CHAPTER FOUR

UMMC BREAST CANCER DATA: DESCRIPTIVE AND SURVIVAL ANALYSIS

Prospective cohort studies of women with breast cancer treated in the University of Malaya Medical Centre (UMMC) Kuala Lumpur are considered. The first cohort consists of patients who are treated as breast cancer patients diagnosed from year 1993 to 1997 and are followed up until December 2002. The second cohort comprises of patients who are diagnosed from year 1998 to 2002 and are followed up until March 2006. Patients in the first cohort underwent surgery and adjuvant chemotherapy under care of general surgery in UMMC but underwent radiotherapy in Hospital Kuala Lumpur. In 1998, oncology services commenced in UMMC, with the return of two trained oncologists administering chemotherapy and radiotherapy. Therefore the initial cohorts were patients treated by general surgeon due to the absence of trained oncologists. The second cohort was those who were jointly treated by a multidisciplinary team. Therefore the cut point of year was chosen. The numbers of patients in the first cohort and the second cohorts are 423 and 965 respectively.

4.1 Description of Data

The patients' information collected consists of race, age, date of diagnosis, and pathological characteristics of tumour. The pathological characteristics considered include site, size, grade of tumour, number of positive lymph nodes, estrogen receptor, and stage of cancer. In addition, the survival times of patients and status of cancer are recorded at the end of the study. The mortality information is confirmed by referring to the record in the National Registry of Births and Deaths.

The first pathological characteristic considered in this study is site. The site is the laterality of the breast which is affected by cancer. The second characteristic is size of tumour. The pathologist grossly measures the tumour after surgery. As we can see from Figure 4.1, tumour size 2 cm in diameter is as big as a 10sen Malaysian coin while 5 cm tumour is about the size of an AA battery. The size of the tumour is also used to determine the stage of cancer.

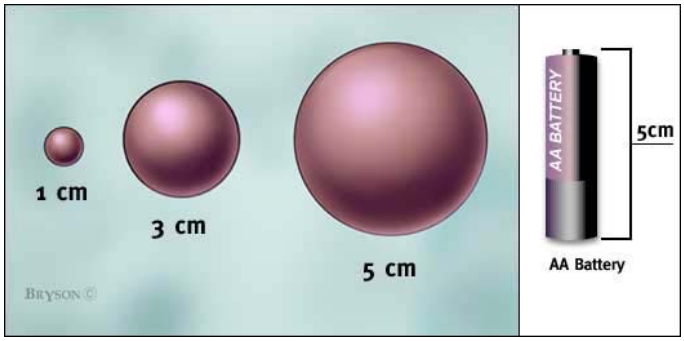


Figure 4.1
Three spheres representing the size of tumour
Source: Weiss (2000)

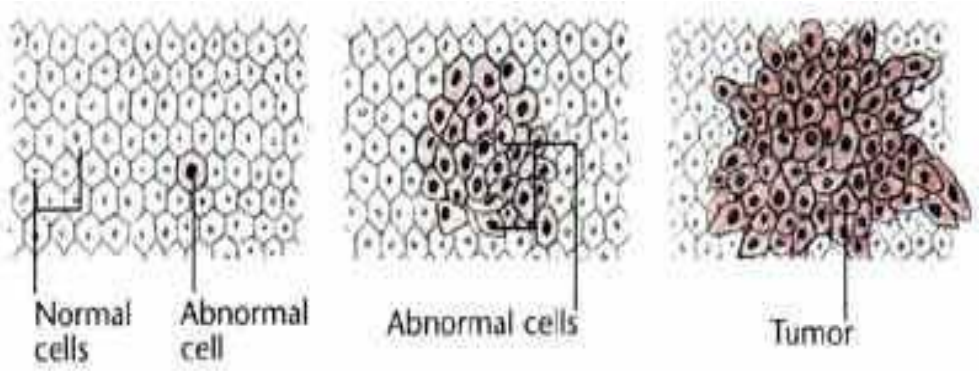
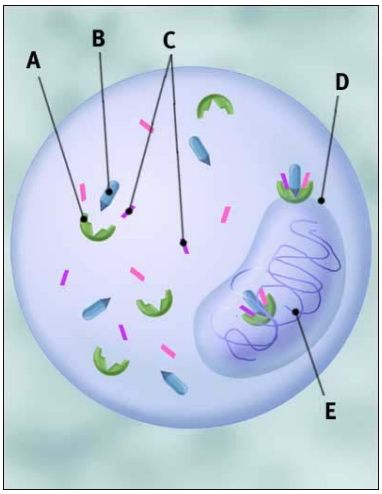


Figure 4.2
Normal Cells and Cancer Cells Structure
Source: Bodai (1999)

The third characteristic is grade of tumour denoted as *GD*. The grade is classified by using the Scraff-Bloom-Richardson system. It identifies the patterns of the cell growth by inspecting the character of the cancer cell and nucleus. To determine

the grade of tumour, pathologists will examine the breast cancer cells and their patterns under a microscope. A sample of breast cells may be taken from a breast biopsy, lumpectomy or mastectomy. The *GD* is divided into three levels; Grade one, two and three. Each level refers to the aggregate of scoring of three different components which are tubular formation (percentage of cancer composed of tubular structures), number of mitosis (rate of cell division) and nuclear pleomorphisms (change in cell size and uniformity). Each of these features is assigned a score ranging from one to three. The scores of each of the features of the cells are then added together for a final sum that ranges between three and nine. A tumour with a final sum of three, four, or five is considered a grade one tumour (well-differentiated). A sum of six or seven is considered as grade two tumour (moderately-differentiated), and a sum of eight or nine is a grade three tumour (poorly-differentiated). The detail can be found in Bodai (1999).



Cell with estrogen receptors, estrogen, and helper proteins.

