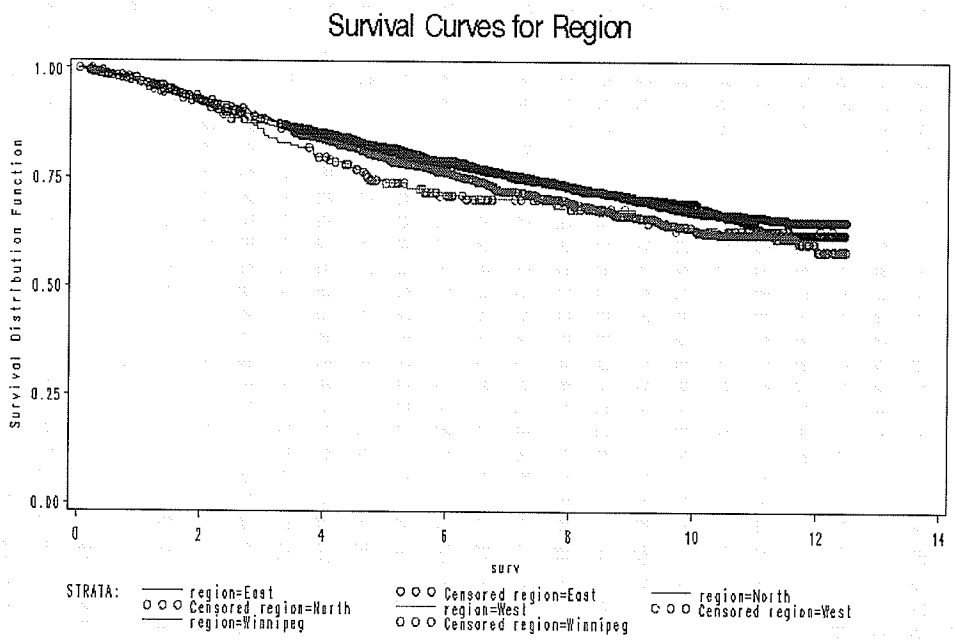


Figure 6.4: Survival Curves for Region

Output 6.5 shows that there is no evidence to conclude that the survival times are different for urban and rural since the $P-value$ is relatively large. The survival curves are not apart to each other in Figure 6.5.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	2.6632	1	0.1027
Wilcoxon	2.0966	1	0.1476
-2Log(LR)	2.7269	1	0.0987

Output 6.5: Test of Difference in Survival Time for Urban and Rural

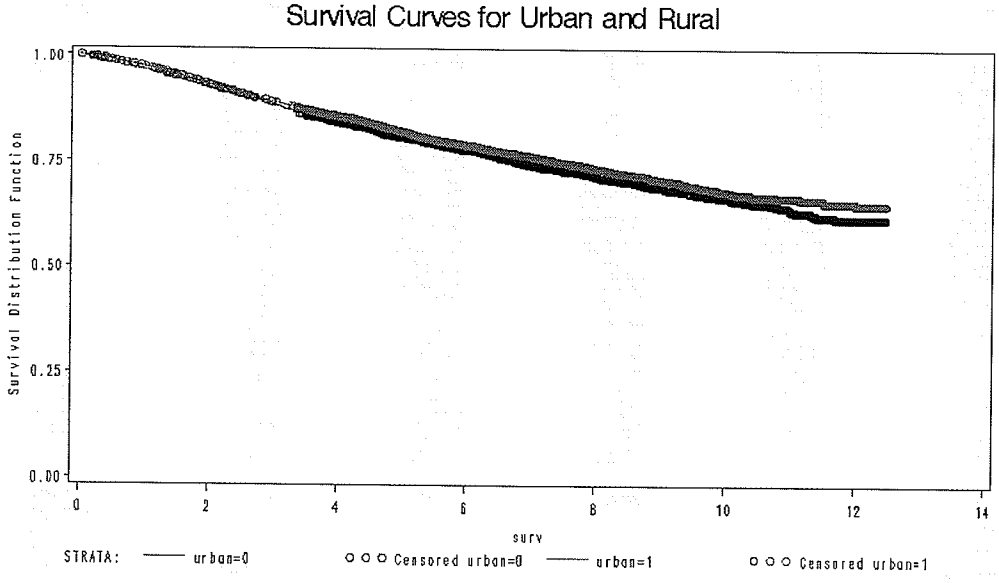


Figure 6.5: Survival Curves for Urban and Rural

Output 6.6 shows that the $P-value$ is significant. We conclude that the survival times are different by income within Urban. Figure 6.6 shows that the survival probabilities are increasing with income increasing. From MCHP website, the average household income of U_5 is almost triple of U_1. Therefore, women with higher income can afford a higher living standard, which means eating better food, accessing to different activities and receiving private health care services.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	65.1607	4	<.0001
Wilcoxon	64.8696	4	<.0001
-2Log(LR)	64.2247	4	<.0001

Output 6.6: Test of Difference in Survival Time for Income within Urban

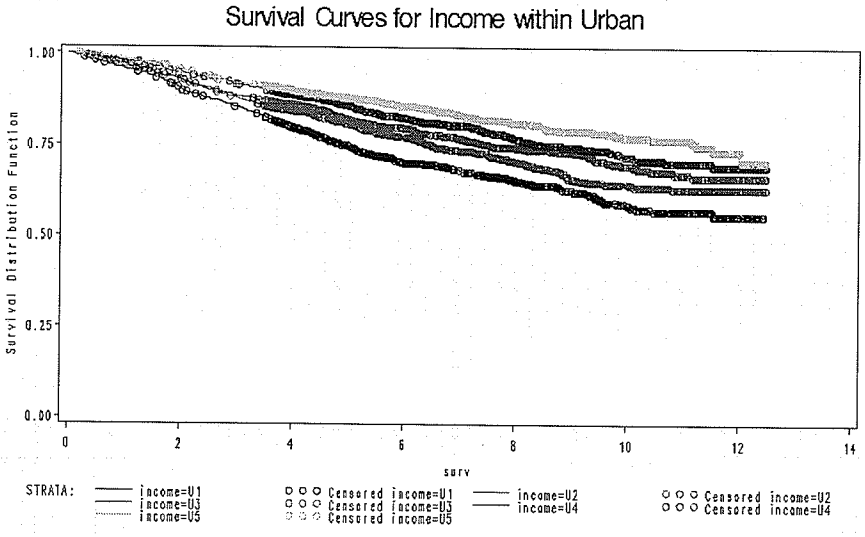


Figure 6.6: Survival Curves for Income within Urban

Output 6.7 shows that the $P-value$ is very large. Therefore, we conclude that the survival times are not different by income within Rural. Figure 6.7 does not show any departure betweens survival curves. The MCHP website does not show significant difference on income from R_1 to R_5. Even though women have different income within rural, they may have similar living standard.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.5798	4	0.3332
Wilcoxon	5.6560	4	0.2264
-2Log(LR)	4.5631	4	0.3351

Output 6.7: Test of Difference in Survival Time for Income within Rural

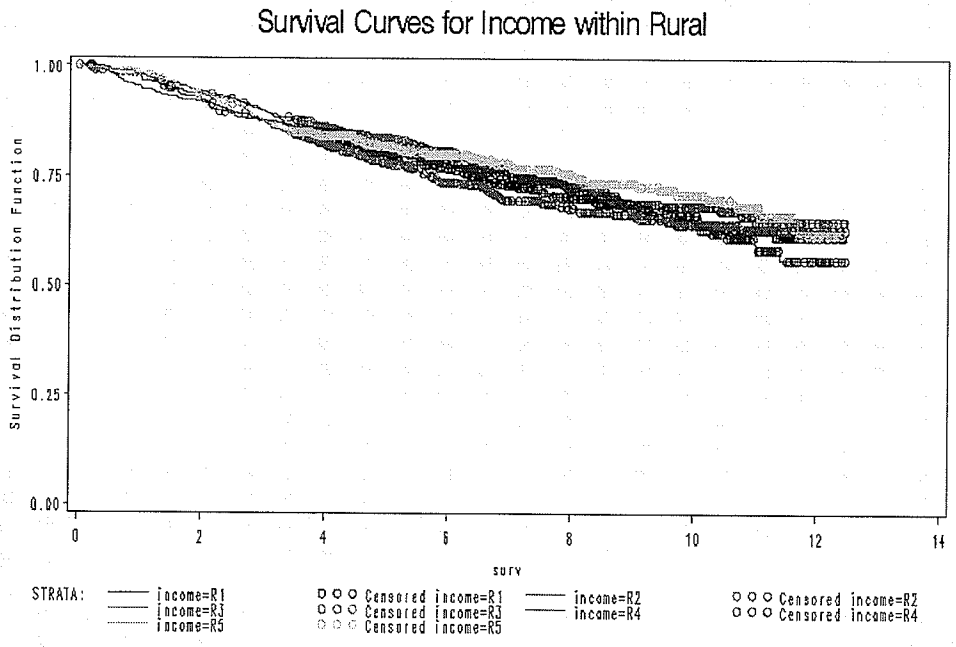


Figure 6.7: Survival Curves for Income within Rural

In summary, we conclude that the survival times are not different by Urban and Rural, and income within rural. The survival times are significantly different by cancer stage and income within urban. Women in higher cancer stage have lower survival probabilities and women with higher income within urban have higher survival probabilities. There is some evidence to conclude that the survival times are different by region.

