Micrometastatic Node-Positive Breast Cancer: An Analysis of Survival Outcomes and Prognostic Impact of the Number of Positive Nodes and the Ratio of Positive to Excised Nodes in Comparison to Node-Negative and Macrometastatic Node-Positive Breast Cancer

By

KAREN HUI LI

B.Sc., Brock University, 2005

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

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Supervisory Committee

Micrometastatic Node-Positive Breast Cancer: An Analysis of Survival Outcomes and Prognostic Impact of the Number of Positive Nodes and the Ratio of Positive to Excised Nodes in Comparison to Node-Negative and Macrometastatic Node-Positive Breast Cancer

By

KAREN HUI LI

B.Sc., Brock University, 2005

Supervisory Committee

Dr. Mary Lesperance, (Department of Mathematics and Statistics)
Supervisor

Dr. Laura Cowen, (Department of Mathematics and Statistics)
Department Member

Dr. Farouk Nathoo, (Department of Mathematics and Statistics)
Department Member

Supervisory Committee

Dr. Mary Lesperance, (Department of Mathematics and Statistics)
Supervisor

Dr. Laura Cowen, (Department of Mathematics and Statistics)
Department Member

Dr. Farouk Nathoo, (Department of Mathematics and Statistics)
Department Member

Abstract

In this study, we examined survival for patients with micrometastases greater than 0.2mm but less than 2mm (pN1a) in comparison to node-negative (pN0) and macrometastatic node-positive (pN1b) patients. Data for patients diagnosed from 1988 to 1998 with TNM pathological T1-2 stage, pN0, and pN1a-b breast cancer with no distant metastasis was provided by Dr. P. Truong from BC Cancer Agency. Results obtained from the Kaplan-Meier estimators and the multivariable Cox Proportional Hazards Model analyses suggested that micrometastatic node-positive patients had worse survival than the node-negative patients, but better survival in comparison to the macrometastatic node-positive patients. Increasing number of positive nodes and larger values of the ratio of positive to excised nodes were significantly associated with worse survival.

Table of Contents

Lists of Tables

Lists of Figures

Acknowledgments

I would like to thank all the people who have helped and inspired me during my Master's study at University of Victoria. I especially want to thank my supervisor Dr. Mary Lesperance for her support, patience and guidance in countless way. Her enthusiasm in research has motivated all her students and her willingness to help has made my research smooth and rewarding. I was delighted to take courses from Dr. Laura Cowen and Dr. Farouk Nathoo and have them as my committee members. The knowledge I have learned from them is critical for my research. Dr. Pauline T. Truong deserves a special thank you as my external committee member. I would like to thank her for providing the dataset, valuable information and support for this research. Furthermore, I want to thank my families and friends who have always been supportive and encouraging. My research is supported by the Department of Mathematics and Statistics, University of Victoria.

Chapter 1

1 Introduction

1.1 Background Information on Breast Cancer

Breast cancer is the most commonly diagnosed cancer in women worldwide and accounts for 4% of mortality among North American women [23]. The established risk factors for breast cancer include age, age at menarche, age at menopause, age at first birth, number of pregnancies, breast-feeding, exogenous hormone use, benign breast disease, breast density, family history, height, adiposity, alcohol consumption, physical activity, and ionizing radiation [3].

When a suspicious lump or abnormality is detected by self-examination or screening mammogram, the following steps will take place to further manage breast cancer: diagnosis, staging, primary therapy, and local or systemic adjuvant therapy if necessary [23]. If the diagnosis is uncertain after mammogram and ultrasound, a biopsy will be performed to take some breast tissue from the suspicious area or remove the lump. Once the tissues are diagnosed as cancerous, a treatment strategy will be planned based on a pathology report containing the gross description of the tissues, the microscopic description, and final diagnosis [23].

There are several staging systems for breast cancer and the most common two are the Stage I, II, III, IV system and the TNM system [23]. The TNM system, developed by Pierre Denoix in 1942, describes the extent of the cancer based on three tumor

morphological attributes: the size/extent of the primary tumor (T), regional lymph node involvement (N, nodal status), and presence or absence of distant metastases (M) [3].

1.2 Pathology Report

Tumor size, tumor type, tumor invasion, tumor extension, grade, lymph node involvement, and estrogen and/or progesterone receptors (ER and/or PR) status are some of the important features in the pathology report that influence a doctor's decision on treatment strategy [23]. *In situ* (noninvasive) cancer is still within the milk ducts (ductal) and/or glands (lobular) of the breast and invasive cancer is in the normal fatty tissue of the breast through the walls of the milk ducts and glands [23]. Noninvasive cancers are more likely to be cured.

According to the TNM staging system, tumor stage T0 and Tis represent no tumor and *in situ* cancer. T stage 1-3 represents invasive cancer with size no larger than 2cm, 2cm to 5cm, and larger than 5cm in diameter respectively. Tumor stage T4 includes patients with a tumor of any size fixed to the chest wall, invading the skin, both skin and chest wall, or inflammatory cancer [23]. Stage T4 patients have the worst prognosis and the tumors may never be removed entirely in primary therapy. A tumor may extend to skin, muscle, and excision margins and invade lymphatic, vascular or perineural spaces. Some of the cancer may be left in the breast after surgery if the cancer is close to the edge of the removed lump instead of the center within a block of normal tissue. The larger the cancer, the more likely it will spread and form tumors in other body parts [23].

According to the microscopic appearance of the tumor, breast cancer can be classified into grade I, II or III representing the degree of aggressiveness. Grades I to III represent well/fairly well differentiated, moderately/partially differentiated, and poorly differentiated from the normal tissues respectively. A higher grade is associated with faster cancer growth, earlier spread of the cancer, and a greater incidence of axillary lymph node invasion [23]. Increasing number of cancerous axillary nodes increases the risk of cancer spreading. When treating a patient with surgery alone, the chance that a cancer spreads or reappears within 5 to 10 years is: 30% to 50% with 1 to 3 cancerous lymph nodes; 50% to 75% with 4 to 9 cancerous lymph nodes; and 75% or greater with 10 or more cancerous lymph nodes [23]. The prognosis of having 1 to 3 lymph nodes that are cancerous is similar to finding cancer in lymph channels or blood vessels. If hormone therapy is recommended, the amount of estrogen receptor (ER) can be measured by chemical testing of tumor tissue to estimate how well the cancer would respond to the anti-estrogen drug. Since the higher the ER level, the more likely the tumor will respond to hormone therapy, ER negative patients may have worse prognosis than ER positive patients [23].

1.3 Background Information on Micrometastasis

Once the primary tumor is developed in the breast, cancerous cells can spread or metastasize beyond the breast via lymphatic systems or blood vessels [23]. There are two main types of breast cancer metastases. If the cancerous cells travel to the axillary lymph nodes, it is considered to be early breast cancer [23]. There is a good chance of recovery when proper surgery and systemic therapies are applied. Based on the number of axillary

lymph nodes involved, lymph node status can be categorized as pN0, pN1, pN2 and pN3 [3]. If the cancerous cells travel beyond the lymph nodes to another part of the body such as lungs, liver, bones or brain, it is considered distant metastasis [23]. At this stage, the cancer is no longer curable, but treatment can prevent the cancer from spreading further.

Micrometastasis refers to a tumor in the axillary lymph nodes with less than 2.0mm in diameter. Node negative subgroup (pN0) refers to patients with a tumor in the nodes of size less than 0.2mm in diameter or none. In the past, micrometastatic disease was classified as nodal status pN1a and considered to have similar prognosis as pN0 disease in the former American Joint Committee on Cancer (AJCC) 5^{th} staging edition. Changes in the 6^{th} edition of AJCC (2002) indicate that a tumor in the node with size larger than 0.2mm and no larger than 2.0mm in diameter is defined as a micrometastasis and is distinguished from pN0 tumor which is no larger than 0.2mm [3]. Macrometastasis refers to a tumor with size lager than 2.0mm in the node. One current controversy in breast cancer research is that recent studies suggest that patients with nodal micrometastasis have worse prognosis than node negative (pN0) patients warranting more aggressive treatment approaches.

Chapter 2

2 Literature Review on studies of Micrometastasis

Nine journal articles related to micrometastatic disease were reviewed and summarized. Attiyeh et al. [1] conducted a study to look at the effects of the level of axillary node involvement, the number of positive nodes, and the extent of metastasis on survival rate. Data on 105 patients with primary operable breast cancer with positive axillary nodes who were treated by radical mastectomy at Memorial Hospital in 1960 was collected for a retrospective study with 14 years follow-up time. There were 18 patients with micrometastatic disease (<2mm) and none of them were found to have micrometastases in more than three positive nodes. The 14-year survival rate for all 105 patients was 43%. The 10 and 14-year survival rates were 75% and 67% for patients who had micrometastasis and 40% and 36% for patients with macrometastasis, respectively.

Chen et al. [5] conducted a study to investigate whether the prognosis of patients with lymph node micrometastases ≤2mm (N1mi) would be intermediate between patients with no positive regional lymph node (N0) and patients with macrometastases in no more than three nodes (N1). The notations for nodal subgroups used in this paper are different from our study. N0, N1mi, and N1 represent pN0, pN1a, and pN1b in our study respectively. All patients diagnosed with invasive ductal and lobular breast cancer from 1992 to 2003 were selected from the surveillance, epidemiology, and end results (SEER) cancer registry. Patients with unknown number of tumor-involved nodes, distant metastases, macrometastases in more than three axillary nodes, and incomplete staging

information were not included. The study population was narrowed to 209,720 patients and 11,405 patients had micrometastasis. Methods of lymph node metastases detection were not specified. The utilization of adjuvant therapies was not available. The demographics (age, gender, and race) were similar when stratified by N-stage. As T stage increased so did nodal involvement. The N1mi patients had a higher rate for ER positive, progesterone receptor (PR) positive, and mixed histology (both ductal and lobular) compared to the other two N-stages. The percentage of N1mi patients increased from 2.3% to 7% from 1992 to 2003 due to a stage migration from N0 patients caused by the increasing use of sentinel node biopsy, a less invasive method for detecting nodal metastasis.