A estrogen receptor
B estrogen
C estrogen helper proteins
D cell nucleus
E DNA (genetic material) inside the cell nucleus

Figure 4.3
Estrogen receptor in the cells
Source: Weiss (2000)

The fourth characteristic is estrogen receptor denoted by *ER*. The receptor for the female hormone estrogen in the cells is shown in Figure 4.3. The receptors are the

eyes and ears of the breast cells, getting messages sent by the hormones and figuring out what to do with these messages. The hormones will tell the receptors to stimulate or "turn on" breast cell growth. Estrogen can increase the normal and abnormal breast cell growth. There are two outcomes on testing the breast cell sensitivity to estrogen hormone. Firstly, the positive estrogen-receptor is due to the cells which are more likely to grow in a high-estrogen environment. Secondly, negative estrogen-receptor refers to cells which are usually not affected by the levels of estrogen in the body. If we have a positive estrogen receptor, then cancers are more likely to respond to anti-estrogen therapies.

The fifth characteristic is the number of positive lymph nodes denoted as *LN*. The *LN* refers to the numbers of armpit or axillary lymph nodes that harbor cancer cells after removal by surgery. The lymph nodes are filtered along the lymphatic system as in Figure 4.4. Their function is to filter out and trap bacteria, viruses, cancer cells, and other unwanted substances. This is to ensure that they are safely eliminated from the body.

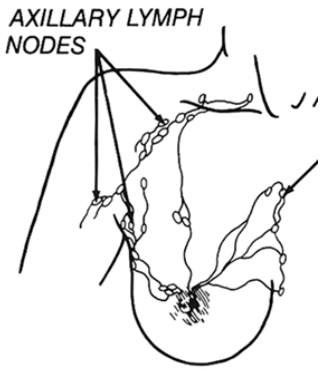


Figure 4.4
The lymph nodes near breast area
Source: Bodai (1999)

The last characteristic is the stage of cancer denoted as *stg*. The *stg* is based on several factors; the size of tumour, invasive or non-invasive type of cancer, the number of cancerous lymph nodes, and whether the cancer has spread beyond the breast area. The staging of the breast cancer may provide some guidance for appropriate treatment regimen for patients. *stg* is separated into five levels as follows:

1. The stage 0 is used to describe non-invasive breast cancers, such as ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS). In stage 0 there is no evidence of cancer cells or non-cancerous abnormal cells breaking out of the part of the breast in which they started or of getting through to or invading neighboring normal tissue.

2. Stage I describes invasive breast cancer. In stage I the cancer cells have broken through or invaded neighboring normal tissue in which the tumour measures up to 2 cm and no lymph nodes are involved.

3. Stage II describes invasive breast cancer and it divided into two levels. Stage IIA and stage IIB.
    a. Stage IIA can be described as no tumour can be found in the breast but cancer cells are found in the axillary lymph nodes which are the lymph nodes under the arm, or the tumour measures 2 cm or less and has spread to the axillary lymph nodes, or the tumour is larger than 2 cm but not larger than 5 cm and has not spread to the axillary lymph nodes.
    b. Stage IIB can be describes as the tumour is larger than 2 cm but no larger than 5 cm and has spread to the axillary lymph nodes, or the tumour is larger than 5 cm but has not spread to the axillary lymph nodes.

4. Stage III describes invasive breast cancer and is divided into three levels. Stage IIIA, stage IIIB, and stage IIIC.
    a. Stage IIIA can be described as no tumour is found in the breast. Cancer is found in axillary lymph nodes that are clumped together or sticking to other structures, or cancer may have spread to lymph nodes near the breastbone, or the tumour is 5 cm or smaller and has spread to axillary lymph nodes that are clumped together or sticking to other structures, or the tumour is larger than 5 cm and has spread to axillary lymph nodes that are clumped together or sticking to other structures.
    b. Stage IIIB can be described as tumour of any size that has spread to the chest wall and/or skin of the breast and may have spread to axillary lymph nodes that are clumped together or sticking to other structures or cancer may have spread to lymph nodes near the breastbone.
    c. Stage IIIC can be describe as there may be no sign of cancer in the breast or, if there is a tumour, it may be any size and may have spread to the chest wall and/or the skin of the breast, and the cancer has spread to lymph nodes above or below the collarbone and the cancer may have spread to axillary lymph nodes or to lymph nodes near the breastbone.

5. Stage IV can be described as the cancer has spread to other organs of the body. It is usually spread to lungs, liver, bone, or brain.

The stage can also be categorized as early stage and advanced stage. Although these terms are not medically precise, they had been used in medical literature. The early stage consists of stage 0, stage I, stage II, and some levels of stage III, while

advanced stage consists of the other stages as described above. In this study, the AJCC 5^{th} edition was used (Fleming *et. al* (1997)).

4.1.1 Data Summary

Summary of the prognostic factors are given in Table 4.1. The *race* of a patient is divided into three levels; level zero is for Chinese, level one is for Indian and level two is for Malay. The other races are not included in the study as the number of patients is too small. The *age* is categorized into three levels; ≤ 40 years as level zero, 40-59 years as level one, and $\geq$ 60 years as level two.

The *size* is divided into three levels. Tumour size ≤ 2 cm is for level zero, size between 2 cm and 5 cm is for level two and lastly, size ≥ 5 cm is for level three. On the other hand, the *site* is labelled according to either the left or right breast or both. Further, the *GD* is separated into three levels. Level zero is for unavailable cancer grade, level one is for cancer grade one and grade two, whilst level two is for serious cancer grade three.

The *ER* is divided into three levels. Level zero is for unavailable status of estrogen receptor. Level one is estrogen-receptor positive and level two is for estrogen-receptor negative. The *LN* are stratified into four levels where the level zero is for unavailable status, level one is for status zero, level two combine the status one, two and three together, while level three represents status more than three. Unavailable status means that surgery was not performed in late stage disease patients while status zero means that examination of all the exceeded lymph nodes did not show any cancer cells. The *stg* is computed according to the American Joint Committee on cancer

system, which divides the factor into four levels; level zero is for stage I, level one is for stage II, level two is for stage III and level three is for stage IV. Stage 0 is not included in this study.

Two other important information recorded are the survival times measured in months and the status of patients. The survival times of patients take the number of months in which the individual enters the study until the date on which the individual dies or was last known to be alive. The date of death is confirmed with records obtained from the National Registration Department Malaysia. Patients who are still alive at the end of study or die because of non-breast cancer death are given status zero, while patients who die because of breast cancer are given status one.