6.3 Identification of Significant Covariates

In this section, we discuss the results of determining whether numerical covariates of waiting time and diagnosis age have significant effect on the hazard or not by the Wald statistic and Breslow approximation for handling tied data discussed in section 3.3.

Output 6.8 shows the results for testing the null hypothesis $\beta_1 = 0$ for the covariate of waiting time. The 95% confidence interval for the hazard ratio is obtained by specifying RISKLIMITS option under MODEL step. The Wald statistic is calculated from equation 3.3.2

$$\left[\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right]^2$$

$$= \left[\frac{0.00159}{0.0003134}\right]^2$$

$$= 25.83$$

The 95% confidence interval for β_1 by equation 3.3.1 is

$$(\hat{\beta}_1 - z_{\alpha/2} se(\hat{\beta}_1), \quad \hat{\beta}_1 + z_{\alpha/2} se(\hat{\beta}_1))$$

$$= (0.00159 - 1.96 \times 0.0003134, \quad 0.00159 + 1.96 \times 0.0003134)$$

$$= (0.00098, \quad 0.00220)$$

From equation 3.1.3, the hazard ratio (HR) of increasing the waiting time by one day is

$$HR$$

$$= \exp(\hat{\beta}_1 \times (wait\_surg_{12} - wait\_surg_{11}))$$

$$= \exp(0.00159 \times 1)$$

$$= 1.002$$

The 95% confidence interval for the hazard ratio by equation 3.3.3 is

$$(\exp(\hat{\beta}_{1L}), \quad \exp(\hat{\beta}_{1U}))$$

$$= (\exp(0.00098), \quad \exp(0.00220))$$

$$= (1.001, \quad 1.002)$$

Because the $P - value < 0.0001$ is significant and zero does not fall within the 95% confidence interval of β_1, we conclude that the waiting times have significant effect on the hazard and should be included in the Cox regression model. Since the parameter estimate is really small and close to zero, we then exclude the waiting time from the Cox regression model. The interpretation of hazard ratio $(HR = 1.002)$ is that for each one-day increase in the waiting time, the risk of death increases by 0.2 percent.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.7234	1	<.0001
Score	26.6152	1	<.0001
Wald	25.8281	1	<.0001

Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	HR	95% HR CI	
wait_surg	1	0.00159	0.0003134	25.8281	<.0001	1.002	1.001	1.002

Output 6.8: Test of the Null Hypothesis: $\beta_1=0$ for Waiting Times

Output 6.9 is the results of testing whether diagnosis age has an effect on the hazard. The $P-value<0.0001$ is very small, therefore diagnosis age has a significant effect on the hazard and should not be excluded from the model. Even though the parameter estimate is small, we apply the diagnosis age in the Cox regression model. The hazard ratio of 1.045 indicates that the risk of death increases by 4.5 percent for each one-year increase in diagnosis age. In another word, older women have higher risk of death compared to younger women.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	593.6650	1	<.0001
Score	582.5610	1	<.0001
Wald	558.2648	1	<.0001

Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	HR	95% HR CI	
dage	1	0.04425	0.00187	558.2648	<.0001	1.045	1.041	1.049

Output 6.9: Test of the Null Hypothesis: $\beta_1=\beta_2=\beta_3$ for Diagnosis Age

In summary, we keep numerical diagnosis age in the Cox regression model since it has a significant effect on the hazard, and exclude the waiting time from the model.

6.4 Model Selection

From sections 6.2 and 6.3, we have concluded that diagnosis age and cancer stage are the significant covariates for Cox regression. After incorporating both covariates into the model, there are 5524 out of 6280 women left because 1296 women with missing values of cancer stage are deleted by SAS automatically.

In section 3.4, we discussed how to select the most appropriate model by the likelihood-ratio for nested models. Output 6.10 is the summary of values of

$-2Log(\hat{L})$ for different models fitted into the data. First, we incorporate the covariates of diagnosis age and cancer stage into the model and consider it as Model (2) discussed in section 3.4. The value of $-2Log(\hat{L})$ for Model (2) is 19867.905. Then we delete diagnosis age from the model (2) and have model (1) with cancer stage. Model (1) is nested within Model (2). The value of $-2Log(\hat{L})$ for Model (1) is 20156.917. The difference of $-2Log(\hat{L})$ between Model (1) and Model (2) is 289.012, which has a chi-square distribution with one degree of freedom. The corresponding $P-value$ is less than 0.0001, which is significant. We therefore conclude that Model (2) is superior to Model (1). We use the same procedures to compare the model with diagnosis age and cancer stage and the model without cancer stage. It leads to an increase of 799.363 for $-2Log(\hat{L})$ and the corresponding $P-value$ is very small. Hence the model with both the diagnosis age and cancer stage is the most appropriate model.

Model Fit Statistics					
Variables in the model	*-2 LOG L*	Difference Of *-2 LOG L*	DF	Difference of DF	Pr>ChiSq
dage, stage1, stage2, stage3, stage4	19867.905		5		<.0001
stage1, stage2, stage3, stage4	20156.917	289.012	4	1	<.0001
dage	20667.268	799.363	1	4	<.0001

Output 6.10: Values of $-2LogL$ for Different Models Fitted into the Data

6.5 Interpretation of the Cox Regression Analysis

After selecting the most appropriate model, we derive the parameter estimates for the Cox regression model. Output 6.11 contains the results of the Cox regression analysis by using Breslow approximation to handle tied data. The Cox regression model can be written as

$$
\begin{aligned}
& h(t \mid X) \\
& = h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_5 x_{5i}) \\
& = h_0(t)\exp(0.03689 \times dage + 0.28091 \times stage1 + 1.23709 \times stage2 \\
& \qquad +2.08924 \times stage3 + 3.41748 \times stage4)
\end{aligned}
$$

The positive signs of the parameter estimates indicate that older women diagnosed with higher cancer stage comparing to stage 0 have higher risk of death. The $P-value$ is very small for testing the null hypothesis of the coefficient of diagnosis age equal to zero, which indicates that diagnosis age has significant effect on the hazard when covariates stage 1 to stage 4 are fixed in the model. The hazard ratio of 1.038 for diagnosis age indicates that a woman has 3.8 percent higher risk of death compared to a one-year younger woman diagnosed with the same cancer stage. The $P-values$ of covariates stage 1 to stage 4 are all significant at 5%. The hazard ratio of stage 1 to stage 4 can not provide the exact comparison of the risk of death between two cancer stages. In order to explain clearly, let's have a look at the hazard ratio of stage 1. What the hazard ratio does here is comparing the risk of death between two groups of women: one group consists of women in stage 1 and the other group

consists of women in all other stages.

Analysis of Maximum Likelihood Estimates					
Variable	Diagnosis Age	Stage 1	Stage 2	Stage 3	Stage 4
DF	1	1	1	1	1
Parameter Estimate	0.03689	0.28091	1.23709	2.08924	3.41748
Standard Error	0.00221	0.12804	0.11998	0.13859	0.14779
Wald Chi-Square	279.0492	4.8132	106.3086	227.2605	534.7456
Pr > ChiSq	<0.0001	0.0282	<0.0001	<0.0001	<0.0001
HR	1.038	1.324	3.446	8.079	30.492
95% HR Confidence Limits	1.033	1.030	2.724	6.157	22.824
	1.042	1.702	4.359	10.600	40.737
Variable Label	Diagnosis Age	1 if stage 1; otherwise 0	1 if stage 2; otherwise 0	1 if stage 3; otherwise 0	1 if Stage 4; otherwise 0

Output 6.11: Results of the Cox Regression Analysis

In section 6.1, we conclude that women in a higher cancer stage have lower survival probability by the non-parametric method. Output 6.12 shows the results of comparing the hazard ratio between cancer stages by semi-parametric method respectively. The positive sign of the parameter estimate indicates that the risk of

death increase with increasing cancer stage. The hazard ratio of stage0_1 is 1.314, which indicates that the risk of death for women in stage 1 is 1.314 times of women in stage 0 when diagnosis age is constant. The same explanation applies to the hazard ratio of stage1_2, stage2_3 and sage3_4. Finally, we can use the following equation to express the relationship of the risk of death for women with the same diagnosis age, but different cancer stages

$$HR_4$$

$$= 3.391 \times (HR_3)$$

$$= 3.391 \times 2.330 \times (HR_2)$$

$$= 3.391 \times 2.330 \times 2.624 \times (HR_1)$$

$$= 3.391 \times 2.330 \times 2.624 \times 1.314 \times (HR_0)$$

From the equation above, we can see that the risk of death for women in stage 4 is the highest and decreases linearly up to stage 0. Once again, we conclude that women in different cancer stages have different survival times and the risk of death increases with increasing cancer stage when diagnosis age is kept constant.