Kaplan-Meier (KM) survival curves and a log-rank test were computed for all patients and patients stratified by T-stage in Chen et al. [5]. Five and ten-year overall survival and corresponding *p*-values for pair-wise log-rank tests comparing N0 and N1 patients with N1mi patients are tabulated in Table 2.1. The unstratified overall survival for N1mi patients was significantly worse than N0 patients and better than N1 patients. When stratified by T-stage, the overall survival of N1mi patients was significantly better than N1 patients in all three T-stages but only significantly worse than N0 patients in T2 stage. Although N1mi patients tended to have poorer survival than N0 patients in T1 and T3 stages, the results were not significant. This could be confounded by the population's heterogeneity, since N-stage was a significant independent prognostic indicator in multivariable analysis when adjusting for other factors. In conclusion, for both unstratified and stratified approaches, five and ten-year survival outcomes for patients with micrometastases were intermediate compared to N0 and N1 patients. Multivariable

analysis using a Cox proportional hazards model was performed adjusting for the following significant prognostic factors selected using a forward stepwise model: gender, age, year of diagnosis, N-stage, T-stage, histology, grade, ER status, PR status, and region of the country. N1mi was a significant prognostic indicator ($p < 0.0001$) and the hazard ratios of N1mi compared with N0 and N1 were 1.35 and 0.82 respectively. Male gender, older age, earlier years of diagnosis, larger T-stage, ductal histology, poorer differentiated grade, negative hormonal receptors, and certain locations were associated with worse prognosis.

Table 2.1: Five and ten-year overall survival by nodal status N-stage for the entire cohort and when stratified by T-stage. *P*-values for pair-wise log-rank tests (H_0: equal survival curves) comparing N0, N1 with N1mi are provided.

		N-stage		
		N0	**N1mi**	**N1**
Entire cohort	**5-year survival**	90%	86%	82%
	10-year survival	76%	71%	65%
	***p*-value**	*<0.001*		*<0.001*
T1	**5-year survival**	92%	91%	88%
	10-year survival	78%	77%	73%
	***p*-value**	*0.07*		*<0.001*
T2	**5-year survival**	84%	80%	77%
	10-year survival	69%	63%	60%
	***p*-value**	*<0.001*		*<0.001*
T3	**5-year survival**	82%	77%	70%
	10-year survival	68%	66%	56%
	***p*-value**	*0.089*		*0.017*

In Clayton and Hopkins [6], 399 infiltrating ductal breast cancer patients with axillary metastases were included from a previous cohort of 1045 breast cancer patients. 62 patients had micrometastasis in which 24 of them had nodal tumor size less than 1.8cm. The mean follow-up was 16.7 years and 87% had a follow-up of more than 10 years. The tumor-related survival rates were calculated by actuarial life-table analysis with generalized Wilcoxon or Savage statistics. Two hundred and forty six patients died due to breast cancer and the tumor-related survival rates of 75%, 50%, and 15% were estimated at 1.5, 3.2, and 9.9 years. Stratified survival curves indicated that measures of the extent of lymph node metastasis were the best predictors of survival, which included the number of lymph node metastases (four groups: 1, 2-3, 4-6, ≥ 7 metastases) ($p < 0.0001$), the size of the largest metastasis ($p < 0.0001$), and the presence of lymph node capsular invasion ($p < 0.0001$). The nodal status was also a significant predictor of survival ($p <0.0001$). The prognosis for patients with micrometastases (1-2 mm in diameter) was slightly worse than node-negative patients ($p = 0.22$). Among patients with nodal tumors less than 1.8 cm in diameter, the survival outcome for those who had micrometastases was significantly worse than those without metastases ($p = 0.05$).

Hartveit and Lilleng [15] collected survival data on 1069 patients with unilateral breast cancer, treated between 1980 and 1989, from the Central Bureau of Statistics in Oslo. The mean post-operative follow-up was 6 years until the end of 1992. Out of the 625 node-negative cases 41 had micrometastases (tumor area ≤ 0.2 cm^2) and out of the 444 node-positive cases 126 had micrometastases. In total, 167 patients (15%) had micrometastases from which 138 had a single micrometastasis and 29 had two or more. The authors state that 17 cases died from breast cancer out of the 126 reported node-

negative cases and 8 cases died from breast cancer out of the 41 reported node-negative cases found on review, while the difference was not significant ($\chi^2 = 0.9$). Micrometastatic patients found from node-negative cases and node-positive cases had similar mean survival times (t = 0.4) of 43.2 months (SD = 27.9) and 39.4 months (SD = 17.1). The 26 patients who were reported to die from breast cancer had significantly smaller micrometastases than those of the 125 patients who were still alive at follow-up ($p < 0.0025$). [Note: 17+8=25. There is a typographical error in the paper.] Among the 138 cases with a single micrometastasis, histological examination showed two variants in some patients: tumor growth confined to the capsular lymphatics and/or the subcapsular sinus and tumor growth in the nodal lymphoid tissue. Micrometastatic patients with and without nodal growth had mean survival times of 50.4 months (SD = 28.6) and 38.1 months (SD = 16.8) respectively. The difference in survival time of micrometastatic patients with and without nodal growth was not significant (t = 1.2). The number of deaths for micrometastatic patients with and without nodal growth was significantly different ($\chi^2 = 9$, $p < 0.0035$). The percentage of death from breast cancer was similar in the following pairs: micrometastases and node-negative ($\chi^2 = 0.1$), micrometastasis with nodal growth and node-negative ($\chi^2 = 0.1$), and micrometastasis without nodal growth and node-positive ($\chi^2 = 0.8$).

Klauber-DeMore et al. [16] included 122 patients with stage II or III breast cancer treated with neoadjuvant chemotherapy at the University of North Carolina at Chapel Hill from 1991 to 2002 with a median follow-up of 5.4 years. Kaplan-Meier survival curves for distant disease-free survival (DDFS) and overall survival (OS) with log-rank trend tests were calculated to examine the postneoadjuvant chemotherapy prognostic

significance of lymph node metastasis size and number of positive nodes. There were 11 patients with micrometastasis who had a similar relapse pattern to patients with metastasis from 2 mm to 2 cm and significantly worse DDFS and OS compared to node-negative patients. The 5-year DDFS and OS for patients with micrometastasis were 42% and 43% with 95% confidence intervals 11%-72% and 7%-76%, respectively.

Data on 1306 patients with a complete axillary dissection were collected in Kuijt et al. [19] from the population-based Eindhoven Cancer Registry in Netherlands to identify a subgroup who can safely avoid axillary dissection. The clinico-pathological features of 489 patients with only one positive lymph node and 817 patients with more than one positive lymph node were compared to examine prognostic factors related to further axillary metastatic involvement. Chi-square tests were used to determine patient and tumor differences between the two groups. Univariable and multivariable logistic regressions were used to examine the effect of the following covariates: age, period of diagnosis, histological type, tumor site, tumor size, tumor grade, lymphovascular invasion, ER and PR status, number of lymph nodes examined, number of positive lymph nodes, metastasis size, extranodal extension, and axillary apex involvement. In both the univariable and multivariable analyses, "tumor size greater than 1 cm, harvesting more than 15 axillary lymph nodes at histopathological examination, metastasis size larger than 2 mm, extranodal extension, and nodal involvement of the axillary apex are independently associated with the occurrence of more than one metastatic axillary lymph node". Tumor size less than 1 cm, presence of a micrometastasis, and no extranodal extension were associated with only one positive axillary lymph node independently. No subgroup could be identified to safely avoid axillary dissection.

In another study, Kuijt et al. [20] collected data on 5196 patients diagnosed with invasive breast cancer from 1975 to 1997 from the population-based Eindhoven Cancer Registry in Netherlands with a follow-up completed in April 2002. Patients were divided into three subgroups: 4377 patients without axillary metastasis (pN0), 179 patients with axillary micrometastasis smaller than 2 mm (pN1a), and 640 patients with a macrometastasis larger than 2 mm in only one lymph node (pN1). The three subgroups were compared using a log-rank test and survival outcomes were computed using the life-table method. Univariable and multivariable Cox proportional hazards regressions were fitted adjusting for potential confounders including age and tumor size. In order to exclude the confounding effect of adjuvant systemic treatment, the Cox models were also fitted to the subset of patients who did not receive chemotherapy or hormonal therapy. Hazard ratios with 95% CI and *p*-values were estimated. Ten-year overall survival for the three subgroups with and without adjuvant systemic treatment were tabulated in Table 2.2. The log-rank test indicated no significant difference in unadjusted survival curves between the subgroups. When patients with adjuvant systemic treatment were excluded, the overall survival for patients in the pN1a subgroup was significantly worse than patients in the pN0 subgroup ($p = 0.019$) but was not significantly different from patients in the pN1 subgroup ($p = 0.13$). In the multivariable Cox model, micrometastasis was a significant predictor of mortality. The hazard ratios (95% confidence interval) of pN1a versus pN0 were 1.32 (1.03-1.69) and 1.51 (1.11-2.06) with and without adjuvant systemic treatment respectively. The hazard ratios of pN1 versus pN1a were 1.02 (0.77-1.34) and 1.27 (0.78-1.83) with and without adjuvant systemic treatment respectively.

Table 2.2: Ten-year survival for pN0, pN1a, and pN1 subgroups with and without adjuvant systemic treatment.

	pN0 (N = 4437)	**pN1a (N = 179)**	**pN1 (N = 640)**
Adjuvant systemic treatment	0.690	0.619	0.595
No adjuvant systemic treatment	0.697	0.561	0.443

Kurosumi et al. [21] reported on resection specimens from 92 patients with invasive ductal breast cancer who were treated with quadrantectomy and axillary lymph node dissection were collected from the Tumor Registry of the Department of Clinical Pathology of Saitama Cancer Center from 1992 to 1994 to study the efficacy of detecting the presence of axillary lymph node metastasis including micrometastasis by examining lymphatic invasion in peritumoral breast tissue. Among the 63 node-negative patients, 3 were determined to have micrometastasis by immunohistochemistry. When they were included, the accuracy of detection increased from 84.8% to 88.0% with a sensitivity of 90.6% and a specificity of 86.7%.

Four hundred and eighty four patients with unilateral breast cancer and positive axillary nodes including micrometastases were included in Lilleng et al. [22]. The patients had a mean post-operative follow-up of 6 years. Survival data was collected from the Central Bureau of Statistics in Oslo; 152 patients had micrometastasis. The nodal tumor-load and the total number of positive nodes were used for analysis. The total axillary tumor-load was calculated from the morphometrically recorded tumor area (cm^2). Actuarial life table estimates of survival, log rank tests and the Cox proportional hazards model including tumor-load and number of positive nodes was used to analyze the data. There was a significantly high positive correlation between the tumor-load and the total

number of positive nodes. Survival outcomes between patients with 1-3 positive nodes and 4-6 positive nodes were significantly different, as well as between 4-6 positive nodes and 7-12 positive nodes. There was no significant linear correlation between survival of the 168 patients who died of breast cancer and number of positive nodes. Patients with tumor-load under 0.0001 cm^2 and over 0.5 cm^2 had high death rate and patients with tumor-load between 0.0001 cm^2 and 0.5 cm^2 had low death risk. There was significant difference in survival of patients with tumor-load between 0.0001 cm^2 and 0.5 cm^2 and tumor-load over 0.5 cm^2. There was a significant positive linear correlation between survival and tumor-load up to 0.25 cm^2 ($r = 0.544$, $p < 0.005$) and a significant negative linear correlation between survival and tumor-load over 0.25 cm^2 ($r = -0.164$, $p < 0.05$). Patients with ≤3 positive nodes and small tumor-load (less than 0.5 cm^2) had significantly better survival ($p < 0.0001$). Patients with small tumor-load under 0.5cm^2 had better prognosis than those with tumor-load greater than 0.5cm^2 regardless of number of positive nodes.