Table 4.1
Description of data

Variables		**Level**
race	Chinese	*race0*
	Indian	*race1*
	Malay	*race2*
age	$\leq$ 40 years	*age0*
	40-59 years	*age1*
	$\geq$ 60 years	*age2*
stg	Stage I	*stg0*
	Stage II	*stg1*
	Stage III	*stg2*
	Stage IV	*stg3*
LN	Undetected Status	*LN0*
	0	*LN1*
	1 to 3	*LN2*
	More than 3	*LN3*
GD	Undetected Status	*GD0*
	1 and 2	*GD1*
	3	*GD2*
ER	Undetected Status	*ER0*
	Positive	*ER1*
	Negative	*ER2*
size	$\leq$ 2 cm	*size0*
	2 cm to 5 cm	*size1*
	$\geq$ 5 cm	*size2*
site	Left or Right	*site0*
	Left and Right	*site1*

4.2 Survival Analysis

Table 4.2 gives the number of patients for every prognostic factor of the first and second cohorts. For both cohorts, Chinese has the highest number of patients who are diagnosed with breast cancer, followed by the Malay and then the Indian. It shows that Chinese women are more susceptible to breast cancer as indicated in the National Cancer Registry (NCR) report. Next, it is found that the majority of breast cancer

Table 4.2
Number of patients with breast cancer in two cohorts

Variables			First Cohort		Second Cohort	
			Frequency	Percentage	frequency	Percentage
race	*race0*	Chinese	264	62.4	611	63.32
	race1	Indian	69	16.3	120	12.44
	race2	Malay	90	21.3	234	24.25
age	*age0*	$\leq$ 40 years	83	19.6	170	17.6
	age1	40-59 years	245	57.9	603	62.5
	age2	$\geq$ 60 years	95	22.5	192	19.9
stg	*stg0*	Stage I	73	17.3	207	21.5
	stg1	Stage II	206	48.7	471	48.8
	stg2	Stage III	74	17.5	171	17.7
	stg3	Stage IV	70	16.5	116	12.0
LN	*LN0*	Undetected Status	86	20.3	196	20.3
	LN1	0	165	39.0	395	40.9
	LN2	1 to 3	97	22.9	207	21.5
	LN3	More than 3	75	17.7	167	17.3
GD	*GD0*	Undetected Status	185	43.7	280	29.0
	GD1	1 and 2	161	38.1	440	45.6
	GD2	3	77	18.2	245	25.4
ER	*ER0*	Undetected Status	256	60.5	147	15.2
	ER1	Positive	78	18.4	467	48.4
	ER2	Negative	89	21.0	351	36.4
size	*size0*	$\leq$ 2 cm	49	11.6	285	29.5
	size1	2 cm to 5 cm	207	48.9	429	44.5
	size2	$\geq$ 5 cm	167	39.5	251	26.0
site	*site0*	Left or Right	413	97.6	945	97.9
	site1	Left and Right	10	2.4	20	2.1

patients are between age 41 and 59 years, followed by age more than 60 and less than 40 years old. That is, patients between 41 to 59 years old are more prone to cancer compared to patients in the other two age groups. This phenomena is also seen in the NCR report. Besides, the majority of breast cancer patients who seek treatment at UMMC are diagnosed as having early stages, which are stage I and stage II for both cohorts. Stage II is the commonest stage. Meantime, there are also patients who are diagnosed to have advanced stage but the number is smaller.

In addition, in both cohorts it can be seen that 20.3% of the breast cancer patients do not undergo any surgery giving the undetected status of *LN*. For patients who underwent surgery, the cancer did not spread to the lymph node in 56% of them.

It can also be seen that 43.7% of patients in the first cohort had unavailable grade of tumour but it decreased in the second cohort to 29%. This suggests that the pathology evaluation on grade of tumour had become part of a routine report. Both cohorts show that most of the patients had low and moderate nature growth of the cancer cell which is 38.1% in the first cohort and 45.6% in the second cohort.

Further, the majority of patients in the first cohort had unavailable estrogen receptor information. Only 21% of the patients had complete information of ER status. In contrast, for the second cohort, the majority of patients have complete estrogen receptor information. Around 48.4% of patients had a positive estrogen receptor while 36.4% of patients had a negative estrogen receptor. Data on estrogen receptor of patients were unavailable for 15.2% of the patients, where these patients did not undergo surgery.

Meanwhile, higher percentage of patients was found to have 2 cm to 5 cm size of tumour in the first cohort and the second cohort; 48.9% and 44.5%, respectively. However, more patients were diagnosed with small sized tumour in the second cohort compared to first cohort. This might be due to better awareness among patients to go for early check-up.

Finally, the majority of patients in the first cohort and the second cohort had unilateral breast cancer. Only 2.4% and 2.1% of breast cancer patients in the first cohort and the second cohort, respectively have tumour in both sides of their breast.

4.2.1 Survival Probability of Breast Cancer Patients

Figure 4.5 gives the plot of overall survival probability for patients in the first and second cohorts. The log-rank tests confirm that the survival of these two cohorts are significantly different (p-value = 0). By comparing the five-year survival probabilities of both cohorts, we can conclude that patients in the second cohort had a much chance of survival compared to the first cohort as given in Table 4.3. Note that the 95% confidence interval (C.I.) of the survival probabilities for both cohorts does not overlap.