Analysis of Maximum Likelihood Estimates								
	Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	HR	Variable Label
Stage 0 & Stage 1	dage	1	0.07103	0.00506	197.2790	<.0001	1.074	
	stage0_1	1	0.27281	0.12808	4.5370	0.0332	1.314	1 if stage 1 0 if stage 0
Stage 1 & Stage 2	dage	1	0.04223	0.00265	254.5918	<.0001	1.043	
	stage1_2	1	0.96466	0.07182	180.4320	<.0001	2.624	1 if stage 2 0 if stage 1
Stage 2 & Stage 3	dage	1	0.03151	0.00263	143.7382	<.0001	1.032	
	stage2_3	1	0.84582	0.08913	90.0529	<.0001	2.330	1 if stage 3 0 if stage 2
Stage 3 & Stage 4	dage	1	0.01160	0.00438	7.0171	0.0081	1.012	
	stage3_4	1	1.22117	0.12485	95.6652	<.0001	3.391	1 if stage 4 0 if stage 3

Output 6.12: Comparison of the HR between Cancer Stages

6.6 Estimation of the Survivor Function

After choosing the most appropriative model, we also estimate the survivor function that is discussed in section 3.5. This can be done easily in SAS by specifying BASELINE in PROC PHREG. Recall that SAS uses the sample means as shown in equation 3.5.8. Output 6.13 shows the portion of the estimation of survivor functions. The sample means for the covariates of diagnosis age, stage 1, stage 2, stage 3 and stage 4 are listed from column 2 to column 6. The observed survival times, the estimated survival probabilities, the logarithm of the survival probabilities (also known as the cumulative hazard function) and the logarithm of the cumulative hazard function are listed from column 7 to column 10 respectively. The estimates survival probabilities decrease. See section 3.5 for details about the calculations.

Obs	Diagnosis age	Stage 1	Stage 2	Stage 3	Stage 4
1	60.4808	0.37744	0.39464	0.060282	0.024258
2	60.4808	0.37744	0.39464	0.060282	0.024258
3	60.4808	0.37744	0.39464	0.060282	0.024258
4	60.4808	0.37744	0.39464	0.060282	0.024258
5	60.4808	0.37744	0.39464	0.060282	0.024258
6	60.4808	0.37744	0.39464	0.060282	0.024258
7	60.4808	0.37744	0.39464	0.060282	0.024258
8	60.4808	0.37744	0.39464	0.060282	0.024258
9	60.4808	0.37744	0.39464	0.060282	0.024258

Obs	surv	s	ls	lls
1	0.0000	1.00000	0.00000	.
2	0.0055	0.99989	-0.00011	-9.11303
3	0.0082	0.99978	-0.00022	-8.41916
4	0.0465	0.99967	-0.00033	-8.01339
5	0.0657	0.99956	-0.00044	-7.72551
6	0.0712	0.99945	-0.00055	-7.50219
7	0.0739	0.99923	-0.00077	-7.16480
8	0.0794	0.99912	-0.00088	-7.03056
9	0.0986	0.99900	-0.00100	-6.91201

Output 6.13: Portion of Estimation of Survivor Functions at Sample Means

6.7 Goodness of Fit Assessment of the PH Assumption

In this section, we apply two different methods to check the goodness-of-fit for the PH model.

The first method is plotting the logarithm of the cumulative hazard functions against the logarithm of observed survival times for the categorical variable of cancer stage. If the PH assumption were true, the plot should yield parallel curves as discussed in section 3.6. Figure 6.8 shows roughly parallel curves except for stage 0 and stage 1. Output 6.12 gives us the hazard ratio of stage0_1 as 1.314. Comparing it to the hazard ratios between other cancer stages, it does not indicate significant difference on survival between stage 0 and stage 1. Hence, there is no strong violation to the PH assumption.

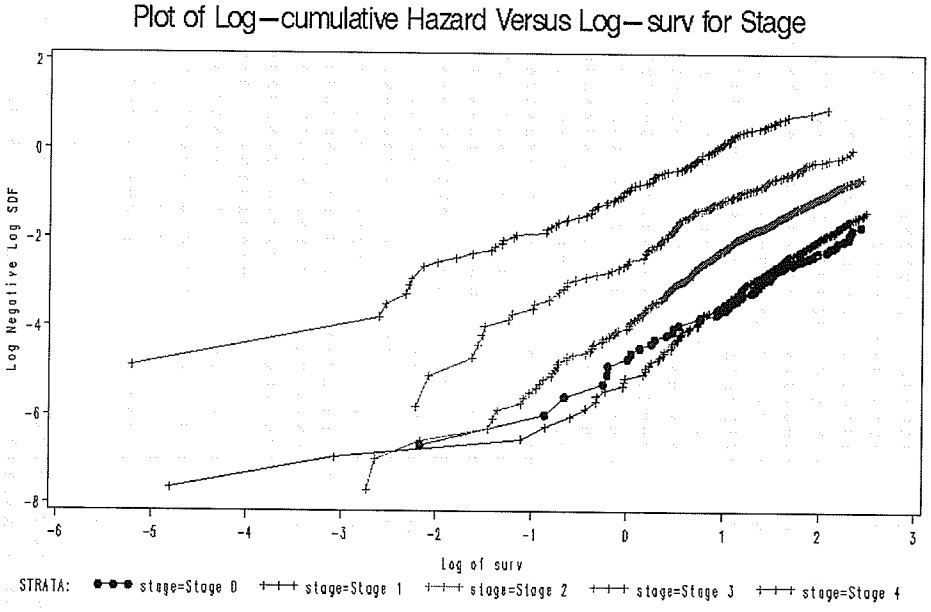
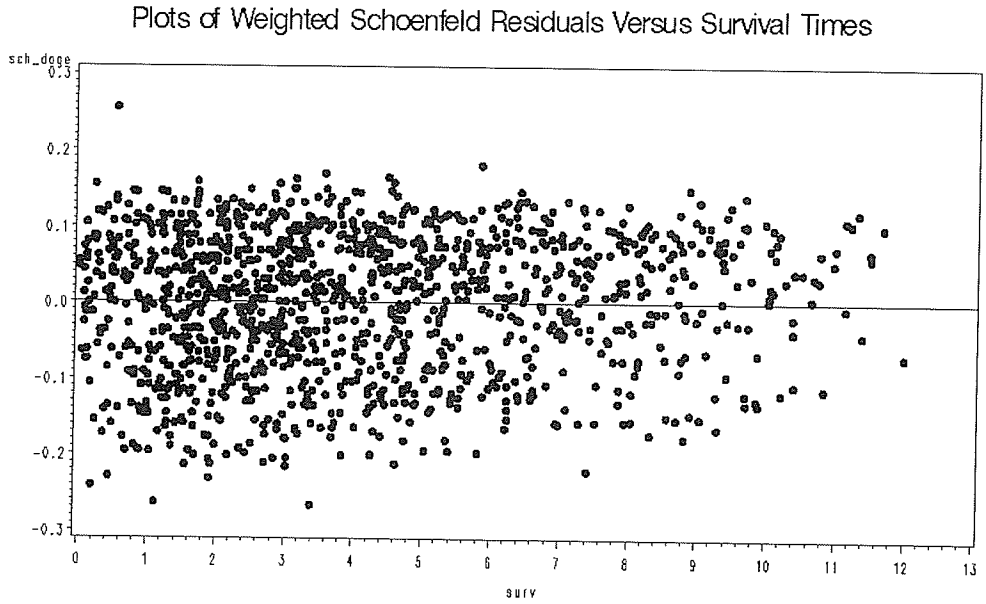
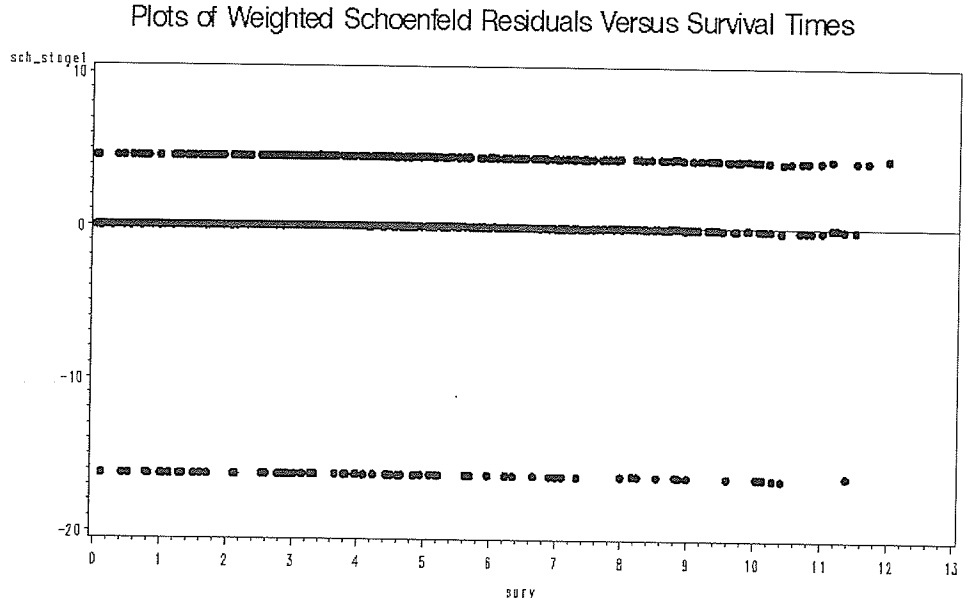