These articles provided solid support for our study. In summary, the survival outcomes of the micrometastatic patients were worse than the node-negative patients, but better compared to the macrometastatic patients. Number of positive lymph nodes and nodal status were important prognostic predictors. Increasing number of positive nodes was associated with worse survival. However, each article had its limitations, such as small sample size of micrometastatic patients, short follow-up times or lacking information on adjuvant therapy. Some of the methodology and results were questionable, such as only looking at patients who died or analyzing linear correlation. None of the

above articles looked at lymph node ratio. The prognostic impact of the lymph node ratio on micrometastatic disease is a novel area that requires further exploration.

Chapter 3

3 Description of Dataset

3.1 Data Collection

Data was provided by Dr. Pauline T. Truong from the BC Cancer Agency (BCCA) Vancouver Island Centre, Breast Cancer Outcomes Unit. The original data set contained 9,638 patients diagnosed from 1988-1998 with AJCC 5th edition stage T1-2, pN0, and pN1a-b breast cancer with no distant metastasis. Patients with unknown number of positive nodes and unknown number of nodes removed were not included in the original data set. Information on treatment for each patient including type of surgery (breast conserving surgery and mastectomy), systemic therapy (hormonal therapy and chemotherapy), and radiation therapy was provided. In addition, the following information was also included: ID number, age at diagnosis, date of diagnosis, cause of death, survival years, histologic type, tumor stage, tumor grade, tumor size, nodal status, number of positive nodes, number of nodes removed, ER status, lymphatics and/or veins invasion status, and locoregional recurrence information.

The revision of the AJCC nodal staging classification suggested that the absolute number of positive nodes in patients was an important prognostic factor. However, the relationship between survival and number of positive nodes had not been fully explored. In addition, the effect of the number of excised nodes on prognosis remained unclear. Recent studies suggested that the lymph node ratio (LNR), which was the ratio of the number of positive nodes versus the number of nodes removed, was associated with

recurrence and survival [30]. Therefore the LNR for each patient was also provided in the data set.

3.2 List of Variables

Dr. Pauline T. Truong provided a detailed data dictionary for all variables used in this study. Some categorical variables were created through the study based on the original variables and study interests. The following variables were used in the analyses:

- *surgn* is TNM surgical N stage also referred to as the nodal status that indicates regional lymph node involvement. A 0 represents no axillary lymph node metastases (the pN0 subgroup), 1A represents micrometastases > 0.2mm but ≤2mm (the pN1a subgroup), and 1B represents metastases > 2mm but < 2cm (the pN1b subgroup).
- *dxage* is the patient's age at diagnosis.
- *dxagecat* is a categorized version of *dxage* with a cut point of age 50. Patients with age less than 50 are categorized as 0 and patients with age greater or equal to 50 are categorized as 1.
- *survyrs* is the number of years of survival of a patient from diagnosis to death or October 31st 2004 if alive. This variable can be used to calculate follow-up time, overall survival (OS), and breast cancer specific survival (BCSS). It can also be referred to as time-to-death or last-follow-up from diagnosis.
- *brdeath* indicates the cause of death. It indicates whether a patient is dead from breast cancer (category 1), dead from a cause other than breast cancer (category 2), alive (category 3), or dead from an unknown cause (category 9).

- *tumsize* is the size of a tumor (cm) at diagnosis based on the following source in the specified order: pathology review, pathology report, staging diagram, mammogram, and pre-operative clinical exam. When a tumor consists of both invasive and *in situ* disease, the size of lesion refers to invasive carcinoma. However if the size of the invasive component is not specified, the size of the whole tumor is recorded.
- *stage1* is TNM pathological T stage. Category 1 means T1 stage, a tumor no more than 2 cm. Category 2 is T2 stage, a tumor of size 2 to 5 cm.
- *histcat* represents histology categories, where 1 denotes ductal, 2 denotes lobular, and 3 denotes other.
- *grade* is the grade of primary tumor that is the histopathological degree of differentiation of the malignancy or the total number of histopathological features translated into a grade. When there is a discrepancy between invasive and *in situ* in a tumor, the invasive grade is recorded. However, if the invasive grade is not commented on and only the *in situ* component is graded, it will be recorded as unknown. A 1 (grade 1, low grade) represents well/fairly well differentiated, 2 (grade 2) represents moderately/partially differentiated, moderately well differentiated, 3 (grade 3, high grade) represents poorly differentiated, and 9 (unknown) represents undetermined/not stated grade of differentiation.
- *gradenew* is defined in the same way as *grade* but with missing values treated as a separate category.

- *erposneg* is estrogen receptor status at diagnosis, where 0 indicates ER status negative, 1 indicates ER status positive, and 9 indicates ER status not done or unknown if done.
- *erposnegnew* is categorized in the same way as *erposneg* but with missing values treated as a separate category.
- *lvi* indicates whether there is an invasion of lymphatics and/or veins at diagnosis. A 1 denotes positive, 2 denotes negative, and 9 denotes unknown.
- *lvinew* is *lvi* with missing values treated as a separate category.
- *mx* indicates initial complete/total mastectomy where 0 denotes no and 1 denotes yes.
- *bcs* indicates initial breast conserving surgery in which 0 represents no and 1 represents yes.
- *MxBcs* is a combined version of *mx* and *bcs*. A 0 indicates patients who have neither or either surgeries and 1 indicates patients who have both mastectomy and breast conserving surgery.
- *systx* denotes the type of initial systemic therapy where 0 represents no initial systemic therapy, 1 represents hormonal therapy alone, 2 represents chemotherapy alone, and 3 represents both hormonal therapy and chemotherapy.
- *systxcat* represents categorized systemic therapy in which 0 includes patients with no initial systemic therapy and 1 includes patients with systemic therapy.
- *posnodes* is the number of positive axillary lymph nodes at diagnosis.

- *nodecat* is a categorization of the number of positive nodes. A 0 indicates node negative, 1 indicates 1-3 positive nodes, and 2 indicates 4 or more positive nodes.
- *nodecat3* is another categorized version of the number of positive nodes, where 0 represents no positive node, 1 represents 1 positive node, 2 represents 2 positive nodes, 3 represents 3 positive nodes, and 4 represents more than 4 positive nodes.
- *noderem* is the total number of axillary nodes removed at diagnosis. Unknown number of removed nodes is recorded as 99 but not included in the original data set.
- *noderemcat* is a categorical version of the total number of removed nodes. If no more than 15 nodes are removed, they are categorized as 0. If more than 15 nodes are removed, they are categorized as 1.
- *lnr* is the lymph node ratio which means the ratio of the number of positive nodes versus the number of removed nodes.
- *lnr20* and *lnr25* are two categorized versions of *lnr* with cut points 0.20 and 0.25 respectively. A 0 represents *lnr* equals 0 in both versions. In *lnr20*, 1 indicates *lnr* from 0.01 to 0.20 and 2 indicates *lnr* greater than 0.20. In *lnr25*, 1 represents *lnr* from 0.01 to 0.25 and 2 represents *lnr* greater than 0.25.
- *lnr3* is another categorization of *lnr* in which 0 means *lnr* equals 0, 1 represents $0.01<lnr\leq 0.10$, 2 represents $0.10<lnr\leq 0.15$, 3 represents $0.15<lnr\leq 0.20$, 4 represents $0.20<lnr\leq 0.25$, 5 represents $0.25<lnr\leq 0.50$, and 6 represents $lnr>0.50$.

- *nodesurg* is a combination of surgical N stage and a categorization of the number of positive nodes (*nodecat*). A 0 indicates that patients are in the pN0 subgroup with no positive nodes, 1 denotes patients in the pN1a subgroup with 1-3 positive nodes, 2 denotes patients in the pN1a subgroup with 4 or more positive nodes, 3 denotes patients in the pN1b subgroup with 1-3 positive nodes, and 4 denotes patients in the pN1b subgroup with 4 or more positive nodes.
- *lnr25surg* is a combination of nodal status and *lnr25*. A 0 represents patients in the pN0 subgroup with *lnr25*=0. Category 1 and category 2 represent patients in the pN1a subgroup with *lnr25*=1 and *lnr25*=2 respectively. Category 3 and category 4 denote patients in the pN1b subgroup with *lnr25*=1 and *lnr25*=2 respectively.
- *lrgstat* is a locoregional relapse indicator where 0 and 1 represents no and yes respectively.
- *lrgsurv* is the number of years from diagnosis to locoregional recurrence or censoring date.

3.3 Data Cleaning

Two patients whose *studynum* were 6611 and 2499 were found to have suspiciously young ages at diagnosis, being 6 and 19 years old respectively. The patient with *studynum* 6611 was excluded from the data set due to the suspicious age, leaving 9637 patients in total in the analyses. There were 7988 patients who had no axillary lymph node metastases (pN0), 491 patients who had micrometastases (pN1a), and 1158 patients who had metastases greater than 0.2 cm (pN1b). The patient with *studynum* 8972 was identified as node-negative but had 1 positive node. The number of positive nodes was changed to 0 after confirmation.

Chapter 4

4 Preliminary Analysis

Patient characteristics of the entire cohort and by nodal subgroups were tabulated in Table 4.1. Baseline characteristics for each of the nodal status subgroups were compared using Pearson chi-square tests of homogeneity for categorical variables. Histology type was the only variable with a large *p*-value. One-way ANOVA F tests were computed for continuous variables. The median follow-up time was 8.2 years. Patients in the pN1a subgroup were younger than the other two subgroups. Mastectomy, systemic therapy use, tumor size, number of positive nodes, lymph node ratio (LNR), and grade of primary tumor increased with increasing nodal status, which was from pN0 to pN1b. The pN1b subgroup was less likely to be ER positive. The likelihood of LVI positive status increased with nodal status.

Table 4.1: Characteristics of the entire cohort and stratified by the three nodal subgroups: pN0, pN1a, and pN1b.