Table 4.3
Five-year probability of overall survival

Prospective Cohort Studies	**S(60)**	**95% C.I.**
First Cohort	0.584	(0.538, 0.634)
Second Cohort	0.757	(0.729, 0.786)

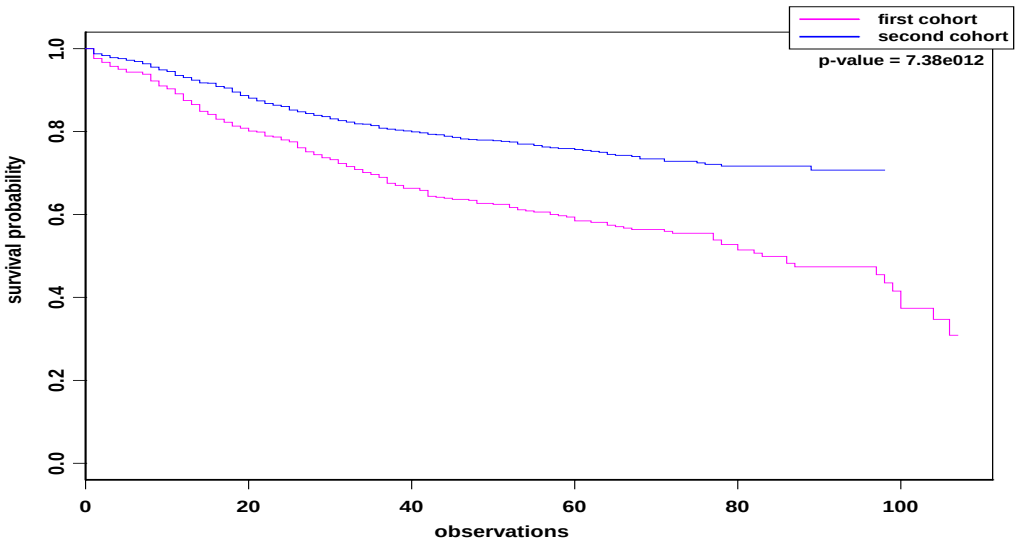
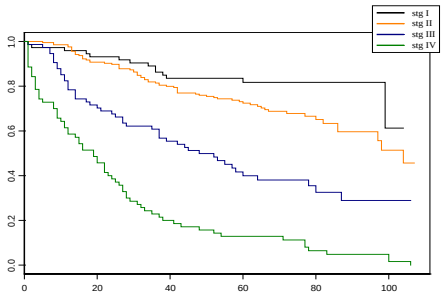


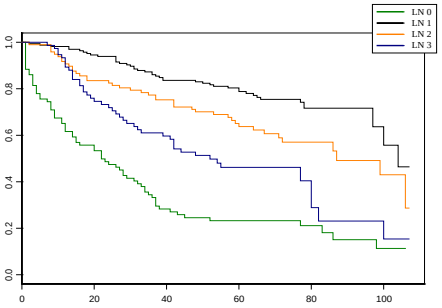
Figure 4.5
Kaplan-Meier plot of overall survival probability for each cohort

Now, we look at the survival probabilities for every prognostic factor. Figure 4.6 gives the Kaplan-Meier plots of each prognostic factor for individuals in the first cohort of breast cancer patients. The survival of patients in different levels of the prognostic factors *stg*, *LN*, *GD*, *site* and *race* were significantly different. Their respective Kaplan-Meier curves do not cross each other and the p-values of the log-rank tests were less than 0.05. While, the two other prognostic factors; *ER*, and *age* gave insignificant results and the Kaplan-Meier curves do cross. Since the KM plot for *size* is crossing, we then consider the Peto-Wilcoxon test which does not receive the PHM condition. The result shows that the factor is positive different is significant. For the second cohort, the survival experience of patients in different levels of all prognostic factors was significantly different. Figure 4.7 gives the Kaplan-Meier plots of each prognostic factor. None of the curves cross each other.

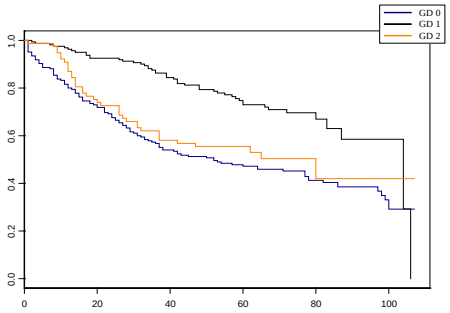
a) *stg* ($\chi_3^2 = 194$, p-value=0)



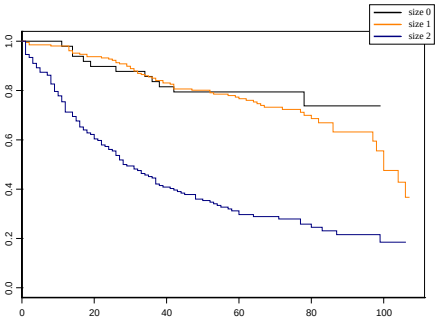
b) *LN* ($\chi_3^2 = 112$, p-value=0)



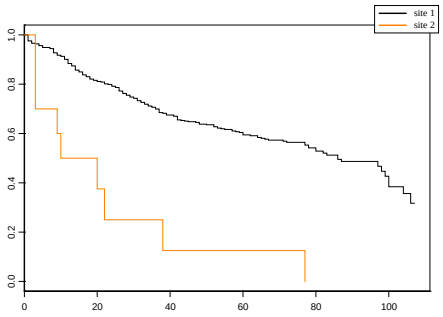
c) *GD* ($\chi_2^2 = 26.5$, p-value=0)



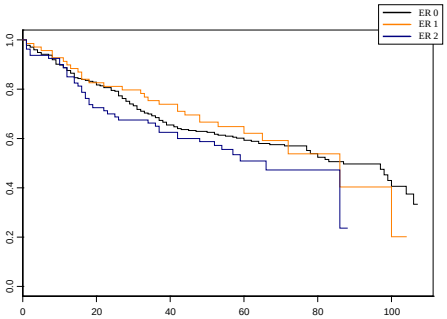
d) *size* ($\chi_2^2 = 104$, p-value=0)



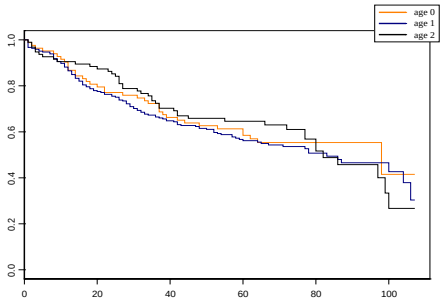
e) *site*($\chi_1^2 = 25.4$, p-value=0)