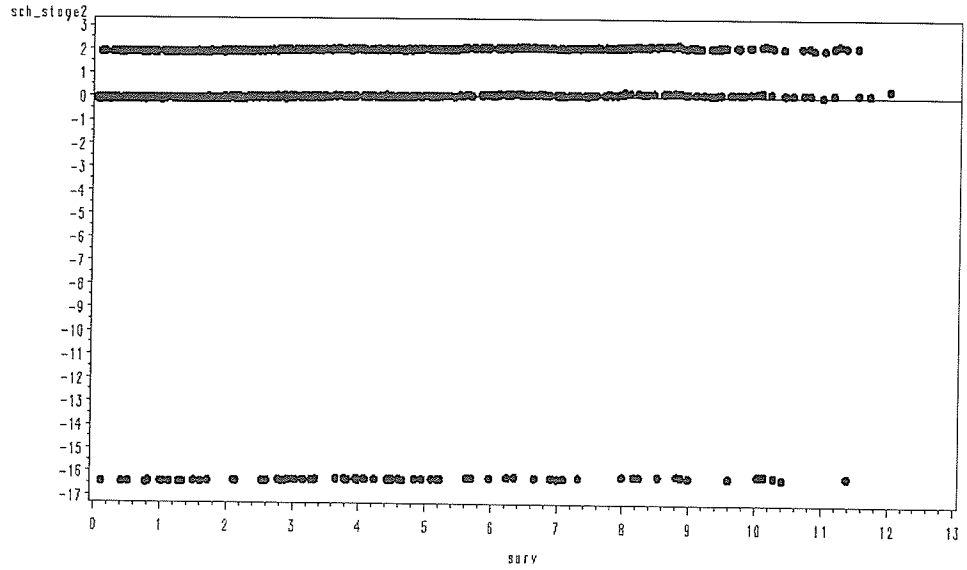
Figure 6.8: Plot of Log-cumulative Hazard Versus Log-surv for Cancer Stage

The second method is using the weighted Schoenfeld residuals discussed in section 3.6. Figure 6.9 shows the plots of the weighted Schoenfeld Resduals versus observed survival times for the diagnosis age, stage 1, stage 2, stage 3 and stage 4 respectively. The plots are roughly symmetrically distributed about zero. Therefore, there is no strong violation against the PH model.

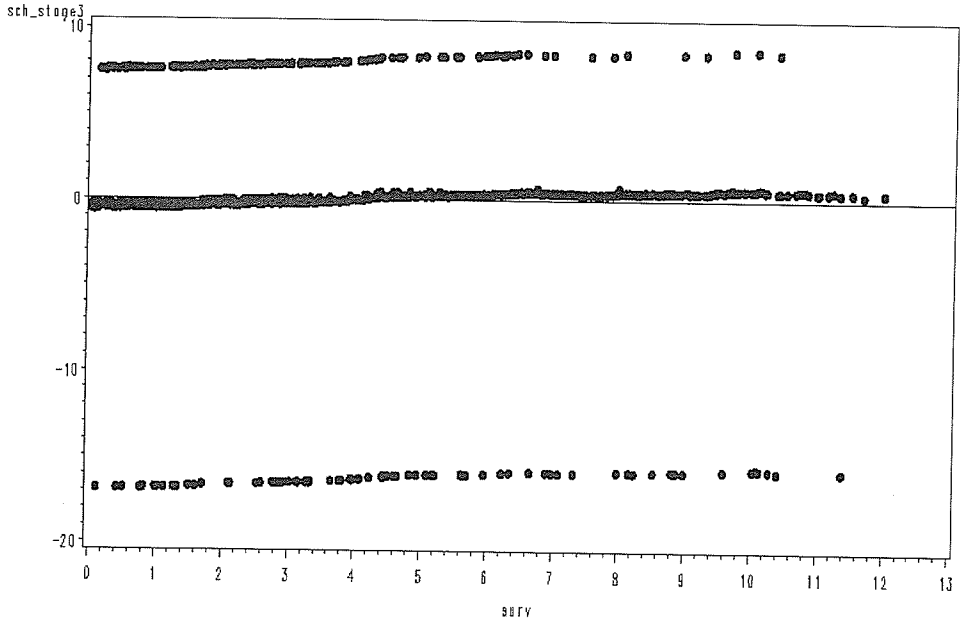




Plots of Weighted Schoenfeld Residuals Versus Survival Times



Plots of Weighted Schoenfeld Residuals Versus Survival Times



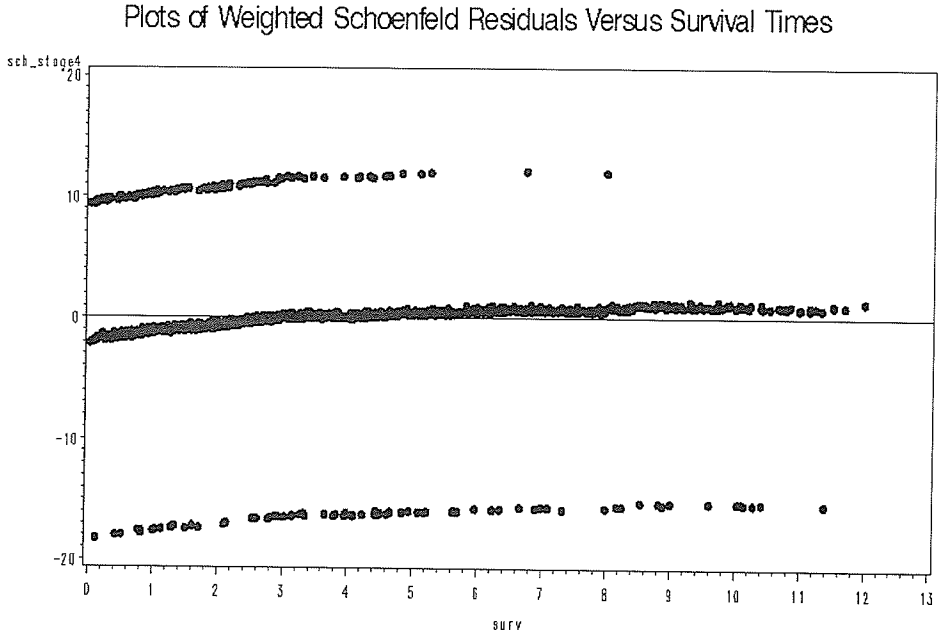


Figure 6.9: Plot of Weighted Schoenfeld Residuals Versus Survival Times

In conclusion, the goodness-of-fit of PH assumption is not violated and the Cox regression model with the covariates of diagnosis age and cancer stage (including stage1 to stage 4) is a good model to describe the survival times.

Chapter 7

Summary and Future Studies

Our objectives for the waiting time analysis are to test whether the waiting times are different by diagnosis age group, cancer stage, region, and between urban and rural. There were 2101 out of 6280 women in the waiting time analysis because 4719 women had surgery on the same date of diagnosis. If we kept waiting times of zero in our analysis, we could not get the estimate of the median waiting times because of extremely right skewness. Therefore, we deleted these data for the waiting time analysis. The estimate of the median waiting times for women with non-zero waiting times was 4.429 weeks. There were no significant differences on waiting times by diagnosis age group, and between urban and rural. We noted that waiting times are significant different by cancer stages. Women with higher cancer stage had longer waiting times because these women needed chemotherapy to shrink tumours. Women from North Manitoba waited longer for their first surgery because the North is a remote living area.

For the survival time analysis, our objectives are to identify significant covariates and select the most appropriate model to describe the survival times. Since there are 4719 out of 6280 women had zero waiting times, we first compared the survival times for two groups of women. Group one are women with zero waiting times and group two are women with non-zero waiting times. We found that the

survival times are not different for these two groups. Hence, we did analysis for all women together. We found that there are no significant difference on survival for waiting times, region, between urban and rural, and income within rural, but the survival times are different by income within urban. The covariates of diagnosis age and cancer stage have significant effects on the hazard and were incorporated into the Cox regression model. With the diagnosis age and cancer stage increasing, the risk of death also increases.

In this study, we also found that the survival times might follow an exponential distribution. My future studies will be using the parametric method to analyze the survival times. Also I will consider whether comorbidities have effects on survival or not.

Appendices

A.1 Variables List

ANDXL, axillary node dissection (1 if AND; 0 otherwise)

CENSOR, indicator (1 if censored data; 0 otherwise)

DAGE, diagnosis age of patient

DDT, diagnosis date

DOC, date of cancellation

DTHDT, death date

ICD9, ICD9 diagnosis code

ICD10, ICD10 diagnosis code

NOPN, number of positive nodes

MASTXL: mastectomy (1 if mastectomy; otherwise)

MPHIN, Manitoba Personal Health Identification Number

PCAD, postal code at diagnosis

PNS, pathology nodal status

PSS, pathology summary stage

PTS, pathology tumour stage

RECONXL: reconstruction (1 if reconstruction; 0 otherwise)

REGION, including North, West, East and Winnipeg

RHA, region of residence (derived from PCAD)

ROC, reason of cancellation

SEGXL, segmental mastectomy (1 if segmental mastectomy; 0 otherwise)

SLNBXL, sentinel lymph node biopsy (1 if SLNB; 0 otherwise)

SEX, sex of patient (Female, Male)

SOT, size of tumour

STAGE, cancer stage

STAGE1, 1 if stage 1; 0 otherwise

STAGE2, 1 if stage 2; 0 otherwise

STAGE3, 1 if stage 3; 0 otherwise

STAGE4, 1 if stage 4; 0 otherwise

STAGE0_1, 1 if stage1; 0 if stage 0

STAGE1_2, 1 if stage2; 0 if stage 1

STAGE2_3, 1 if stage3; 0 if stage 2

STAGE3_4, 1 if stage4; 0 if stage 3

STATUS, alive or dead

SUPI, scrambled unique personal identifier

SURV, survival times in year

TCD9, ICD9 treatment procedure codes

TCD10, ICD10 treatment procedure codes

TDATE, treatment date

UPI, unique personal identifier

URBAN, 1 if women from urban; 0 if women from rural

UTI, unique tumour identifier

WAIT_SURG1, wait times to the first surgery in week

WAIT_SURG, wait times to the first surgery in week

WAIT_ZERO, 1 if wait times is equal to zero; 0 otherwise.