	Entire Cohort N=9637 (%)	**pN0 N=7988 (%)**	**pN1a N=491 (%)**	**pN1b N=1158 (%)**	**P value Chi-square test or F test***
Follow up Time in years Median (min, max)	8.1516 (.12, 15.82)	8.3847 (.12, 15.82)	7.2827 (.67, 15.71)	6.9528 (.21, 15.79)	<.001*
Age in years					
Median (min, max)	59 (19,95)	60 (19,95)	53 (24,89)	55 (24,90)	
< 50	2949 (30.6)	2274 (28.5)	217 (44.2)	458 (39.6)	<.001
≥ 50	6688 (69.4)	5714 (71.5)	274 (55.8)	700 (60.4)	
Type of Surgery					
Mastectomy	3456 (35.9)	2596 (32.5)	243 (49.5)	617 (53.3)	<.001
BCS	6642 (68.9)	5779 (72.3)	271 (55.2)	592 (51.1)	<.001
Systemic Therapy					
Chemotherapy alone	1524 (15.8)	999 (12.5)	160 (32.6)	365 (31.5)	<.001
Hormone therapy alone	2662 (27.6)	2065 (25.9)	183 (37.3)	414 (35.8)	
Both	848 (8.8)	387 (4.8)	128 (26.1)	333 (28.8)	
None	4603 (47.8)	4537 (56.8)	20 (4.1)	486 (4.0)	
Categorized Systemic Therapy					
no initial systemic therapy	4603 (47.8)	4537 (56.8)	20 (4.1)	46 (4.0)	<.001
systemic therapy	5034 (52.2)	3451 (43.2)	471 (95.9)	1112 (96.0)	
Tumor size in cm Median (min, max)	1.500 (.1, 5.0)	1.500 (.1, 5.0)	2.000 (.1, 5.0)	2.500 (.2, 5.0)	<.001*
T stage					
T1	6718 (69.7)	6022 (75.4)	259 (52.7)	437 (37.7)	<.001
T2	2919 (30.3)	1966 (24.6)	232 (47.3)	721 (62.3)	

# Positive Nodes					
Median (min, max)	.00 (0, 44)		1.00 (1, 22)	3.00 (1, 44)	
0	7988 (82.9)	7988 (100.0)	0 (.0)	0 (.0)	<.001
1-3	1139 (11.8)	0 (.0)	445 (90.6)	694 (59.9)	
≥ 4	510 (5.3)	0 (.0)	46 (9.4)	464 (40.1)	
# Removed Nodes					
Median (min, max)	10.00 (1, 52)	10.00 (1, 50)	11.00 (1, 36)	11.00 (1, 52)	
<=15	8046 (83.5)	6744 (84.4)	380 (77.4)	922 (79.6)	<.001
>15	1591 (16.5)	1244 (15.6)	111 (22.6)	236 (20.4)	
LNR					
Median (min, max)	.0000 (.00, 1.00)		.1111 (.03, 1.00)	.2857 (.03, 1.00)	
0.01-0.20	803 (48.7)	0 (100.0)	372 (75.8)	431 (37.2)	<.001
>0.20	846 (51.3)	0 (.0)	119 (24.2)	727 (62.8)	
0.01-0.25	935 (56.7)	0 (100.0)	409 (83.3)	526 (45.4)	<.001
>0.25	714 (43.3)	0 (.0)	82 (16.7)	632 (54.6)	
(.01, .10]	380 (3.9)	0 (.0)	218 (44.4)	162 (14.0)	<.001
(.10, .15]	232 (2.4)	0 (.0)	102 (20.8)	130 (11.2)	
(.15, .20]	191 (2.0)	0 (.0)	52 (10.6)	139 (12.0)	
(.20, .25]	132 (1.4)	0 (.0)	37 (7.5)	95 (8.2)	
(.25, .50]	290 (3.8)	0 (.0)	45 (9.2)	245 (21.2)	
>.50	424 (3.6)	0 (.0)	37 (7.5)	387 (33.4)	
ER Status					
Negative	2016 (23.9)	1621 (23.5)	102 (22.2)	293 (27.3)	.018
Positive	6413 (76.1)	5275 (76.5)	357 (77.8)	781 (72.7)	
# unknown	1208	1092	32	84	

Histology Category					
ductal	8751 (90.8)	7245 (90.7)	452 (92.1)	1054 (91.0)	.331
lobular	764 (7.9)	634 (7.9)	37 (7.5)	93 (8.0)	
other	122 (1.3)	109 (1.4)	2 (.4)	11 (.9)	
Grade of primary tumor					
Grade 1	1571 (17.2)	1449 (19.1)	53 (11.1)	69 (6.2)	<.001
Grade 2	4403 (48.1)	3695 (48.8)	248 (52.1)	460 (41.5)	
Grade 3	3182 (34.8)	2428 (32.1)	175 (36.8)	579 (52.3)	
# unknown	481	416	15	50	
LVI Status					
Negative	6955 (75.1)	6322 (82.1)	253 (53.5)	380 (34.7)	<.001
Positive	2309 (24.9)	1374 (17.9)	220 (46.5)	715 (65.3)	
# unknown	373	292	18	63	
Death Status					
death from breast cancer	1137 (11.8)	697 (8.7)	92 (18.7)	348 (30.1)	<.001
death from other than breast cancer	1009 (10.5)	858 (10.7)	30 (6.1)	121 (10.4)	
alive	7448 (77.3)	6397 (80.1)	369 (75.2)	682 (58.9)	
unknown cause of death	43 (.4)	36 (.5)	0 (.0)	7 (.6)	

Table 4.2 contains crosstabulations of LNR with the numbers of positive nodes in the pN1a subgroup. With cut point 0.25, patients with 1, 2, and 3 positive nodes had higher proportions of LNR ≤.25. Among patients with four or more positive nodes, a higher proportion of them had LNR >0.25. The distribution of LNR ≤0.25 decreased from 1 positive node to ≥4 positive nodes. In the case of cut point 0.20, advancing LNR was correlated with advancing number of positive nodes.

Table 4.2: Distributions of lymph node ratios by the numbers of positive nodes in the pN1a subgroup.

	Number of Positive Nodes in the pN1a Subgroup Total N=491			
Lymph Node Ratio	**1 Positive Nodes N (%) N=336**	**2 Positive Nodes N (%) N=76**	**3 Positive Nodes N (%) N=33**	**≥4 Positive Nodes N (%) N=46**
≤0.20	316 (94.0)	48 (63.2)	7 (21.2)	1 (2.2)
>0.20	20 (6.0)	28 (36.8)	26 (78.8)	45 (97.8)
≤0.25	326 (97.0)	56 (73.7)	21 (63.6)	6 (13.0)
>0.25	10 (3.0)	20 (26.3)	12 (36.4)	40 (87.0)

Chapter 5

5 Survival Analysis

5.1 Definition

Time to event data arises in a number of fields, such as health, biology, medicine, epidemiology, engineering, economics, and demography. Survival analysis typically involves modeling and analyzing time to event data. A survival time is the time until some specified event occurs, usually referred to as a failure [17]. Suppose T is a nonnegative continuous random variable from a homogeneous population representing the lifetimes of individuals. Let $F(x) = P(T \le x)$ denote the cumulative distribution function of T with corresponding probability density function *f(x)*. The survival function is the probability that an individual survives up to time t [17]. The survival function is the complement of the cumulative distribution function and the integral of the probability density function (p.d.f.) [29, 17]. It is defined as

$$S(t) = P(T \ge t) = 1 - F(t) = \int_t^{\infty} f(x)dx \ .$$

The survival function is a monotone decreasing function that equals one at time zero. The p.d.f. can be expressed as

$$f(t) = \frac{F(t)}{dt} = -\frac{dS(t)}{dt} \ .$$

The hazard function represents the conditional failure rate at $T = t$ given that the individual survived up to time t and is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d \ln[S(t)]}{dt} \ .$$

The hazard function is a nonnegative function that can take on many different shapes, such as constant, increasing, decreasing, bathtub-shaped, hump-shaped, etc. [17]. A related quantity is the cumulative hazard function, defined as

$$H(t) = \int_0^t h(u)du = -\ln\left[S(t)\right].$$

Censoring is one of the problems one can encounter when analyzing time to event data. There are different types of right censoring including Type I censoring and Type II censoring. Let $T_1, T_2, \cdots, T_n$ be independently and identically distributed with probability density function *f(x)* and survival function *S(x)* and C_i be a fixed censoring time for the ith individual. Instead of T_i, we observe a pair (Y_i, δ_i) where

$$Y_i = \min(T_i, C_i) \text{ and } \delta_i = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } T_i > C_i \end{cases}.$$

In Type I censoring, an observation is made only if it occurs prior to some predetermined time, which could vary from individual to individual [17]. Type II censoring means that the study would stop when the failure of the first *r* individuals occurs, where *r* is a pre-specified integer [17].

The Kaplan-Meier (KM) estimator is a product-limit estimator of survival for right-censored data. It is a non-parametric distribution free method for estimating survival functions. The KM curve is a right continuous step function that steps down at uncensored observations only [29]. Assume that events occur at *D* distinct times $t_1 < t_2 < \cdots < t_D$, and there are d_i events and n_i individuals at risk at time t_i. The Kaplan-Meier estimator is defined as

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_i \\ \prod_{t_i \le t} (1 - \frac{d_i}{n_i}) & \text{if } t \ge t_i \end{cases}.$$

The variance is estimated by the Greenwood's formula defined as

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i (n_i - d_i)}.$$

5.2 Survival Results

Survival times for the study were defined as number of years from diagnosis to death or October 31st 2004 if alive in this data set. Kaplan-Meier breast cancer-specific survival (BCSS), overall survival (OS), and locoregional recurrence (LRR) were compared for the pN0, pN1a, and pN1b nodal subgroups using the log-rank test in SPSS 15.0. For breast cancer-specific survival, a patient was considered to have the event if she died from breast cancer.

Ten-year unadjusted KM breast cancer-specific survival were 90% (*SE* = .39%), 77% (*SE* = 2.40%), and 66% (*SE* = 1.64%) and overall survival were 79% (*SE* = .53%), 71% (*SE* = 2.52%), and 57% (*SE* = 1.69%) for patients in the pN0, pN1a, and pN1b subgroups respectively (log-rank $p<0.0001$ for both BCSS and OS). Ten-year unadjusted KM LRR for the three nodal subgroups were 92%, 90%, and 85% respectively (log-rank $p<0.0001$). BCSS, OS, and LRR were examined within each of the three nodal subgroups by the number of positive nodes, the number of excised nodes, and LNR (Table 5.2.1-5.2.3). In both pN1a and pN1b subgroups, increasing number of positive nodes was significantly associated with worse survival. Ten-year survival was significantly better

with smaller values of LNR with both cut points 0.20 and 0.25 in both pN1a and pN1b subgroups.