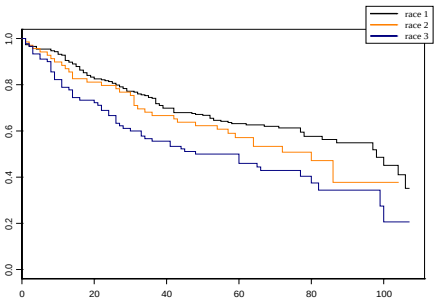
f) *ER*($\chi_2^2 = 2.1$, p-value=0.342)



g) *age* ($\chi_2^2 = 0.5$, p-value=0.76)

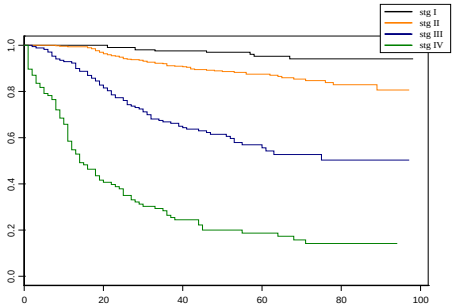


h) *race* ($\chi_2^2 = 12$, p-value=0.0025)

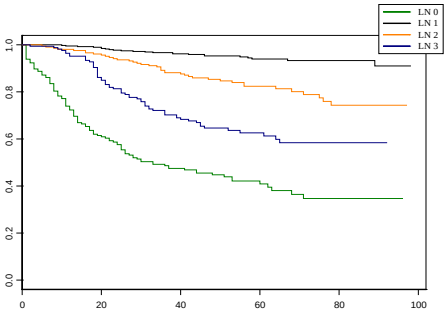


Figures 4.6
Kaplan-Meier plot of variables for first cohort
y-axis is survival probability, x-axis is observation

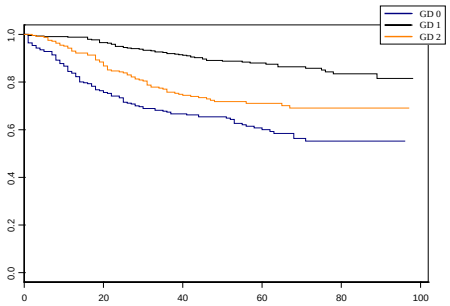
a) *stg* ($\chi_3^2 = 531$, p-value=0)



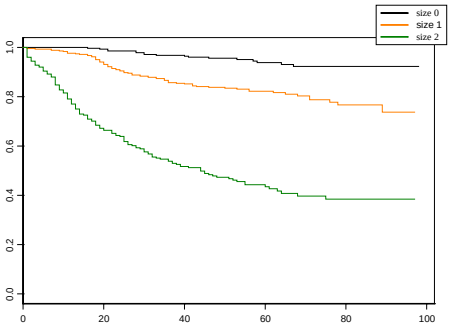
b) *LN* ($\chi_3^2 = 282$, p-value=0)



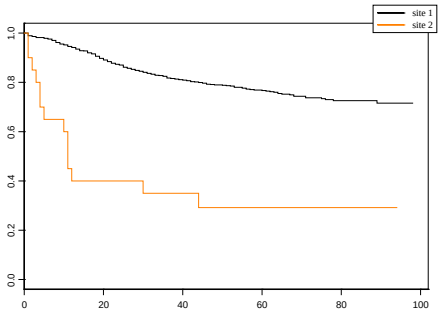
c) *GD* ($\chi_2^2 = 75.7$, p-value=0)



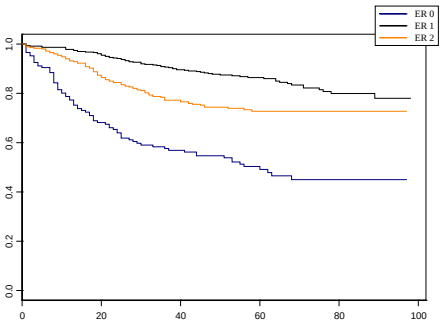
d) *size* ($\chi_2^2 = 243$, p-value=0)



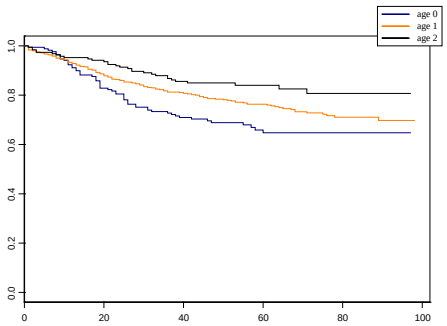
e) *site* ($\chi_1^2 = 50.4$, p-value=0)



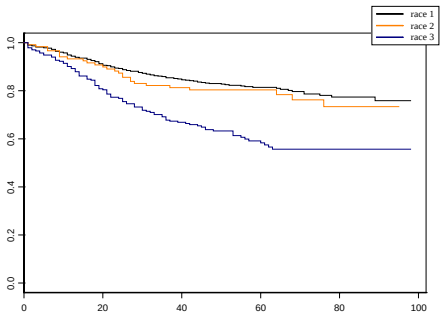
f) *ER*($\chi_2^2 = 95.2$, p-value=0)



g) *age* ($\chi_2^2 = 12.5$, p-value=0.0020)



g) *race* ($\chi_2^2 = 42.7$, p-value=0)



Figures 4.7
Kaplan-Meier plot of variables for second cohort
y-axis is survival probability, x-axis is observation

Table 4.4 gives the five-year probabilities of breast cancer patients for each prognostic factor in both cohorts. The five-year probabilities of survival showed that stage I has a better chance of survival compared to other stages, while patients diagnosed with stage IV had significantly worse probability of survival compared to the other groups. We observed that the early stage of cancer had a better survival compared to the advanced stage of cancer. Therefore, it is important that breast cancer patients come early for check-up upon noticing any symptom of breast cancer.