A.2 ICD9 (ICD10) Diagnosis Codes for Breast Cancer

Diagnosis codes: ICD9 (ICD10)	Description
174(C50)	Malignant neoplasm of female breast
174.0(C50.0)	Nipple and areola
174.1(C50.1)	Central portion
174.2(C50.2)	Upper-inner quadrant
174.3(C50.3)	Lower-inner quadrant
174.4(C50.4)	Upper-outer quadrant
174.5(C50.5)	Lower-outer quadrant
174.6(C50.6)	Axillary tail breast
174.8(C50.8)	Other specified site of female breast -Ectopic sites, Midline of breast, Inner breast, Outer breast, Lower breast, Upper breast, Malignant neoplasm of contiguous or overlapping sites of breast whose point of origin cannot be determined
174.9(C50.9)	Breast(female), unspecified
233(D05)	Carcinoma in situ of breast and genitourinary system
233.0(D05.0,DO5.1,D05.7,D05.9)	Breast

A.3 ICD9 (ICD10) Surgery Treatment Procedure Codes

Procedure name	ICD9 codes	ICD10 codes
Sentinel Lymph Node Biopsy	40.29	2MD71LA
Segmental Mastectomy	85.20-85.23, 85.25	1YM87DA,1YM87GB,'1YM87LA, 1YM87LAXXA, 1YM87UT, 1YK87LA,1YK87LAXXA, 1YK87LAXXB, 1YK87LAXXE
Axillary Node Dissection	40.3, 40.51	1MD87LA,1MD89LA, 1MD89LAXXA,1MD89LAXXE, 1MD89LAXXF,1MD89LAXXG, 1MD89LAXXN
Simple Mastectomy	85.41, 85.42	1YM89LA, 1YM89LAXXA
Radical Mastectomy	85.45–85.48	1YM91TR,1YM91TRXXA, 1YM91TRXXE, 1YM91WP, 1YM91WPXXA,1YM91WPXXE, 1YM92LAXXG,1YM92LAXXF, 1YM92LAXXE,1YM92LAPME, 1YM92LAPMF,1YM92LAPMG, 1YM92LATPG,1YM92LATPE, 1YM92LATPF,1YM92LAQFE, 1YM92LAQFF,1YM92LAQFG, 1YM92TRXXE, 1YM92TRXXF, 1YM92TRXXG,1YM92TRPME, 1YM92TRPMF,1YM92TRPMG, 1YM92TRTPE,1YM92TRTPF, 1YM92TRTPG,1YM92TRQFE, 1YM92TRQFF,1YM92TRQFG, 1YM92WPXXE,1YM92WPXXF, 1YM92WPXXG,1YM92WPPME, 1YM92WPPMF,1YM92WPPMG, 1YM92WPTPE,1YM92WPTPF, 1YM92WPTPG, 1YM92WPQFE, 1YM92WPQFF, 1YM92WPQFG
Modified Radical Mastectomy	85.43, 85.44	1YM91LA,1YM91LAXXA, 1YM91LAXXE
Reconstruction	85.7, 85.70-85.79, 85.8, 85.80-85.89, 85.33-85.36	1YM80LAPM,1YM80LAPMA, 1YM80LAPME,1YM80LAPMF, 1YM80LAPMG,1YM80LAQF,

		1YM80LAQFG,1YM80LAQFF, 1YM80LAQFA, 1YM80LAQFE,1YM80LATP, 1YM80LATPG,1YM80LATPF, 1YM80LATPA, 1YM80LATPE,1YM80LA, 1YM80LAXXG,1YM80LAXXF, 1YM80LAXXA, 1YM80LAXXE, 1YM88LAPM,1YM88LAPMG, 1YM88LAPMF,1YM88LAPME, 1YM88LAQF,1YM88LAQFF, 1YM88LAQFG,1YM88LAQFGE, 1YM88LATP,1YM88LATPG, 1YM88LATPF,1YM88LATPFE, 1YM88LAXXF,1YM88LAXXG, 1YM88LAXXE,1YM90LAXXG, 1YM90LAXXF,1YM90LAXXE, 1YM90LAPM, 1YM90LAPMG, 1YM90LAPMF,1YM90LAPME, 1YM90LATP,1YM90LATPG, 1YM90LATPF,1YM90LAQF, 1YM90LAQFG,1YM90LAQFF, 1YM90LAQFE