Ten-year unadjusted KM BCSS and OS for the entire cohort and the three nodal subgroups by systemic therapy were tabulated in Table 5.2.4 and Table 5.2.5. When treating each type of systemic therapy separately, patients with no systemic therapy, hormone therapy alone, and chemotherapy alone had the highest ten-year BCSS in the pN0, pN1a, and pN1b subgroups, respectively. Survival of the pN1a subgroup was intermediate of the pN0 and pN1b subgroups. The highest ten-year OS were obtained by patients with both systemic therapies and chemotherapy alone in pN0 and pN1b subgroups, while the results for the pN1a subgroup were not significant. When categorizing systemic therapy into no systemic therapy and systemic therapy use, results for the pN1a subgroup were not significant either. Patients with no systemic therapy had better survival in the pN0 subgroup, but worse survival in the pN1b subgroup. In the entire cohort, patients who had no systemic therapy had both the best ten-year BCSS and OS regardless of how systemic therapy use was categorized.

Table 5.2.1: Ten-year Kaplan-Meier breast cancer-specific survival by the number of positive nodes, number of excised nodes, and LNR.

	10-year KM BCSS (standard error)		
	pN0	**pN1a**	**pN1b**
# Positive Nodes			
0	.9028 (.0039)		
1		.8142 (.0264)	.7748 (.0270)
2		.7657 (.0663)	.7080 (.0351)
3		.6525 (.0961)	.5981 (.0470)
p-value, log-rank test		*.0001*	*<.0001*
0	.9028 (.0039)		
1 – 3		.7942 (.0241)	.7139 (.0199)
≥ 4		.5516 (.0835)	.5785 (.0272)
p-value, log-rank test		*.0001*	*<.0001*
# Excised Nodes			
≤ 15	.9007 (.0043)	.7548 (.0286)	.6569 (.0185)
> 15	.9128 (.0090)	.7952 (.0441)	.6652 (.0354)
p-value, log-rank test	*.2438*	*.9127*	*.8483*
LNR			
0	.9028 (.0039)		
(.01, .10]		.8557 (.0296)	.7809 (.0392)
(.10, .15]		.7850 (.0492)	.7827 (.0406)
(.15, .20]		.7546 (.0783)	.7697 (.0383)
(.20, .25]		.6000 (.1352)	.6831 (.0584)
(.25, .50]		.6331 (.0806)	.6298 (.0371)
> 0.50		.5475 (.0957)	.5386 (.0299)
p-value, log-rank test		*.0001*	*<.0001*
0.01-0.20		.8219 (.0248)	.7775 (.0228)
>0.20		.6044 (.0564)	.5879 (.0218)
p-value, log-rank test		*<.0001*	*<.0001*
0.01-0.25		.8065 (.0250)	.7613 (.0214)
>0.25		.5899 (.0628)	.5740 (.0234)
p-value, log-rank test		*<.0001*	*<.0001*

Table 5.2.2: Ten-year Kaplan-Meier overall survival by the number of positive nodes, number of excised nodes, and LNR.

	Ten-year KM OS (standard error)		
	pN0	**pN1a**	**pN1b**
# Positive Nodes			
0	.7943 (.0053)		
1		.7573 (.0288)	.6899 (.0297)
2		.7008 (.0658)	.5869 (.0391)
3		.5348 (.1009)	.5056 (.0476)
p-value, log-rank test		*.0035*	*<.0001*
0	.7943 (.0053)		
1 – 3		.7298 (.0259)	.6160 (.0215)
≥ 4		.5516 (.0835)	.4943 (.0266)
p-value, log-rank test		*.0042*	*.0001*
# Excised Nodes			
≤ 15	.7878 (.0058)	.6899 (.0299)	.5582 (.0189)
> 15	.8273 (.0120)	.7666 (.0456)	.5952 (.0370)
p-value, log-rank test	*.0011*	*.3165*	*.3079*
LNR			
0	.7943 (.0053)		
(.01, .10]		.8171 (.0320)	.6747 (.0436)
(.10, .15]		.7091 (.0543)	.7027 (.0445)
(.15, .20]		.6651 (.0771)	.6927 (.0446)
(.20, .25]		.4965 (.1288)	.5501 (.0596)
(.25, .50]		.5710 (.0806)	.5379 (.0379)
> 0.50		.5475 (.0957)	.4551 (.0288)
p-value, log-rank test		*.0004*	*<.0001*
0.01-0.20		.7633 (.0269)	.6880 (.0257)
>0.20		.5559 (.0559)	.4951 (.0215)
p-value, log-rank test		*<.0001*	*<.0001*
0.01-0.25		.7435 (.0270)	.6627 (.0238)
>0.25		.5606 (.0619)	.4868 (.0230)
p-value, log-rank test		*<.0001*	*<.0001*

Table 5.2.3: Ten-year Kaplan-Meier locoregional recurrence by the number of positive nodes, number of excised nodes, and LNR.

	10-year KM LRR (standard error)		
	pN0	**pN1a**	**pN1b**
# Positive Nodes			
0	.9211 (.0036)		
1		.8884 (.0223)	.8667 (.0215)
2		.9487 (.0298)	.8742 (.0248)
3		.9374 (.0429)	.7728 (.0442)
p-value, log-rank test		*.5660*	*.2100*
0	.9211 (.0036)		
1 – 3		.9019 (.0180)	.8507 (.0157)
≥ 4		.9000 (.0478)	.8399 (.0216)
p-value, log-rank test		*.6971*	*.8143*
# Excised Nodes			
≤ 15	.9207 (.0039)	.8935 (.0201)	.8496 (.0138)
> 15	.9246 (.0090)	.9305 (.0296)	.8357 (.0313)
p-value, log-rank test	*.2261*	*.2577*	*.8364*
LNR			
0	.9211 (.0036)		
(.01, .10]		.8889 (.0291)	.8474 (.0339)
(.10, .15]		.9253 (.0306)	.8476 (.0329)
(.15, .20]		.9389 (.0342)	.9031 (.0316)
(.20, .25]		.9394 (.0418)	.8878 (.0355)
(.25, .50]		.8507 (.0646)	.8856 (.0245)
> 0.50		.8879 (.0529)	.7783 (.0259)
p-value, log-rank test		*.7561*	*.0017*
0.01-0.20		.9062 (.0194)	.8727 (.0192)
>0.20		.8924 (.0314)	.8305 (.0167)
p-value, log-rank test		*.3313*	*.0362*
0.01-0.25		.9086 (.0813)	.8749 (.0172)
>0.25		.8714 (.0411)	.8215 (.0185)
p-value, log-rank test		*.1468*	*.0118*

Table 5.2.4: Ten-year BCSS according to the three nodal subgroups and systemic treatment.

	Ten-year BCSS (standard error)			
	Entire Cohort N=9637 %	**pN0 N=7988 %**	**pN1a N=491 %**	**pN1b N=1158 %**
Systemic Therapy				
Chemotherapy alone	.8042 (.0113)	.8669 (.0118)	.6938 (.0436)	.6803 (.0270)
Hormone therapy alone	.8486 (.0087)	.8935 (.0086)	.8153 (.0364)	.6498 (.0280)
Both	.7727 (.0177)	.8687 (.0194)	.7921 (.0474)	.6701 (.0310)
None	.9131 (.0048)	.9175 (.0047)	.7579 (.1098)	.5096 (.0845)
p-value, log-rank test	<.0001	<.0001	.0317	.0217
No Systemic Therapy	.9131 (.0048)	.9175 (.0047)	.7579 (.1098)	.5096 (.0845)
Systemic Therapy	.8223 (.0065)	.8829 (.0065)	.7661 (.0245)	.6650 (.0166)
p-value, log-rank test	<.0001	<.0001	.9018	.0028

Table 5.2.5: Ten-year OS according to the three nodal subgroups and systemic treatment.

	Ten-year OS (standard error)			
	Entire Cohort N=9637 %	**pN0 N=7988 %**	**pN1a N=491 %**	**pN1b N=1158 %**
Systemic Therapy				
Chemotherapy alone	.7802 (.0119)	.8464 (.0127)	.6701 (.0448)	.6488 (.0274)
Hormone therapy alone	.6924 (.0109)	.7396 (.0119)	.6968 (.0403)	.4788 (.0280)
Both	.7437 (.0185)	.8473 (.0210)	.7745 (.0495)	.6318 (.0317)
None	.7990 (.0067)	.8039 (.0067)	.7200 (.1106)	.3466 (.0768)
p-value, log-rank test	<.0001	<.0001	.0613	<.0001
No Systemic Therapy	.7990 (.0067)	.8039 (.0067)	.7200 (.1106)	.3466 (.0768)
Systemic Therapy	.7273 (.0075)	.7815 (.0085)	.7088 (.0258)	.5753 (.0172)
p-value, log-rank test	<.0001	.0110	.9090	<.0001

Chapter 6

6 The Cox Proportional Hazards Models

6.1 Introduction to the Cox Proportional Hazards Models

The Cox Proportional Hazards (PH) model is often used to determine the relationship between survival time and various covariates of interest associated with experimental subjects, such as age, weight, blood pressure, levels of treatment, and so on. Let $Z = (Z_1,...,Z_p)'$ represent a vector of covariates that are also referred to as regression variables, regressors, risk factors, or explanatory variables [17]. Given the assumption that covariates Z affect the hazard multiplicatively, the hazard function for the Cox PH model is

$$h(t \mid Z) = \exp(\beta_1 Z_1 + \cdots + \beta_p Z_p) \cdot h_0(t) = \exp(\beta' Z) \cdot h_0(t),$$

where $h_0(t)$ is an arbitrary baseline hazard rate that is the same for all individuals and $\beta = (\beta_1,...,\beta_p)'$ is a parameter vector [17]. Since the baseline hazard rate is treated non-parametrically and only the covariate effect is treated parametrically, the Cox PH model is called a semi-parametric model [17]. It is relatively easy to handle mathematically without specifying a parametric form for the baseline hazard and provides a fairly wide family of distributions. The survival function and p.d.f. can be expressed as

$$S(t \mid Z) = \exp\left[-\exp(\beta' Z)\int_0^t h_0(x)dx\right] = \left\{\exp\left[-\int_0^t h_0(x)dx\right]\right\}^{\exp(\beta' Z)} = \left[S_0(t)\right]^{\exp(\beta' Z)}$$

and

$$f(t \mid z) = h(t \mid Z) \cdot S(t \mid Z) = \exp(\beta' Z) \cdot h_0(t) \cdot \left[S_0(t)\right]^{\exp(\beta' Z)}.$$

In order to interpret the effects of covariates, suppose Z_j changes to $Z_j + 1$, then

$$\exp(\beta' Z) = \exp\left[\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_j (Z_j + 1) + \cdots + \beta_r Z_r\right] = e^{\beta_j} \cdot \exp(\beta' Z).$$

Therefore, e^{β_j} is the relative risk associated with one unit increase in Z_j. In another words, a one unit increase in Z_j increases the hazard by a factor of e^{β_j}. If $\hat{\beta}_j > 0$ $(e^{\hat{\beta}_j} > 1)$, then an increase in Z_j results in an increased hazard. If $\hat{\beta}_j < 0$ $(e^{\hat{\beta}_j} < 1)$, then an increase in Z_j results in a decreased hazard. The hazard ratio for two different covariates Z_1 and Z_2 is

$$\frac{h(t \mid Z_1)}{h(t \mid Z_2)} = \frac{\exp(\beta' Z_1) h_0(t)}{\exp(\beta' Z_2) h_0(t)} = \frac{\exp(\beta' Z_1)}{\exp(\beta' Z_2)},$$

a constant with respect to time, which is referred to as the proportional hazards property [29]. A hazard ratio grater than one means the hazard for Z_1 is greater than Z_2 and a hazard ratio less than one means the hazard for Z_1 is less than Z_2.