In addition, we observe that patients who do not undergo the surgery (*LN0*) had the worst probability to survive compared to patients who underwent surgery (*LN1*, *LN2*, and *LN3*). Meanwhile, patients who were diagnosed with no lymph node affected by the tumour (*LN1*) had a better chance of survival than those whose lymph nodes are affected by the tumour (*LN2* and *LN3*). This shows that the spread of cancer cells to the lymph nodes decreases the survival probability of patients with breast cancer. Patients who did not have the axillary surgery are mainly patients with stage IV disease, when treatment is mainly with chemotherapy and not surgery. Further, we find that patients with positive estrogen receptor had better survival probability compared to the negative and undetected status estrogen receptor groups.

In general, for the grade of tumour, better survival was observed in the second cohort then in the first cohort. Patients in the lower and moderate grades of tumour had better survival compared to the rest. That is, the cancer cell is not as aggressive as that in grade three. In addition, the five-year probabilities of prognostic factor *size* were close for size less than 2 cm and size between 2 and 5 cm and was lower for size of tumour more than 5 cm. At the same time, the five-year probabilities of *site* indicate that the chance of survival was better when only one side of breast was affected. As for

Table 4.4
Five-year probability survival for both cohorts

Prognostic Factor		First Cohort S(60)	First Cohort 95% C.I	Second Cohort S(60)	Second Cohort 95% C.I
stg (stage of cancer)	*stg0* (stage I)	0.817	(0.731, 0.913)	0.952	(0.919, 0.986)
	stg1 (stage II)	0.664	(0.644, 0.789)	0.875	(0.844, 0.907)
	stg2 (stage III)	0.399	(0.299, 0.534)	0.556	(0.480, 0.644)
	stg3 (stage IV)	0.129	(0.070, 0.237)	0.187	(0.124, 0.281)
LN (Number of positive Lymph Node)	*LN 0* (undetected status)	0.232	(0.157, 0.344)	0.409	(0.339, 0.492)
	LN 1 (0 of *LN*)	0.788	(0.727, 0.854)	0.940	(0.910, 0.966)
	LN 2 (1 - 3 of *LN*)	0.637	(0.546, 0.744)	0.824	(0.770, 0.881)
	LN 3 (> 3 of *LN*)	0.462	(0.359, 0.595)	0.626	(0.553, 0.708)
ER (estrogen receptor)	*ER 0* (undetected status)	0.593	(0.537, 0.655)	0.491	(0.413, 0.584)
	ER 1 (positive *ER*)	0.621	(0.511, 0.755)	0.864	(0.832, 0.897)
	ER 2 (negative *ER*)	0.509	(0.402, 0.643)	0.728	(0.680, 0.779)
GD (grade of tumour)	*GD 0* (undetected status)	0.472	(0.405, 0.550)	0.600	(0.541, 0.666)
	GD 1 (grade 1 and 2)	0.730	(0.662, 0.805)	0.880	(0.848, 0.913)
	GD 2 (grade 3)	*0.554*	(0.453, 0.678)	0.711	(0.654, 0.773)

Table 4.4 (continued)

Prognostic Factor		First Cohort S(60)	First Cohort 95% C.I	Second Cohort S(60)	Second Cohort 95% C.I
size (size of tumour)	*size0* ($\leq$ 2 cm)	0.794	(0.689, 0.917)	0.938	(0.908, 0.970)
	size1 (2 cm – 5 cm)	0.767	(0.710, 0.828)	0.822	(0.785, 0.861)
	size2 ($\geq$ 5 cm)	0.297	(0.233, 0.377)	0.435	(0.374, 0.506)
site (breast affected)	*site0* (left or right)	0.594	(0.548, 0.645)	0.767	(0.739, 0.796)
	site1 (left and right)	0.125	(0.021, 0.762)	0.292	(0.145, 0.585)
race (race of patients)	*race0* (Chinese)	0.632	(0.575, 0.694)	0.814	(0.782, 0.847)
	race1 (Indian)	0.571	(0.464, 0.704)	0.804	(0.735, 0.879)
	race2 (Malay)	0.459	(0.366, 0.577)	0.583	(0.518, 0.656)
age (age of patients)	*age0* ($\leq$ 40 years)	0.585	(0.487, 0.703)	0.647	(0.575, 0.729)
	age1 (40 – 60 years)	0.561	(0.501, 0.629)	0.763	(0.729, 0.800)
	age2 ($\geq$ 60 years)	0.646	(0.555, 0.751)	0.840	(0.787, 0.897)

race and *age*, Malay patients have lower chance of survival compared to other races (Chinese and Indian) while those in age group more than 60 years had a better chance of survival in five-year probability.

4.3 Modeling the Local Breast Cancer Data

In section 4.2, descriptive analysis on local breast cancer data has been carried out. In this section we are interested into identifying the prognostic factors that are important to the survival of the breast cancer patients using Cox PHM. Firstly, we investigate whether the proportional hazard assumption is satisfied for each factor. We then employ the stepwise method to determine the 'best' Cox PHM.

4.3.1 Proportional Hazard Assumption

Firstly we investigate the assumption of proportional hazard model for each variable. It can be seen that the proportional hazard assumption is satisfied for *stg*, *LN*, *GD*, *site* and *race* factors in the first cohort as shown in Figure 4.6 (see Section 4.2.1) as the curves of the corresponding plots do not cross each other. However for *size*, *ER* and *age* factors, the Kaplan-Meier curves are quite close to each others. Thus, for these factors we employ the goodness-of-fit test to investigate the assumption.

Table 4.5 gives the p-value of Cox's time dependant covariate test for *size*, *ER* and *age* factors. The test is suggested by *Brelow et al.* (1984) for assessment of the PHA by testing the interaction term in the model. Let the probabilities are not weighted with ranks of time as *Null* hypothesis. The result shows that *p*-values for factors *ER* (*0.2* and *0.72*), *size* (*0.24* and *0.52*) and *age* (*0.92* and *0.22*) are greater than 0.05

therefore the coefficient β in the model is constant and as conclusion the proportional hazard assumption is met for ER, size and age factors.