A.4 SAS Codes of Data Cleaning

```
proc format;
/*
agegrpff, classify numerical values of diagnosis age into 3 categories
$inc96-$inc03, assign income quintile according to post code at diagnosis
mothdiff, classify the numerical values of the difference between the first
          diagnosis date and other diagnosis dates into 4 categories
$mothdiffl, descript full name of 'mothdiff'
$nodord, rank the nodal stages, no grouping
$proccci, classify ICD10 procedure codes to each surgery group
$procd, classify ICD9 procedure codes to each surgery group
$procgp, descript full name of surgery procedures
$regff, classify RHAs into four regions
$rhaf, assign RHA codes Income codes according to post code at diagnosis
$stage, classify pathology summary stage
$staord, rank the summary stage, no grouping
$tuord, rank the tumour stages
wk, classify numerical values of wait times into 5 categories
yr, classify numerical values of survival times into 3 categories
*/
value agegrpff
      0-<49       = '00'
      50-<69      = '50'
      70-high     = '70+';

value mothdiff
      0           = '0'
      1-89        = '3'
      90-< 180    = '6'
      180-high    = '9';

value $mothdiffl
      '0' = 'Same Date'
      '3' = '<3 Mths'
      '6' = '3-<6 Mths'
      '9' = '>6 Mths';
value $nodord
      'n0'      = 1
      'n1 '     = 2
      'n1a'     = 3
      'n1b'     = 4
      'n2'      = 5
      Other     = 0;
```

```
value $proccci
        '2MD71LA'                                                    =' 1'
        '1YM87DA','1YM87GB', '1YM87LA','1YM87LAXXA','1YM87UT'
        '1YK87LA','1YK87LAXXA','1YK87LAXXB','1YK87LAXXE'             =' 2'
        '1MD87LA','1MD89LA','1MD89LAXXA','1MD89LAXXE',
        '1MD89LAXXF','1MD89LAXXG','1MD89LAXXN'                       =' 3'
        '1YM89LA','1YM89LAXXA'                                       =' 4'
        '1YM91LA','1YM91LAXXA','1YM91LAXXE'                          =' 5'
        '1YM91TR','1YM91TRXXA','1YM91TRXXE',
        '1YM91WP','1YM91WPXXA','1YM91WPXXE',
        '1YM92LAXXG', '1YM92LAXXF','1YM92LAXXE',
        '1YM92LAPME','1YM92LAPMF','1YM92LAPMG',
        '1YM92LATPG','1YM92LATPE','1YM92LATPF',
        '1YM92LAQFE','1YM92LAQFF','1YM92LAQFG',
        '1YM92TRXXE','1YM92TRXXF','1YM92TRXXG',
        '1YM92TRPME','1YM92TRPMF','1YM92TRPMG',
        '1YM92TRTPE','1YM92TRTPF','1YM92TRTPG',
        '1YM92TRQFE','1YM92TRQFF','1YM92TRQFG',
        '1YM92WPXXE','1YM92WPXXF','1YM92WPXXG',
        '1YM92WPPME','1YM92WPPMF','1YM92WPPMG',
        '1YM92WPTPE','1YM92WPTPF','1YM92WPTPG',
        '1YM92WPQFE','1YM92WPQFF','1YM92WPQFG'                       =' 6'
        '1SZ87LA','1SZ87LAXXA','1SZ87LAXXE','1SZ87LAXXF',
        '1YM80LAPM','1YM80LAPMA','1YM80LAPME','1YM80LAPMF',
        '1YM80LAPMG', '1YM80LAQF', '1YM80LAQFG','1YM80LAQFF',
        '1YM80LAQFA', '1YM80LAQF','1YM80LAQFG','1YM80LAQFF',
        '1YM80LAQFA','1YM80LAQFE', '1YM80LATP','1YM80LATPG',
        '1YM80LATPF','1YM80LATPA','1YM80LATPE','1YM80LA',
        '1YM80LAXXG','1YM80LAXXF','1YM80LAXXA','1YM80LAXXE',
        '1YM88LAPM','1YM88LAPMG','1YM88LAPMF','1YM88LAPME',
        '1YM88LAQF','1YM88LAQFF','1YM88LAQFG','1YM88LAQFGE',
        '1YM88LATP', '1YM88LATPG','1YM88LATPF','1YM88LATPFE',
        '1YM88LAXXF','1YM88LAXXG','1YM88LAXXE','1YM90LAXXG',
        '1YM90LAXXF','1YM90LAXXE','1YM90LAPM','1YM90LAPMG',
        '1YM90LAPMF','1YM90LAPME','1YM90LATP','1YM90LATPG',
        '1YM90LATPF','1YM90LAQF','1YM90LAQFG', '1YM90LAQFF',
        '1YM90LAQFE'                                                 =' 7'
        other                                                        ='00';

value $procd
        '4029'                                                       = ' 1'
        '8520','8521','8522','8523', '8525'                         = ' 2'
        '403', '4051'                                                = ' 3'
        '8541','8542'                                                = ' 4'
```

```
        '8543','8544'                                                    = ' 5'
        '8545','8546','8547','8548'                                      = ' 6'
        '857', '8570'-'8579','858','8580'-'8589','8533'-'8536'           = ' 7'
        other                                                            = '00';

 value $procgp
        ' 1'    =   'Sentinel Lymph Node Biopsy'
        ' 2'    =   'Segmental mastectomy'
        ' 3'    =   'Axillary node dissection'
        '4'     =   'Simple mastectomy'
        '5'     =   'Modified radical mastecotmy'
        '6'     =   'Radical mastectomy'
        '7'     =   'Reconstruction'
        '00'    =   'Other';

 value $regff
        '10'                = 'Winnipeg'
        '90', '80','70'     ='North'
        '60','45','30'      ='West'
        '40','30','25','20'='East'
        Other               =   ' ';

 value $staord
        '0 '    = 1
        'i '    = 2
        'iia '  = 3
        'iib '  = 4
        'iiia'  = 5
        'iiib'  = 6
        'iv'    = 7
        other = 0;

value $stage
        '0'                 = 'Stage 0'
        'i'                 = 'Stage 1'
        'iia','iib'         = 'Stage 2'
        'iiia', 'iiib', 'iiic' = 'Stage 3'
        'iv','yiv'          = 'Stage 4'
        other               = '   ';

value $tuord
        't0 '   = 1
        'tis'   = 2
        't1', 't1mic' = 3
```

```
        't1a' = 4
        't1b' = 5
        't1c' = 6
        't2' = 7
        't3' = 8
        't4' = 9
        't4a' = 10
        't4b' = 11
        't4c' = 12
        't4d' = 13
        other = 0;

  value wk
        0   -< 14 = '   < 2 weeks'
        14-< 28 = '2 - 4 weeks'
        28-< 56 = '4 - 8 weeks'
        56-< 84 = '8 -12 weeks'
        84-high = '   >12 weeks';

 value yr
        0 -< 5   = ' < 5 years'
        5 -< 10 = ' 5-10 years'
        10-high = ' >10 years';


*Select the Cohort from December 2006 Dataset

data cohort_list;
   set data_06(keep=upi uti icd9 icd10 ddt sex pcad status nopn pts pns pss sot
                    dthdt dage mphin
                    where=((( substr(icd9,1,3)='174' or icd9 = '2330') OR
                    (substr(icd10,1,3) = 'C50' OR substr(icd10,1,3) = 'D05'))
                    AND (1995<=year(ddt)<=2003) and (sex='F')
                   AND (substr(pcad,1,1)='R')));
   dy=year(ddt)
run;


*Update Death Date

data data_07;
   set data_07;
   status1=status;
   dthdt1=dthdt;
keep upi uti dthdt1 status1;
run;
```

```
proc sort data=data_07;
   by upi uti ;
run;

proc sort data= cohort_list;
   by upi uti;
run;

data upcohort;
   merge cohort_list (in=m1) data_07 (in=m2);
   by upi uti;
   if m1 ;
run;

data upcohort1;
   set upcohort;
   if status1^= ' ' then status2=status1;
      else status2=statust;
   if dthdt1^=. then dthdt2=dthdt1;
      else dthdt2=dthdt;
   if dthdt^=ddt;
   drop dthdt status dthdt1 status1 ;
run;

data upcohort2;
   set upcohort1;
   dthdt=dthdt2;
   status=status2;
   drop dthdt2 status2;
run;


*Select One Tumour Per Woman

data tum;
   set upcohort2;
   dropflag=0;
   label dropflag='Reason for dropping tumour';
run;

proc sort data= tum;
   by upi ddt;
run;
```

```
data dxdf;
   set tum;
   by upi ddt;
   retain firstddt 0;
   format firstddt yymmdd10.;
   firstcr=first.upi;
   lastcr=last.upi;
   firstdx=first.ddt;
   lastdx=last.ddt;
   if (first.upi) or (ddt=firstddt) then diff=put(0,mdiff.);
      else diff=put(ddt-firstddt,mothdiff.);
   if first.upi then firstddt=ddt;
   format diff $mothdiffl.;
   staord = put(pss,staord.);
   norod = put(pns,nodord.);
   tuord = put(pts,tuord.);
   format firstddt yymmdd10.;
run;

proc sort data=dxdf;
   by upi;
run;

* If the diagnosis date (ddt) is greater than 6 months after the first ddt
  → remove the later tumours.(dropflag=1;
data remove1 drop1;
   set dxdf;
   if diff^='9' then output remove1;
      else do;
      dropflag=1;
      output drop1;
   end;
run;

data mult1;
   set remove1;
   by upi;
   if (^first.upi or ^last.upi);
run;

proc sort data=mult1;
   by upi ddt;
run;
```

```
proc sort data=mult1;
   by upi descending staord;
run;

*Check the pathological summary stages (pss), if both pss are known and
 different→remove the tumours with the lower stage.(dropflag=2);
data drop2;
   set mult1;
   by upi;
   retain prestaord;
   if first.upi then prestaord=staord;
   else if prestaord^='0' & staord^='0' then if prestaord^=staord then do;
   dropflag=2;
   output drop2;
   end;
run;

proc sort data=remove1;
   by upi uti;
run;

data remove2;
   merge remove1 (in=m1) drop2 (in=m2);
   by upi uti;
   if (^m2);
run;

proc sort data=remove2;
   by upi uti;
run;

data mult2;
   set remove2;
   by upi;
   if (^first.upi or ^last.upi);
run;

proc sort data=mult2;
   by upi ddt;
run;

proc sort data=mult2;
   by upi descending nodord;
run;
```

```
*Check the pathological nodal status (pns), if both pns are known and different
  →remove the tumours with lower pns.(dropflag=3)*/
 data drop3;
    set mult2;
    by upi;
    retain prenodord;
    if first.upi then prenodord=nodord;
    else if prenodord ^='0' & nodord ^='0' then if prenodord ^=nodord then do
    dropflag=3;
    output drop3;
    end;
 run;

 data remove3;
    merge remove2 (in=m1) drop3 (in=m2);
    by upi uti;
    if (^m2);
 run;

 proc sort data=remove3;
    by upi uti;
 run;

 data mult3;
    set remove3;
    by upi;
    if nopn=' ' then nopn ='0';
    if (^first.upi or ^last.upi);
 run;

 proc sort data=mult3;
    by upi ddt;
 run;

 proc sort data=mult3;
    by upi descending nopn;
 run;

 *Check the number of positive nodes (nopn), if one has higher positive nodes and
    others have lower or blank→remove tumours with lower/blank nopn.(dropflag=4);
 data drop4;
    set mult3;
    by upi;
```

```
    retain prevnpos;
    if first.upi then prevnpos=nopn;
    else if prevnpos^=nopn then do;
    dropflag=4;
    output drop4;
    end;
 run;

 data remove4;
    merge remove3 (in=m1) drop4 (in=m2);
    by upi uti;
    if (^m2);
 run;

 proc sort data=remove4;
    by upi uti;
 run;

 data mult4;
    set remove4;
    by upi;
    if (^first.upi or ^last.upi);
 run;

 proc sort data=mult4;
    by upi ddt;
 run;

 proc sort data=mult4;
    by upi descending tuord;
 run;

 *Check the pathological tumour stage (pts), if pts are known and different→
  remove tumours with the lower pts.(dropflag=5);
 data drop5;
    set mult4;
    by upi;
    retain pretuord;
    if first.upi then pretuord=tuord;
    else if (pretuord^=tuord)&( tuord^='0') then do;
    dropflag=5;
    output drop5;
    end;
 run;
```

```
data remove5;
   merge remove4 (in=m1) drop5 (in=m2);
   by upi uti;
   if (^m2);
run;

proc sort data=remove5;
   by upi uti;
run;

data mult5;
   set remove5;
   by upi;
   if (^first.upi or ^last.upi);
run;

proc sort data=mult5;
   by upi ddt;
run;

*Check size of tumour size(sot), if one tumour has a smaller size→remove
 tumours with smaller size.(dropflag=6);
data mult5;
   set mult5;
   length nsize stsize endsize 4.;
   if substr(sot,1,1)='c' then do;
   stsize=indexc(sot,'0123456789.');
   endsize =indexc(sot,'m')-2;
   nsize=put(substr(sot,stsize,endsize-stsize+1),4.);
   end;
      else do;
      stsize=indexc(sot,'0123456789.');
      endsize =indexc(sot,'c')-1;
   if stsize=0 then nsize=0;
      else do;
         if endsize=-1 then endsize =length(sot);
         nsize=put(substr(sot,stsize,endsize-stsize+1),4.);
         end;
      end;
   format nsize 4.2;
run;

proc sort data=mult5;
```

```
    by upi descending tuord descending nsize;
 run;

 data drop6;
    set mult5;
    by upi;
    retain pretuord prevnsiz ;
    if first.upi then do;
    prevnsiz=nsize;
    pretuord=tuord;
    end;
    else if (pretuord ^= '2' & tuord^= '2' ) then if (prevnsiz^=nsize) then do;
    dropflag=6;
    output drop6;
    end;
 run;

 data remove6;
    merge remove5 (in=m1) drop6 (in=m2);
    by upi uti;
    if (^m2);
run;

proc sort data=remove6;
     by upi ddt;
run;

data mult6;
   set remove6;
   by upi;
   if (^first.upi or ^last.upi);
run;

proc sort data=mult6;
   by upi ddt;
run;

* Randomly select one tumour left from the above steps.(dropflag=7);
data drop7;
   set mult6;
   by upi;
   length numb prevnumb 4.;
   retain prevnumb;
   if first.upi then do;
```

```
    numb=put(10*ranuni(10),3.0);
    prevnumb=numb;
    end;
    else numb=prevnumb+1;
run;

data drop7;
   set drop7(where=(mod(numb,2)=0));
   dropflag=7;
run;

data alldrop;
   set drop1 drop2 drop3 drop4 drop5 drop6 drop7;
run;

data merdrop;
   set alldrop (keep=upi ddt uti dropflag);
run;

proc sort data= tum;
   by upi ddt uti;
run;

proc sort data=merdrop;
   by upi ddt uti;
run;

data brcamult error;
   merge tum (in=m1) merdrop (in=m2);
   by upi ddt uti;
   if m1 then output brcamult;
   else output error;
run;


*Create Scrambled Unique Personal Identifier
data getphin;
   set brcamult;
   where dropflag = 0;
   keep upi uti mphin ddt ;
run;

data nophin fndphin;
   set getphin;
```

```
    if mphin = ' ' then output nophin;
    if mphin ^= ' ' then output fndphin;
run;

data studynum;
   set fndphin;
   supi = _n_;
run;

*Attach Information on Termination of Coverage
proc sort data= studynum;
   by supi;
run;

proc sort data= cov_mh;
   by supi;
run;

data cohort2;
   merge studynum (in=m1) cov_mh (in=m2);
   by supi;
   if m1;
run;

*Attach Treatment Information
data txmt;
   set txmt (keep=upi uti ticd9 ticd10 tdate);
   tdt=input(tdate,date9.);
   txyr=year(tdt);
   format tdt yymmdd10.;
   drop tdate;
   txproc = put(ticd9,$procd.);
   txprocci=put(ticd10,$proccci.);
   if txproc^ ='00' then txpc=txproc;else txpc=txprocci;
   keep upi uti txpc tdt;
run;

proc sort data=txmt;
   by upi uti;
run;

proc sort data= cohort2;
```

```
    by upi uti;
 run;

 data cr_txmt;
    merge cohort2(in=m1) txmt (in=m2);
    by upi uti;
    if m1 ;
 run;


 *Select the First Surgery for Each Woman
 data surg;
    set cr_txmt;
    if txpc in (' 1',' 2',' 3',' 4',' 5',' 6',' 7');
 run;

 data crflags;
    set surg;
    if txpc in (' 4',' 5',' 6') then mastx = 1; else mastx = 0;
    if txpc =' 1' then slnbx = 1; else slnbx = 0;
    if txpc in (' 3)then andx = 1; else andx = 0;
    if txpc = ' 2' then segx = 1; else segx = 0;
    if txpc = '7' then reconx = 1; else reconx = 0;
    wait_surg 1= abs(tdt-ddt)/7;
 run;

 proc sort data=crflags out=use;
    by upi wait_surg1;
 run;

 data ana_cohort;
    set use;
    by upi wait_surg1;
    if first.upi ;
 run;

 data onerec;
    set use;
    by upi;
    retain andxl mastxl segxl slnbxl reconxl;
    if first.upi then do
    andxl = 0;
    mastxl = 0;
    segxl = 0;
```

```
    slnbxl = 0;
    reconxl = 0;
    end;
    if andx = 1 then andxl = 1;
    if slnbx = 1 then slnbxl = 1;
    if mastx = 1 then mastxl = 1;
    if segx = 1 then segxl = 1;
    if reconx = 1 then reconxl = 1;
    if last.upi;
    keep upi andxl mastxl segxl slnbxl reconxl;
run;

proc sort data=onerec;
   by upi;
run;

proc sort data=ana_cohort;
   by upi;
run;

data ana_cohort1;
   merge onerec(in=m1) ana_cohort(in=m2);
   by upi;
   if m1;
run;
```