Estimates of the parameters β are obtained from maximizing a partial likelihood function based on conditional probabilities which are free of the baseline hazard [29]. Let t_j denote a time at which a death occurs and let $R(t_j)$ denote the risk set at time t_j. A risk set contains all the individuals at risk at time t_j. Assume no censoring and suppose failures at times $t_1, t_2, ..., t_r (r \le n)$ are observed, so that t_j is the j^{th} ordered death time [29]. Let Z_j denote the vector of covariates associated with the individual who dies at t_j. Then for each j,

$$L_j(\beta) = P\{\text{individual with } Z_j \text{ dies at } t_j \mid \text{one death in } R(t_j) \text{ at } t_j\}$$

$$= \underline{P\{\text{individual with } Z_j \text{ dies at } t_j \mid \text{individual in } R(t_j)\}} \ .$$

$$P\{\text{one death at } t_j \mid R(t_j)\}$$

The term $P\{\text{one death at } t_j \mid R(t_j)\}$ can be expressed as $\sum_{l \in R(t_j)} P\{T_l = t_j \mid T_l \geq t_j\}$.

Since

$$\begin{aligned} P\{\text{one death in } [t_j, t_j + \Delta t_j) \mid R(t_j)\} &= \sum_{l \in R(t_j)} P\{T_l \in [t_j, t_j + \Delta t_j) \mid T_l \geq t_j\} \\ &\approx \sum_{l \in R(t_j)} h(t_j \mid Z_l)\Delta t_j \\ &= \sum_{l \in R(t_j)} h_0(t_j)\exp(\beta' Z_l)\Delta t_j \,, \end{aligned}$$

it follows that

$$L_j(\beta) = \frac{h_0(t_j)\exp(\beta' Z_j)}{\sum_{l \in R(t_j)} h_0(t_j)\exp(\beta' Z_j)} = \frac{\exp(\beta' Z_j)}{\sum_{l \in R(t_j)} \exp(\beta' Z_j)} .$$

The partial likelihood function is defined as

$$L(\beta) = \prod_{j=1}^{r} L_j(\beta') = \prod_{j=1}^{r} \frac{\exp(\beta' Z_j)}{\sum_{l \in R(t_j)} \exp(\beta' Z_j)} .$$

Define the partial log-likelihood function $l(\beta)$ and the score vector $U(\beta)$ which are the derivatives with respect to β_j. The Maximum Likelihood Estimator of β can be obtained by solving the r equations in $U(\beta) = 0$.

6.2 The Multivariable Cox PH Models Analyses

Multivariable analyses of BCSS and OS were carried out using the Cox proportional hazards models to determine which factors were significantly associated with survival. Hazard ratios with 95% confidence intervals and *p*-values for each covariate were obtained. Four Cox PH models were fit for the entire cohort and under each of the three nodal subgroups. The two BCSS models and the two OS models all contain the following covariates: age, type of surgery, T stage, number of excised nodes, systemic therapy, ER status, histology type, grade of primary tumor, and LVI status. One of the BCSS and OS models contained number of positive nodes and the other one contained LNR.

The multivariable Cox PH model analyses indicated that T stage, systemic therapy, number of positive nodes, LNR with cut point 0.25, estrogen receptor (ER) status, histologic type, tumor grade, and LVI status were significant factors associated with both BCSS and OS (Table 6.2.1). When missing values were categorized and retained in the models, systemic therapy was not significant for BCSS (Table 6.2.2). The significance level for the other covariates stayed the same as those in Table 6.2.1, cases with missing values removed. Age was significantly associated only with OS in both cases. The multivariable analysis of only BCSS was also carried out to estimate the effects of the covariates of interest on survival under each nodal subgroup with and without cases with missing values removed (Table 6.2.3 and 6.2.4). T stage, number of positive nodes, LNR with cut point 0.25, ER status, grade of primary tumor (except the pN1a subgroup), and LVI status were significantly associated with BCSS under all three subgroups with and without missing values removed. The Cox model hazard ratio

estimates do not converge for categories with no events. Both histology type and primary tumor grade contained categories with no events in the pN1a subgroup. Since only primary tumor grade was a significant indicator of survival in both the pN0 and pN1b subgroup, we combined grade 1 and grade 2 and performed the multivariable Cox PH analysis for BCSS again with and without missing values removed (Table 6.2.5). T stage, number of positive nodes, LNR, and ER status were significant in all models. The number of positive nodes and lymph node ratio were strong prognostic indicators in all models.

In order to compare the goodness-of-fit between models with number of positive nodes and models with LNR, the log-likelihood ratio test statistics are tabulated at the bottom of tables (Table 6.2.1-6.2.2). The log-likelihood ratio test statistics is defined as $-2[l(\beta_0) - l(\hat{\beta})] = 2[l(\hat{\beta}) - l(\beta_0)]$. Based on the partial log-likelihood function defined in the previous section, $l(\beta_0)$ is a constant. Therefore the larger the value from the log-likelihood ratio test statistics, the larger the value for $l(\hat{\beta})$, indicating a better fit. Results suggested the Cox PH models with LNR are better than the models with number of positive nodes.

Table 6.2.1: The Cox Proportional Hazards multivariable analyses of breast cancer-specific survival and overall survival with number of positive nodes or LNR as a covariate, cases with missing values removed. *P*-values from the Wald test for equal hazards are recorded.

	BCSS Hazard Ratio (95% CI) p value		**OS Hazard Ratio (95% CI) p value**	
	With number of positive nodes N=7814	**With LNR N=7814**	**With number of positive nodes N=7814**	**With LNR N=7814**
Age < 50 vs. ≥ 50 y	.895 (.782, 1.026) .112	.901 (.787, 1.032) .134	.499 (.445, .560) <.001	.500 (.446, .562) <.001
Type of Surgery Both Mx&Bcs vs. not both	.941 (.687, 1.290) .707	.952 (.695, 1.304) .759	.976 (.781, 1.221) .834	.980 (.783, 1.225) .856
T stage T2 vs. T1	1.745 (1.510, 2.017) <.001	1.734 (1.500, 2.004) <.001	1.663 (1.492, 1.854) <.001	1.655 (1.485, 1.845) <.001
# Excised Nodes <15 vs. ≥15	1.160 (.980, 1.373) .084	1.091 (.922, 1.291) .308	1.193 (1.047, 1.359) .008	1.152 (1.011, 1.312) .033
Systemic Therapy systemic therapy vs. no systemic therapy	.788 (.664, .936) .007	.793 (.668 .942) .008	.772 (.683, .872) <.001	.775 (.686, .875) <.001
# Positive Nodes				
pN1a $_{1\text{-}3}$ vs. pN0	1.736 (1.333, 2.60)		1.305 (1.046, 1.628)	
pN1a $_{\geq 4}$ vs. pN0	4.263 (2.562, 7.095)		2.440 (1.500, 3.969)	
pN1b $_{1\text{-}3}$ vs. pN0	2.218 (1.822, 2.700)		1.743 (1.484, 2.047)	
pN1b $_{\geq 4}$ vs. pN0	3.042 (2.467, 3.751) <.001		2.293 (1.924, 2.732) <.001	
LNR				
pN1a $_{\leq 0.25}$ vs. 0		1.632 (1.226, 2.194)		1.233 (.974, 1.561)

$pN1a_{>0.25}$ vs. 0		3.677 (2.434, 5.556)		2.259 (1.552, 3.290)
$pN1b_{\leq 0.25}$ vs. 0		1.837 (1.461, 2.310)		1.540 (1.280, 1.852)
$pN1b_{>0.25}$ vs. 0		3.190 (2.643, 3.850) <.001		2.340 (2.001, 2.737) <.001
Estrogen Receptor Status positive vs.negative	.692 (.598, .799) <.001	.692 (.599, .799) <.001	.845 (.752, .949) .004	.845 (.753, .949) .004
Histology Category lobular vs. ductal	.563 (.398, .796)	.563 (.398, .795)	.768 (.627, .942)	.768 (.627, .942)
other vs. ductal	.222 (.083, .595) <.001	.224 (.083, .600) <.001	.517 (.298, .897) .003	.519 (.299, .901) .003
Grade of primary tumour grade 2 vs. grade 1	2.117 (1.527, 2.933)	2.136 (1.542, 2.960)	1.217 (1.031, 1.436)	1.223 (1.036, 1.443)
grade 3 vs. grade 1	3.440 (2.472, 4.787) <.001	3.452 (2.481, 4.803) <.001	1.520 (1.275, 1.812) <.001	1.522 (1.276, 1.814) <.001
LVI Status positive vs.negative	1.660 (1.426, 1.934) <.001	1.639 (1.407, 1.909) <.001	1.471 (1.310, 1.652) <.001	1.464 (1.303, 1.644) <.001
Log likelihood ratio test (d.f.) **H_0: All hazard ratios are 1**	706 (15)	722 (15)	572 (15)	583 (15)

Table 6.2.2: The Cox Proportional Hazards multivariable analyses of breast cancer-specific survival and overall survival with number of positive nodes or LNR as a covariate, cases with missing values retained. *P*-values from the Wald test for equal hazards are recorded.