Table 4.5
Cox's time dependant covariate test for ER, size and age factors in the first cohort

Prognostic Factors	***ER***		***size***		***age***	
	ER1	***ER2***	***size1***	***size2***	***age1***	***age2***
p-values	0.20	0.72	0.24	0.52	0.92	0.22

On the other hand Figure 4.7 (see in Section 4.2.1) gives the Kaplan-Meier plots for the second cohort. All the factors clearly satisfied the assumption as the Kaplan-Meier curves do not cross each other. Thus, it is reasonable to model the breast cancer data of both cohorts with a proportional hazard model, in particular, the Cox PHM.

4.3.2 Modelling

The analysis on breast cancer data is investigated further to identify significant prognostic factors through survival regression modeling. The baseline levels chosen are *stg*0 (stage I), *LN*1 (no cancer in lymph nodes), *GD*0 (undetected status), *ER*0 (undetected status), *size*0 ($\leq$ 2 cm), *site*0 (left or right breast area), *age*0 ($\leq$ 40 years), and *race*0 (Chinese).

Tables 4.6 and 4.7 provide the step-by-step search to find the 'best' *Cox* PHM for the first and second cohorts of breast cancer data sets respectively. In Step 1 of the model selection for the first cohort, we found that *stg* should included first in the model with a reduction values of $-2\log\hat{L}$ = 138.456 and *p-value* = 0. The next step is to fit model by including a factor at a time with *stg* factor remains in the model. As shown in

Table 4.6
Variables selection on first cohort

	Model (d.f.)	-2 Log L	Reduction	p-value
	Null	2217.474	-	-
Step 1	***stg* (3)**	2079.018	**138.456**	**0**
	LN (3)	2124.16	93.314	0
	ER (2)	2215.44	2.034	0.362
	GD (2)	2189.508	27.966	0.000
	size (2)	2124.38	93.094	0
	site (1)	2203.732	13.742	0.000
	race (2)	2206.388	11.086	0.004
	age (2)	2216.926	0.548	0.760
Step 2	***stg + LN* (3)**	2056.002	**23.016**	**0.000**
	stg + ER (2)	2076.428	2.590	0.274
	stg + GD (2)	2067.848	11.170	0.004
	stg + size (2)	2058.964	20.054	0.000
	stg + site (1)	2078.376	0.642	0.423
	stg + race (2)	2072.316	6.702	0.035
	stg + age (2)	2078.106	0.912	0.634
Step 3	*stg + LN + ER* (2)	2053.514	2.488	0.2882
	stg + LN + GD (2)	2047.112	8.890	0.0117
	***stg + LN + size* (2)**	2045.43	**10.572**	**0.0051**
	stg + LN + site (1)	2056.00	0.002	0.9643
	stg + LN + race (2)	2049.534	6.468	0.0394
	stg + LN + age (2)	2054.308	1.694	0.4287
Step 4	*stg + LN + size + ER* (2)	2042.742	2.688	0.2608
	***stg + LN + size + GD* (2)**	2038.144	**7.286**	**0.0262**
	stg + LN + size + site (1)	2045.414	0.016	0.8993
	stg + LN + size + race (2)	2040.276	5.154	0.0760
	stg + LN + size + age (2)	2044.284	1.146	0.5638
Step 5	*stg + LN + size + GD + ER* (2)	2035.972	2.172	0.3376
	stg + LN + size + GD + site (1)	2038.108	0.036	0.8495
	stg + LN + size + GD + race (2)	2033.902	4.242	0.1199
	stg + LN + size + GD + age (2)	2036.62	1.524	0.4667

Table 4.7
Variables selection in second cohort

	Model (d.f.)	-2 Log L	Reduction	p-value
	Null	3093.512	-	-
Step 1	***stg* (3)**	2769.496	**324.016**	**0**
	LN (3)	2855.146	238.366	0
	ER (2)	3016.294	77.218	0
	GD (2)	3018.936	74.576	0.000
	size (2)	2890.472	203.040	0
	site (1)	3068.4	25.112	0.000
	race (2)	3056.272	37.240	0.000
	age (2)	3081.048	12.464	0.002
Step 2	***stg* + *LN* (3)**	2726.094	**43.402**	**0.0000**
	stg + *ER* (2)	2749.398	20.098	0.0000
	stg + *GD* (2)	2748.754	20.742	0.0000
	stg + *size* (2)	2766.718	2.778	0.249
	stg + *site* (1)	2765.928	3.568	0.059
	stg + *race* (2)	2753.628	15.868	0.0004
	stg + *age* (2)	2764.068	5.428	0.066
Step 3	***stg* + *LN* + *ER* (2)**	2704.37	**21.724**	**0.0000**
	stg + *LN* + *GD* (2)	2710.79	15.304	0.0005
	stg + *LN* + *size* (2)	2721.962	4.132	0.1267
	stg + *LN* + *site* (1)	2723.992	2.102	0.1471
	stg + *LN* + *race* (2)	2716.018	10.076	0.0065
	stg + *LN* + *age* (2)	2721.202	4.892	0.0866
Step 4	*stg* + *LN* + *ER* + *GD* (2)	2698.59	5.780	0.0556
	stg + *LN* + *ER* + size (2)	2701.052	3.318	0.1903
	stg + *LN* + *ER* + site (1)	2702.28	2.090	0.1483
	***stg* + *LN* + *ER* + race (2)**	2697.216	**7.154**	**0.0280**
	stg + *LN* + *ER* + *age* (2)	2698.298	6.072	0.0480
Step 5	*stg* + *LN* + *ER* + *race* + *GD* (2)	2691.998	5.218	0.0736
	stg + *LN* + *ER* + *race* + *size* (2)	2693.784	3.432	0.1798
	stg + *LN* + *ER* + *race* + *site* (1)	2695.34	1.876	0.1708
	stg + *LN* + *ER* + *race* + *age* (2)	2691.64	5.576	0.0615

Step 2, the reduction value of $-2\log\hat{L}$ by including *LN* factor in the model is the largest compared to the others ($-2\log\hat{L}$ = 23.016 with *p-value* = 0). In Step 3 and 4 similar rule is followed and the important factors identified are *size* and *GD* respectively. The model selection terminate at Step 5 when there is insignificant reductions of $-2\log\hat{L}$ observed.