A.5 SAS Codes of Defining Variables

```
*Define Variables
data inc;
   set ana_cohort1;
   rha=put(pcad, $rha04f.);
   region=put(rha, $regff.);
   urbrha=0;
   if rha in ('10','15') then urbrha=1;
   label urbrha='1=WPG/BRAN';
   if year(ddt) in (1995, 1996) then income=put((put(urbrha,1.)||pcad),$inc96.);
   if year(ddt)=1997 then income=put((put(urbrha,1.)||pcad),$inc97.);
   if year(ddt)=1998 then income=put((put(urbrha,1.)||pcad),$inc98.);
   if year(ddt)=1999 then income=put((put(urbrha,1.)||pcad),$inc99.);
   if year(ddt)=2000 then income=put((put(urbrha,1.)||pcad),$inc00.);
   if year(ddt)=2001 then income=put((put(urbrha,1.)||pcad),$inc01.);
   if year(ddt)=2002 then income=put((put(urbrha,1.)||pcad),$inc02.);
```

```
    if year(ddt)=2003 then income=put((put(urbrha,1.)||pcad),$inc03.);
    stage=put(pss, $stage.);
   if income in ('R1','R2','R3','R4','R5') then urban=0;
     else if income in ('U1','U2','U3','U4','U5') then urban=1;
          else urban=' ';
    keep upi stage income region dage doc roc dthdt status wait_surg1 urban;
run;
```

A.6 SAS Codes of the Waiting Time Analysis

```
*Program of Wait Time Analysis
data wait_times;
   set inc;
   agegrp=put(dage, agegrpff.);
   if tdt>'31dec03'd then do;
   censor=0;
   wait_surg=abs('31dec03'd-ddt)/7;
   end;
   if tdt<='31dec03'd then do;
   censor=1;
   wait_surg=abs(tdt-ddt)/7;
   end;
   if wait_surg=0 then wait_zero=1;
      else wait_zero=0;
   keep stage income region agegrp censor wait_surg urban;
run;

proc freq;
   tables wait_zero;
run;

data wait_times;
   set wait_times;
   if wait_zero=0;
run;

proc freq;
   tables andxl mastxl segxl slnbxl reconxl censor wait_surg agegrp stage income
         region /list missing;
   format wait_surg wk.;
run;

proc lifetest data= wait_times outsurv=a method=km plots=(s) ;
```

```
    time wait_surg*censor(0);
    title1 'Kaplan-Meier Estimates of the Survivor Function';
run;
proc print data=a;
run;

proc lifetest data= wait_times method=km plots=(s) ;
    strata agegrp;
    time wait_surg* censor(0);
    title1 'Waiting Time Curves for Age Group';
run;

proc lifetest data= wait_times method=km plots=(s) ;
    where stage^='';
    strata stage;
    time wait_surg* censor(0);
    title1 "Waiting Time Curves for Stage';
    symbol1 v=none color=blue line=1;
    symbol2 v=none color=red line=2;
   symbol3 v=none color=green line=10;
   symbol4 v=none color=purple line=10;
   symbol5 v=none color=yellow line=10;
run;

proc lifetest data= wait_times method=km plots=(s) ;
    where region^=' ';
    strata region;
    time wait_surg* censor(0);
    title1 'Waiting Time Curves for Region';
run;

proc lifetest data=wait_times method=km plots=(s);
    where urban^='';
    strata urban;
    time wait_surg*censor(0);
    title1 'Waiting Time Curves for Urban and Rural';
run;
```