	BCSS Hazard Ratio (95% CI) p value		**OS Hazard Ratio (95% CI) p value**	
	With number of positive nodes N=9637	**With LNR N=9637**	**With number of positive nodes N=9637**	**With LNR N=9637**
Age < 50 vs. ≥ 50 y	.880 (.776, .997) .045	.884 (.780, 1.002) .054	.487 (.438, .541) <.001	.487 (.439, .541) <.001
Type of Surgery Both Mx&Bcs vs. not both	.959 (.736, 1.250) .758	.958 (.735, 1.248) .752	1.032 (.886, 1.230) .726	1.028 (.863, 1.226) .754
T stage T2 vs. T1	1.853 (1.620, 2.119) <.001	1.833 (1.603, 2.097) <.001	1.692 (1.533, 1.867) <.001	1.680 (1.522, 1.854) <.001
# Excised Nodes <15 vs. ≥15	1.136 (.973, 1.327) .106	1.069 (.916, 1.247) .398	1.209 (1.074, 1.360) .002	1.171 (1.041, 1.317) .009
Systemic Therapy systemic therapy vs. no systemic therapy	.863 (.737, 1.012) .069	.868 (.741, 1.017) .080	.820 (.735, .916) <.001	.822 (.737, .918) <.001
# Positive Nodes				
pN1a $_{1\text{-}3}$ vs. pN0	1.641 (1.275, 2.113)		1.282 (1.041, 1.579)	
pN1a $_{\geq 4}$ vs. pN0	3.877 (2.459, 6.114)		2.163 (1.397, 3.349)	
pN1b$_{1\text{-}3}$ vs. pN0	2.187 (1.821, 2.627)		1.812 (1.566, 2.098)	
pN1b $_{\geq 4}$ vs. pN0	3.023 (2.496, 3.662) <.001		2.357 (2.014, 2.759) <.001	
LNR				
pN1a $_{\leq 0.25}$ vs. 0		1.490 (1.134, 1.959)		1.193 (.953, 1.494)

$pN1a_{>0.25}$ vs. 0		3.578 (2.482, 5.158)		2.130 (1.527, 2.972)
$pN1b_{\leq 0.25}$ vs. 0		1.786 (1.440, 2.215)		1.602 (1.352, 1.897)
$pN1b_{>0.25}$ vs. 0		3.173 (2.670, 3.770)		2.403 (2.088, 2.766)
		<.001		<.001
Estrogen Receptor Status				
positive vs. negative	.700 (.610, .804)	.700 (.610, .804)	.857 (.769, .955)	.857 (.769, .955)
unknown vs. negative	.701 (.563, .871)	.704 (.566, .875)	.873 (.752, 1.015)	.874 (.752, 1.015)
	<.001	<.001	.020	.019
Histology Category				
lobular vs. ductal	.670 (.510, .881)	.670 (.510, .881)	.864 (.733, 1.017)	.864 (.734, 1.018)
other vs. ductal	.273 (.112, .611)	.275 (.123, .616)	.675 (.441, 1.033)	.677 (.442, 1.036)
	<.001	<.001	.044	.045
Grade of primary tumour				
grade 2 vs. grade 1	1.932 (1.451, 2.573)	1.947 (1.462, 2.592)	1.200 (1.035, 1.391)	1.204 (1.038, 1.395)
grade 3 vs. grade 1	3.128 (2.341, 4.179)	3.145 (2.354, 4.202)	1.505 (1.287, 1.760)	1.508 (1.290, 1.763)
unknown vs. grade1	2.155 (1.463, 3.173)	2.156 (1.465, 3.175)	1.362 (1.094, 1.694)	1.363 (1.096, 1.696)
	<.001	<.001	<.001	<.001
LVI Status				
positive vs. negative	1.649 (1.432, 1.898)	1.634 (1.419, 1.881)	1.438 (1.294, 1.598)	1.435 (1.292, 1.594)
unknown vs. negative	1.074 (.773, 1.492)	1.035 (.745, 1.439)	1.131 (.914, 1.400)	1.113 (.899, 1.378)
	<.001	<.001	<.001	<.001
Log likelihood ratio test (d.f.)	849 (18)	870 (18)	720 (18)	734 (18)
H_0: All hazard ratios are 1				

Table 6.2.3: The Cox Proportional Hazards multivariable analyses of breast cancer-specific survival with number of positive nodes or LNR as a covariate within the pN0, pN1a and pN1b subgroups, cases with missing values removed. *P*-values from the Wald test for equal hazards are recorded.

	BCSS Hazard Ratio (95% CI) p value				
	pN0 subgroup	**pN1a subgroup**		**pN1b subgroup**	
	N=6390	**With number of positive nodes N=436**	**With LNR N=436**	**With number of positive nodes N=988**	**With LNR N=988**
Age < 50 vs. ≥ 50 y	.966 (.808, 1.155) .705	.972 (.623, 1.514) .899	1.047 (.672, 1.633) .838	.762 (.599, .968) .026	.765 (.603, .971) .028
Type of Surgery Both Mx&Bcs vs. not both	1.030 (.694, 1.530) .883	.894 (.320, 2.499) .831	.783 (.278, 2.203) .643	.914 (.512, 1.634) .763	.945 (.529, 1.688) .848
T stage T2 vs. T1	1.950 (1.622, 2.343) <.001	1.814 (1.132, 2.907) .013	1.697 (1.052, 2.738) .030	1.431 (1.104, 1.854) .007	1.427 (1.102, 1.849) .007
# Excised Nodes <15 vs. ≥15	1.249 (.988, 1.580) .064	1.317 (.798, 2.172) .281	1.076 (.645, 1.794) .780	1.043 (.787, 1.382) .769	.922 (.697, 1.219) .567
Systemic Therapy systemic therapy vs. no systemic therapy	.743 (.611, .903) .003	1.318 (.464, 3.743) .605	1.169 (.411, 3.322) .770	.677 (.397, 1.156) .153	.680 (.399, 1.158) .155
# Positive Nodes ≥ 4 vs. 1-3		2.500 (1.412, 4.425) .002		1.388 (1.095, 1.759) .007	
LNR $pN1a_{>.25}$ vs. $pN1a_{\leq 0.25}$			2.457 (1.483, 4.073) <.001		1.791 (1.401, 2.290) <.001
Estrogen Receptor Status positive vs. negative	.793 (.654, .960)	.351 (.222, .555)	.369 (.235, .577)	.648 (.503, .834)	.643 (.500, .827)

	.018	<.001	<.001	.001	.001
Histology Category lobular vs. ductal other vs. ductal	.544 (.346, .854) .140 (.035, .564) .001	No event in category other	No event in category other	.479 (.245, .938) .683 (.167, 2.792) .087	.475 (.243, .928) .714 (.175, 2.918) .084
Grade of primary tumour grade 2 vs. grade 1 grade 3 vs. grade 1	1.876 (1.316, 2.674) 3.240 (2.255, 4.657) <.001	No event in category grade 1	No event in category grade 1	2.339 (.946, 5.785) 3.697 (1.501, 9.102) <.001	2.411 (.975, 5.964) 3.767 (1.530, 9.274) <.001
LVI Status positive vs. negative	1.816 (1.497, 2.203) <.001	1.634 (1.036, 2.578) .035	1.547 (.979, 2.443) .061	1.449 (1.096, 1.916) .009	1.396 (1.057, 1.843) .019

Table 6.2.4: The Cox Proportional Hazards multivariable analyses of breast cancer-specific survival with number of positive nodes or LNR as a covariate within the pN0, pN1a and pN1b subgroups, cases with missing values retained. *P*-values from the Wald test for equal hazards are recorded.

	BCSS Hazard Ratio (95% CI) p value				
	pN0 subgroup	**pN1a subgroup**		**pN1b subgroup**	
	N=7988	**With number of positive nodes N=491**	**With LNR N=491**	**With number of positive nodes N=1158**	**With LNR N=1158**
Age < 50 vs. ≥ 50 y	.942 (.800, 1.108) .469	.974 (.635, 1.496) .906	1.061 (.691, 1.631) .786	.757 (.606, .946) .014	.758 (.608, .946) .014
Type of Surgery Both Mx&Bcs vs. not both	.933 (.668, 1.304) .684	.991 (.395, 2.484) .984	.857 (.341, 2.158) .744	1.037 (.634, 1.695) .886	1.027 (.627, 1.680) .917
T stage T2 vs. T1	2.045 (1.727, 2.422) <.001	1.723 (1.099, 2.703) .018	1.604 (1.016, 2.533) .043	1.563 (1.228, 1.991) <.001	1.540 (1.210, 1.960) <.001
# Excised Nodes <15 vs. ≥15	1.178 (.955, 1.454) .126	1.387 (.845, 2.276) .196	1.102 (.663, 1.829) .708	1.053 (.810, 1.368) .701	.924 (.712, 1.199) .553
Systemic Therapy systemic therapy vs. no systemic therapy	.833 (.697, .996) .045	1.343 (.473, 3.813) .580	1.154 (.406, 3.277) .778	.563 (.350, .907) .018	.566 (.352, .910) .019
# Positive Nodes $pN1a_{\geq 4}$ vs. $pN1a_{1\text{-}3}$		2.684 (1.582, 4.555) <.001		1.408 (1.131, 1.753) .002	
LNR $pN1a_{>.25}$ vs. $pN1a_{\leq 0.25}$			2.779 (1.731, 4.464) <.001		1.820 (1.445, 2.291) <.001
Estrogen Receptor Status positive vs. negative	.785 (.655, .941)	.350 (.222, .553)	.370 (.237, .578)	.680 (.533, .868)	.675 (.529, .860)

unknown vs. negative	.718 (.551, .936) .012	.307 (.122, .772) <.001	.352 (.143, .867) <.001	.927 (.598, 1.437) .005	.924 (.596, 1.431) .004
Histology Category lobular vs. ductal other vs. ductal	 .645 (.453, .920) .229 (.085, .616) .001	 No events in category other	 No events in category other	 .610 (.367, 1.014) .530 (.130, 2.160) .113	 .611 (.368, 1.014) .541 (.133, 2.201) .116
Grade of primary tumour grade 2 vs. grade 1 grade 3 vs. grade 1 unknown vs. grade 1	 1.690 (1.244, 2.298) 2.846 (2.080, 3.894) 1.642 (1.046, 2.578) <.001	 No events in category grade 1	 No events in category grade 1	 2.631 (1.067, 6.484) 4.311 (1.755,10.586) 4.400 (1.599,12.107) <.001	 2.681 (1.088, 6.609) 4.377 (1.783,10.745) 4.292 (1.558,11.823) <.001
LVI Status positive vs. negative unknown vs. negative	 1.789 (1.498, 2.137) 1.159 (.747, 1.798) <.001	 1.751 (1.125, 2.724) .266 (.036, 1.953) .012	 1.661 (1.066, 2.588) .219 (.030, 1.611) .015	 1.404 (1.080, 1.826) .967 (.566, 1.654) .021	 1.365 (1.051, 1.774) .906 (.530, 1.549) .028

Table 6.2.5: The Cox Proportional Hazards multivariable analyses of breast cancer-specific survival with number of positive nodes or LNR as a covariate with combined grade 1 and grade 2 in the pN1a subgroup, cases with and without missing values removed. *P*-values from the Wald test for equal hazards are recorded.