Similar steps are applied to the second cohort data set. It is *stg* and *LN* are again identified to be important in Step 1 and Step 2 respectively. In Step 3, *ER* factor have the largest reduction of $-2\log\hat{L}$ compared to *GD*, *size*, *site*, *race* and *age*. The next important factor is *race*, where the reduction of $-2\log\hat{L}$ is 7.154 with *p-value* = 0.0280. Therefore the 'best' fitted models for both cohorts are found to be as follows:

First Cohort:

$$h_i(t) = \exp(\hat{\beta}_1 x_{stgII,i} + \hat{\beta}_2 x_{stgIII,i} + \hat{\beta}_3 x_{stgIV,i} + \hat{\beta}_4 x_{LN0,i} + \hat{\beta}_5 x_{LN2,i} + \hat{\beta}_6 x_{LN3,i} + \hat{\beta}_7 x_{size1,i} + \hat{\beta}_8 x_{size2,i} + \hat{\beta}_9 x_{GD1,i} + \hat{\beta}_{10} x_{GD2,i})\, h_0(t)$$

Second Cohort:

$$h_i(t) = \exp(\hat{\beta}_1 x_{stgII,i} + \hat{\beta}_2 x_{stgIII,i} + \hat{\beta}_3 x_{stgIV,i} + \hat{\beta}_4 x_{LN0,i} + \hat{\beta}_5 x_{LN2,i} + \hat{\beta}_6 x_{LN3,i} + \hat{\beta}_7 x_{ER1,i} + \hat{\beta}_8 x_{ER2,i} + \hat{\beta}_9 x_{race1,i} + \hat{\beta}_{10} x_{race2,i})\, h_0(t)$$

4.3.3 Cox Proportional Hazards Model Analysis

Table 4.8 gives the estimates of parameter and hazard ratio with the standard errors of Cox PHM for both cohorts of local breast cancer data. Two common significant factors found in both cohorts are the *stg* and *LN* factors. Generally, the hazard is greater for the advanced stage of cancer in both cohorts. Also, the values of hazard ratio are greater in the second cohort. For example, the chance of survival for patients in stage I is 13

Table 4.8

Cox PHM result for both cohorts respect to significance prognostic factors

Prognostic Factors	First Cohort				Second Cohort			
	Estimate	**S.E. of estimate**	**Hazard**	**C.I of hazard**	**Estimate**	**S.E. of estimate**	**Hazard**	**C.I. of hazard**
***stg* (*cancer stage*)**								
*stg*1 (stage II)	0.2822	0.325	1.326	(0.702, 2.51)	0.592	0.384	1.805	(0.851, 3.84)
*stg*2 (stage III)	0.7367	0.378	2.089	(0.996, 4.38)	1.625	0.392	5.079	(2.355, 10.95)
*stg*3 (stage IV)	1.5022	0.374	4.492	(2.159, 9.35)	2.614	0.398	13.656	(6.257, 29.81)
***LN* (*number of lymph node*)**								
*LN*0 (undetected status)	0.8316	0.246	2.297	(1.419, 3.72)	1.679	0.286	5.361	(3.061, 9.39)
*LN*2 (1-3 of *LN*)	0.1885	0.236	1.207	(0.760, 1.92)	0.744	0.284	2.103	(1.205, 3.67)
*LN*3 (>3 of *LN*)	0.4971	0.244	1.644	(1.020, 2.65)	1.183	0.274	3.264	(1.908, 5.58)
***size* (*tumour size*)**								
*size*1 (2 cm to 5 cm)	0.0518	0.347	1.053	(0.533, 2.08)	-	-	-	-
*size*2 (≥5 cm)	0.6389	0.367	1.894	(0.923, 3.89)				
***GD* (*tumour grade*)**								
*GD*1 (Grade 1 & 2)	-0.2455	0.194	0.782	(0.534, 1.15)	-	-	-	-
*GD*2 (Grade 3)	0.3647	0.217	1.440	(0.942, 2.20)				
***ER* (*estrogen receptor*)**								
*ER*1 (positive)	-	-	-	-	-0.146	0.204	0.864	(0.578, 1.29)
*ER*2 (negative)					0.533	0.210	1.739	(1.152, 2.63)
race (race of patients)								
*race*1 (Indian)	-	-	-	-	0.101	0.222	1.107	(0.716, 1.71)
*race*2 (Malay)					0.382	0.142	1.466	(1.109, 1.94)

times higher than those in stage IV in the second cohort, but only IV times higher in the first cohort. This is true since more patients come forward with early stage of cancer in the second cohort enabling better chance of survival from the breast cancer.

For the *LN* factor, the undetected status as listed in Table 4.8 refers to patients who do not undergo the operation due to the advanced stage of cancer. It is not surprising to have the hazard of patients in this stage to be five times greater than patients who diagnosed with no lymph nodes affected by cancer after undergoing surgery (LN1).

The other two significant prognostic factors identified in the modeling are different for both cohorts; *size* and *GD* for the first cohort, *ER* and *race* for the second cohort. Results indicate that, in the first cohort, patients with large tumour size (≥5 cm) have almost twice higher risk of death compared to those with small tumour size (2 cm to 5 cm), while patients diagnosed with grade one and two of breast cancer have a better chance of surviving than patients with undetected status and grade 3 of breast cancer.

As for the second cohort, the estrogen receptor (*ER*) in the body plays an important role in determining the survival of patients. Patients who are *ER* positive have better survival then those with undetected status and *ER* negative, while the hazard of patients from different races only differ slightly, though the Chinese has a better chance of survival compared to the others.

4.4 Summary

The analysis indicates that patients in the second cohort are more aware of the breast

cancer issues because more patients were diagnosed at an earlier stage. Variable *stg* and *LN* were found to be significant prognostic factors for both cohorts, whereas *size* and GD are important for the first cohort and ER and race for the second cohort. In the literature, the two factors *stg* and *LN* are known to be important in determining the survival of breast cancer patients (see Haybittle et al. (1982) and Aryandono et al. (2006)).