A.7 SAS Codes of the Survival Time Analysis

*Program of the Survival Analysis

```
data inc1;
   set inc;
```

```
    if stage = 'Stage 1' then stage1 = 1;else stage1=0;
    if stage = 'Stage 2' then stage2 = 1;else stage2=0;
    if stage = 'Stage 3' then stage3 = 1;else stage3=0;
    if stage = 'Stage 4' then stage4 = 1;else stage4=0;
run;

data surv_ana;
   set inc1;
   if (status='d' and dthdt^=.) then do ;
   surv=(dthdt-ddt)/365.25;
   censor=1;
   end;
   if (status='a' and doc^=. ) then do ;
      if roc ='2' then do;
      surv=(doc-ddt)/365.25;
      censor=1;
   end;
      else do;
      surv=(doc-ddt)/365.25;
      censor=0;
      end;
   end;
   if (status='a' and doc =.) then do ;
   surv=('30JUN07'D-ddt)/365.25;
   censor=0;
   end;
   surv=abs(surv);
   wait_surg=wait_surg1;
   keep stage region surv censor wait_surg dage stage1-stage4 urban;
run;

proc lifetest data=surv_ana method=km;
   strata wait_zero;
   time surv*censor(0);
   title1 'Survivor Curves for Wait_zero';
run;

proc lifetest data=surv_ana method=km plots=(s,ls);
   time surv*censor(0);
   title1 'Kaplan-Meier Estimates of the Survivor Function';
run;


proc lifetest data= surv_ana method=km plots=(s,lls) ;
```

```
    where stage^='';
    strata stage;
    time surv* censor(0);
    title1 'Survival Curves for Cancer Stage';
run;

proc lifetest data= surv_ana method=km plots=(s) ;
    where region^='';
    strata region;
    time surv* censor(0);
    title1 'Survival Curves for Region';
run;

proc lifetest data= surv_ana method=km plots=(s) ;
    where r_u^='';
    strata r_u;
    time surv* censor(0);
    title1 'Survival Curves for Urban and Rural';
run;

proc lifetest data= surv_ana method=km plots=(s) ;
    where income in ('U1','U2','U3','U4','U5');
    strata urban;
    time surv* censor(0);
    title1 'Survival Curves for Income within Urban';
run;

proc lifetest data= surv_ana method=km plots=(s) ;
    where income in ('R1','R2','R3','R4','R5');
    strata urban;
    time surv* censor(0);
    title1 'Survival Curves for Income within Rural';
run;

proc phreg data= surv_ana;
    model surv*censor(0)= wait_surg /ties=breslow risklimits;
    title 'Cox Regression Model with Wait_surg';
run;

proc phreg data= surv_ana;
    model surv*censor(0)= dage /ties=breslow risklimits;
    title 'Cox Regression Model with Diagnosis Age';
run;
```

```
proc phreg data= surv_ana;
   model surv*censor(0)=dage stage1-stage4 /ties=breslow risklimits ;
   title 'Cox Regression Model with Diagnosis Age and Stage';
run;

proc phreg data= surv_ana;
   model surv*censor(0)= stage1-stage4 /ties=breslow risklimits ;
   title 'Delete Diagnosis Age from Cox Regression Model';
run;

proc phreg data= surv_ana;
   model surv*censor(0)= dage /ties=breslow risklimits ;
   title 'Delete Cancer Stage from Cox Regression Model';
run;

data stage_01 stage_12 stage_23 stage_34;
   set surv_ana;
   if stage in ('Stage 0','Stage 1') then output stage_01;
   if stage in ('Stage 1','Stage 2') then output stage_12;
   if stage in ('Stage 2','Stage 3') then output stage_23;
   if stage in ('Stage 3','Stage 4') then output stage_34;
run;

*Compare stage0 and stage1;
data stage_011;
   set stage_01;
   if stage ='Stage 1' then stage0_1=1;else stage0_1=0;
run;

proc phreg data=stage_011;
   model surv*censor(0)= dage stage0_1 ;
   title 'Comparison of Stage0 and Stage1';
   label stage0_1 ='1 if stage1';
run;

*Cmpare stage1 and stage2;
data stage_121;
   set stage_12;
   if stage ='Stage 2' then stage1_2=1;else stage1_2=0;
run;

proc phreg data=stage_121;
   model surv*censor(0)= dage stage1_2 ;
   title 'Comparision of Stage1 and Stage2';
```

```
    label stage1_2='1 if stage2';
run;

*Compare stage2 and stage3;
data stage_231;
   set stage_23;
   if stage ='Stage 3' then stage2_3=1;else stage2_3=0;
run;

proc phreg data=stage_231;
   model surv*censor(0)= dage stage2_3 ;
   title 'Comparision of Stage2 and Stage3';
   label stage2_3='1 if stage3';
run;

*Compare stage3 and stage4;
data stage_341;
   set stage_34;
   if stage ='Stage 4' then stage3_4=1;else stage3_4=0;
run;

proc phreg data=stage_341;
   model surv*censor(0)= dage stage3_4 ;
   title 'Comparision of Stage3 and Stage4';
   label stage3_4='1 if stage4';
run;

*Estimation of survivor function;
proc phreg data=surv_ana;
   model surv*censor(0)= dage stage1-stage4 /ties=breslow risklimits;
   baseline out=a survival=s logsurv=ls loglogs=lls ;
   title 'Estimation of the Survivor Functions';
run;

proc print data=a;
run;

*Check goodness-of-fit for the PH assumption by weighted Schoenfeld residuals;
proc phreg data= surv_ana;
   model surv*censor(0)=dage stage1-stage4 /ties=breslow risklimits;
   output out=c WTRESSCH=schdxage schstage1 schstage2 schstage3 schstage4;
   title 'Weighted Schoenfeld Residuals';
run;
```

```
proc gplot data=c;
   symbol value=dot    interpol=none color=blue;
   plot schdxage*surv schstage1*surv schstage2*surv schstage3*surv schstage4*surv
         /vref=0 ;
   title 'Plots of Weighted Schoenfeld Residuals Versus Survival Times';
run;
```

Bibliography

Allison, P. (1995). *Survival Analysis Using the SAS® System: A Practical Guide*, SAS Institute INC., Cary, North Caroline.

Andersen, P. K. (1982). Testing Goodness of Fit of Cox's Regression Model. *Biometrics*, 38, 67-77.

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

Armitage, P. (1959). The Comparison of Survival Curves. *Journal of the Royal Statistical Society, Series A*, 122, 279-300.

Armitage, P., Berry, G. and Matthews, J. N.S. (2001). *Statistical Methods in Medical Research*, 4th ed., Blackwells, Oxford.

Armitage, P. (1981). Importance of Prognostic Factors in the Analysis of Data from Clinical Trials. *Controlled Clinical Trials*, 1, 347-353.

Armitage, P. and Gehan, E. A. (1974). Statistical Methods for the Identification and Use of Prognostic Factors. *International Journal of Cancer*, 13, 16-35.

Breslow, N. (1974). Covariance Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review*, 43, 43-54.

Breslow, N. E. (1975). Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review*, 43, 45-48.

Breslow, N. E. and Day, N.E. (1987). *Statistical Methods in Cancer Research.*Vol. 2,

The Design and Analysis of Cohort Studies, I.A.R.C., Lyon, France.

Collett, D. (2003). *Modeling Binary Data*, 2nd ed., Chapman & Hall, London.

Cox, D. R. (1970). *Analysis of Binary Data*. Methuen, London.

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

Cox, D. R. and Hinkley, D.V. (1974). *Theoretic Statistics*, Chapman & Hall, London.

Cox, D. R. and Oaks, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London.

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society, Series B*, 30, 248-275.

Cox, D. R. and Snell, E. J. (1989), *The Analysis of Binary Data*, 2nd ed., Chapman & Hall , London.

Crowley, J. and Hu, M. (1977). Covariance Analysis of Heart Transplant Survival Data. *Journal of the American Statistical Association*, 72, 27-36.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data, *Journal of the American Statistical Association*, 72, 557-565.

Elandt-Johnson, R. C. and Johnson, N. L. (1980), *Survival Models and Data Analysis*. Wiley, New York.

Gail, M. H., Lubin, J. H. and Rubinstein, L. V. (1981). Likelihood Calculations for Matched Case-Control Studies and Survival Studies with Tied Death Times. *Biometrika*, 68, 703-707.

Gill, R. and Schumacher, M. (1987). A Simple Test of the Proportional Hazards

Assumption. *Biometrika*, 74, 289-300.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional Hazards Test in Diagnostics Based on Weighted Residuals. *Biometrika*, 81, 515-526.

Harris, E. K. and Albert, A. (1991). *Survivorship Analysis for Clinical Studies*. New York: Marcel Dekker, Inc.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.

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

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

Lawless, J. F. (1982). *Statistical Methods and Model for Lifetime Data*, New York: John Wiley & Sons, Inc.

Lee, E. T. and Wang, J. W. (2003), *Statistical Methods for Survival Data Analysis*, 3rd ed., New York: John Wiley & Sons, Inc.

Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, 22, 719-748.

Mantel, N. and Myers, M. (1971). Problems of Convergence of Maximum Likelihood Iterative Procedures in Multiparameter Situation. *Journal of the American Statistical Association*, 66, 484-491.

Marubini, E. and Valsecchi, M. G. (1995). *Analyzing Survival Data from Clinical*

Trials and Observational Studies. New York: John Wiley & Sons, Inc.

Moreau, T., O'Quigley, J. and Mesbah, M. (1985). A Global Goodness-of-Fit Statistic for the Proportional Hazards Model. *Applied Statistics*, 34, 212-218.

Nelson, W. (1972), Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14, 945-966.

Pierce, M., Borges, W. H., Heyn, R., Wolfe, J. and Gilbert, E.S. (1969). Epidemiological Factors and Survival Experience in 1770 Children with Acute Leukemia. *Cancer*, 23, 1296-1304.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics*. 34, 57-67.

Prentice, R. L. and Kalbfleisch, J. D. (1979). Hazard Rate Models with Covariates. *Biometrics*, 35, 25-39.

Prentice, R. L. and Marek, P. (1979). A Quantitative Discrepancy between Censored Data Rank Tests. *Biometrica*, 35, 861-867.

Riffenburgh, R. H. and Johnstone, P. A. (2001). Survival Patterns of Cancer Patients. *Cancer*, 91(12), 2469-2475.

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

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-Based Residuals and Survival Models. *Biometrica*, 77, 147-160.

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley & Sons, Inc.

Wei, L. J. (1984). Testing Goodness of Fit for Proportional Hazards Model with

Censored Observations. *Journal of the American Statistical Association*, 79, 649-652.

Wilcoxon, F. (1945). Individual Comparison by Ranking Methods. *Biometrics*, 1, 80-83.

http://www.cancer.org/docroot/CRI/content/CRI_2_4_3X_How_is_breast_cancer_staged_5.asp.

http://icd9.chrisendres.com/

http://www.infobreastcancer.ca/treatmt.htm

http://www.phac-aspc.gc.ca/ccdpc-cpcmc/bc-cds/index_e.html

http://www.umanitoba.ca/centres/mchp/concept/concept.frame.shtml.

http://www.umanitoba.ca/centres/mchp/concept/dict/icd10ca_icd9cm.html