	BCSS Hazard Ratio (95% CI) p value			
	With missing values removed		**With missing values retained**	
	With number of positive nodes N=436	**With LNR N=436**	**With number of positive nodes N=491**	**With LNR N=491**
Age < 50 vs. ≥ 50 y	1.019 (.650, 1.597) .934	1.101 (.702, 1.728) .675	.977 (.637, 1.498) .913	1.068 (.696, 1.641) .736
Type of Surgery Both Mx&Bcs vs. not both	.689 (.212, 2.241) .536	.638 (.196, 2.079) .456	.951 (.375, 2.411) .915	.830 (.324, 2.121) .697
T stage T2 vs. T1	1.782 (1.104, 2.877) .018	1.669 (1.026, 2.716) .039	1.711 (1.091, 2.685) .019	1.589 (1.005, 2.513) .048
# Excised Nodes <15 vs. ≥15	1.249 (.756, 2.064) .386	1.023 (.611, 1.713) .930	1.362 (.830, 2.237) .221	1.079 (.648, 1.796) .770
Systemic Therapy systemic therapy vs. no systemic therapy	1.217 (.428, 3.457) .712	1.091 (.384, 3.100) .870	1.304 (.460, 3.700) .617	1.117 (.393, 3.171) .836
# Positive Nodes $pN1a_{\geq 4}$ vs. $pN1a_{1\text{-}3}$	2.646 (1.495, 4.684) .001		2.793 (1.642, 4.750) <.001	
LNR $pN1a_{>.25}$ vs. $pN1a_{\leq 0.25}$		2.498 (1.495, 4.172) <.001		2.822 (1.755, 4.538) <.001
Estrogen Receptor Status Positive vs.negative	.406 (.249, .663)	.421 (.259, .684)	.384 (.237, .622)	.407 (.253, .654)

Unknown vs. negative	<.001	<.001	.286 (.110, .745) <.001	.343 (.136, .836) .001
Grade of primary tumour				
grade 3 vs. grade 1 and 2	1.354 (.852, 2.151)	1.317 (.825, 2.100)	1.324 (.848, 2.067)	1.297 (.826, 2.036)
unknown vs. grade 1 and 2	.199	.248	2.298 (.621, 8.513) .264	2.084 (.585, 7.421) .327
LVI Status				
positive vs. negative	1.584 (.977, 2.518)	1.477 (.926, 2.354)	1.718 (1.103, 2.676)	1.619 (1.036, 2.532)
unknown vs. negative	.051	.101	.194 (.023, 1.631) .011	.173 (.022, 1.394) .017

6.3 Residual Analysis

Residual analyses were performed on the four Cox PH models to examine model adequacy on four aspects. First, a martingale residual is the difference between the observed number of events and the expected number of events over time [17]. Normally, martingale residuals are plotted with each given covariate in turn left out of the model to determine the best functional form for that given covariate to explain its influence on survival adjusting for other covariates. In our case, the covariates are categorical and hence we are not concerned with their functional form in the model. Second, since the martingale residuals are often highly skewed, normalized transforms of the martingale residuals, called deviance residuals, were examined to identify observations that were poorly predicted by the fitted model as characterized with large deviance residuals [27]. Third, the graph of influence by observation number was checked to identify influential observations based on their residuals and their distances from the center of the predicted space [27]. Fourth, the rescaled Schoenfeld residuals were plotted for each category of covariates to assess the proportional hazards assumption. A Schoenfeld residual is the difference between the covariate value at a failure time and its expected value [17]. The rescaled Schoenfeld residuals were used in the analysis due to multiple predictors.

The residual plots for the BCSS Cox model with number of positive nodes are illustrated in Figure 6.3.1-6.3.4 as examples. The residual plots for the other three Cox models are in Appendix A. The deviance residuals plot with a disjunction between censored observations and uncensored observations in Figure 6.3.2 suggests no outliers. The estimated changes in the scaled coefficients as a result of removing each observation from the fitted model in turn are used as measures of influence. In Figure 6.3.3, all

estimated changes are less than 0.6 in absolute value meaning no observation has a large effect on the estimates. The rescaled Schoenfeld residuals and the smooth curves in Figure 6.3.4 and the statistical test results in Table 6.3.1 suggest that the proportional hazard assumption holds for all covariates since the smooth curves are flat at 0, except estrogen receptor status. The rescaled Schoenfeld residuals plot of ER positive suggests a dependency of residuals on time. The relationship between ER status and survival time is further investigated using time-dependent models in Section 6.4.

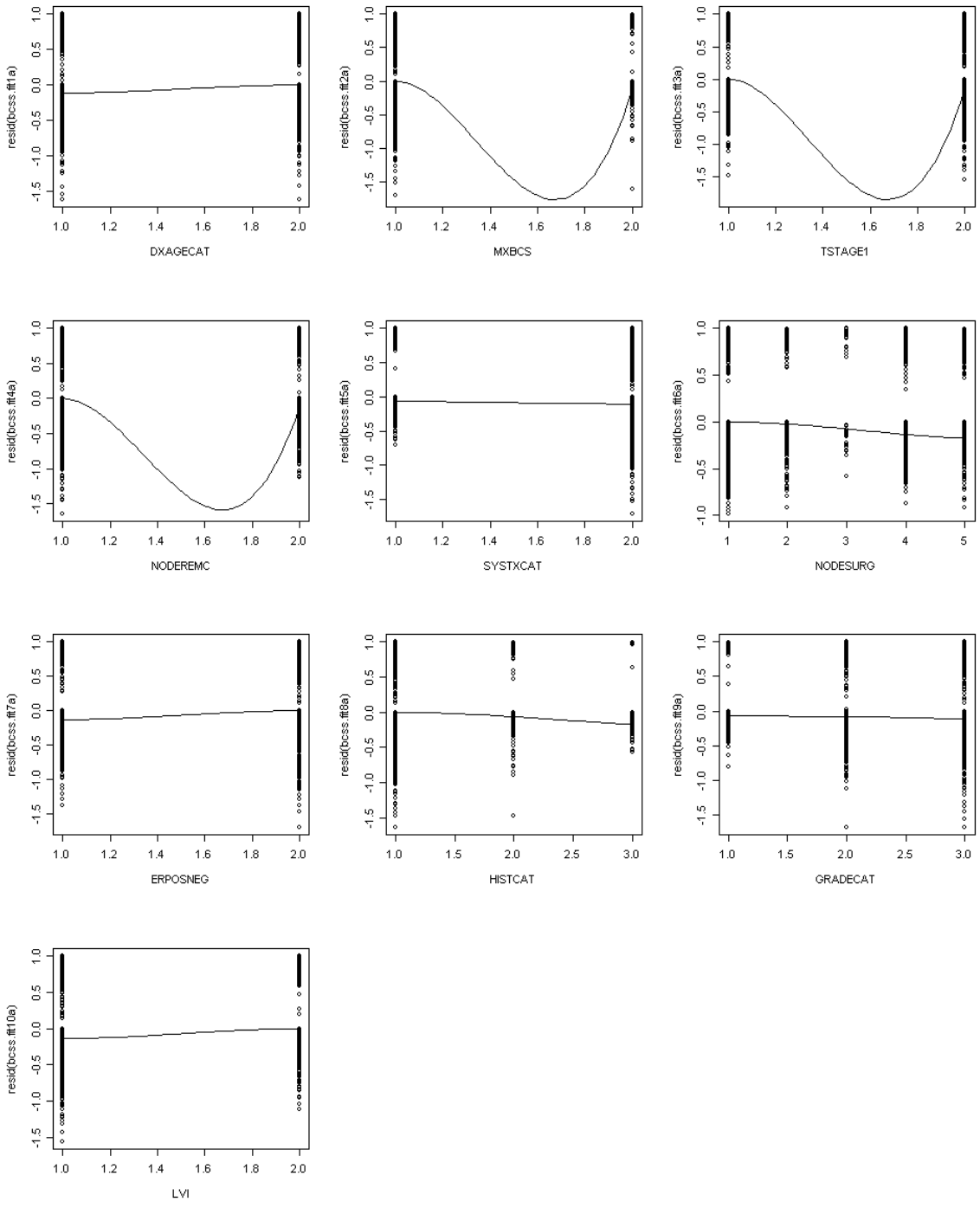


Figure 6.3.1: The martingale residuals plots for the BCSS model with number of positive nodes.

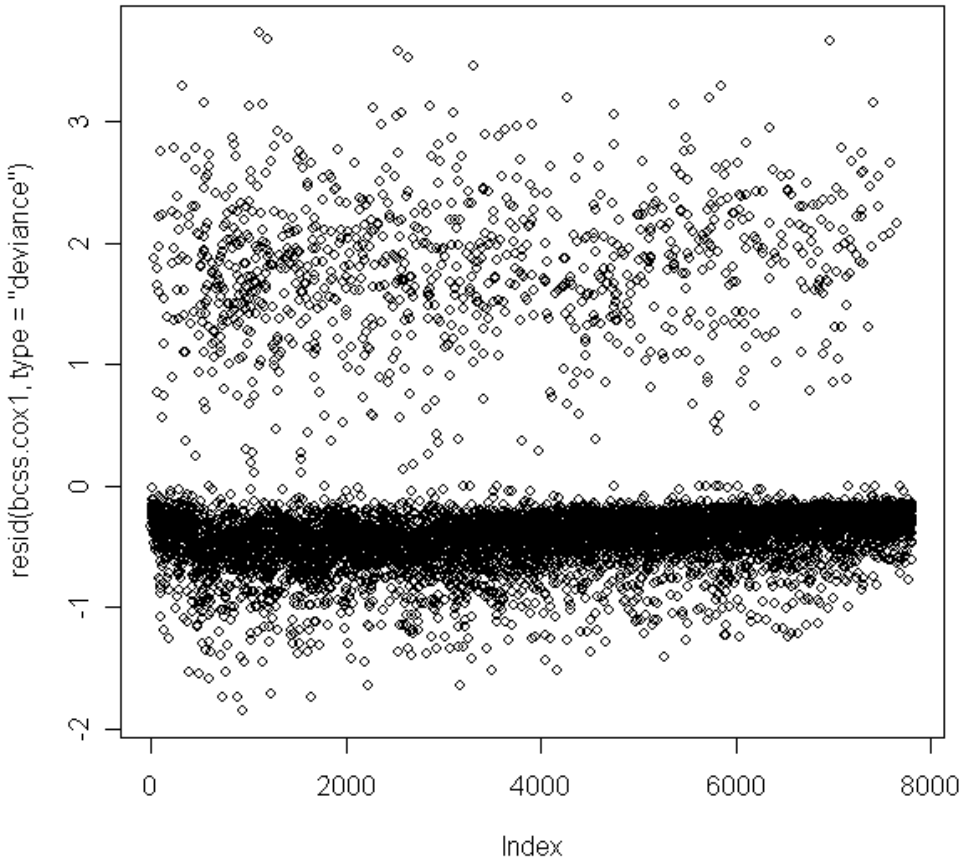
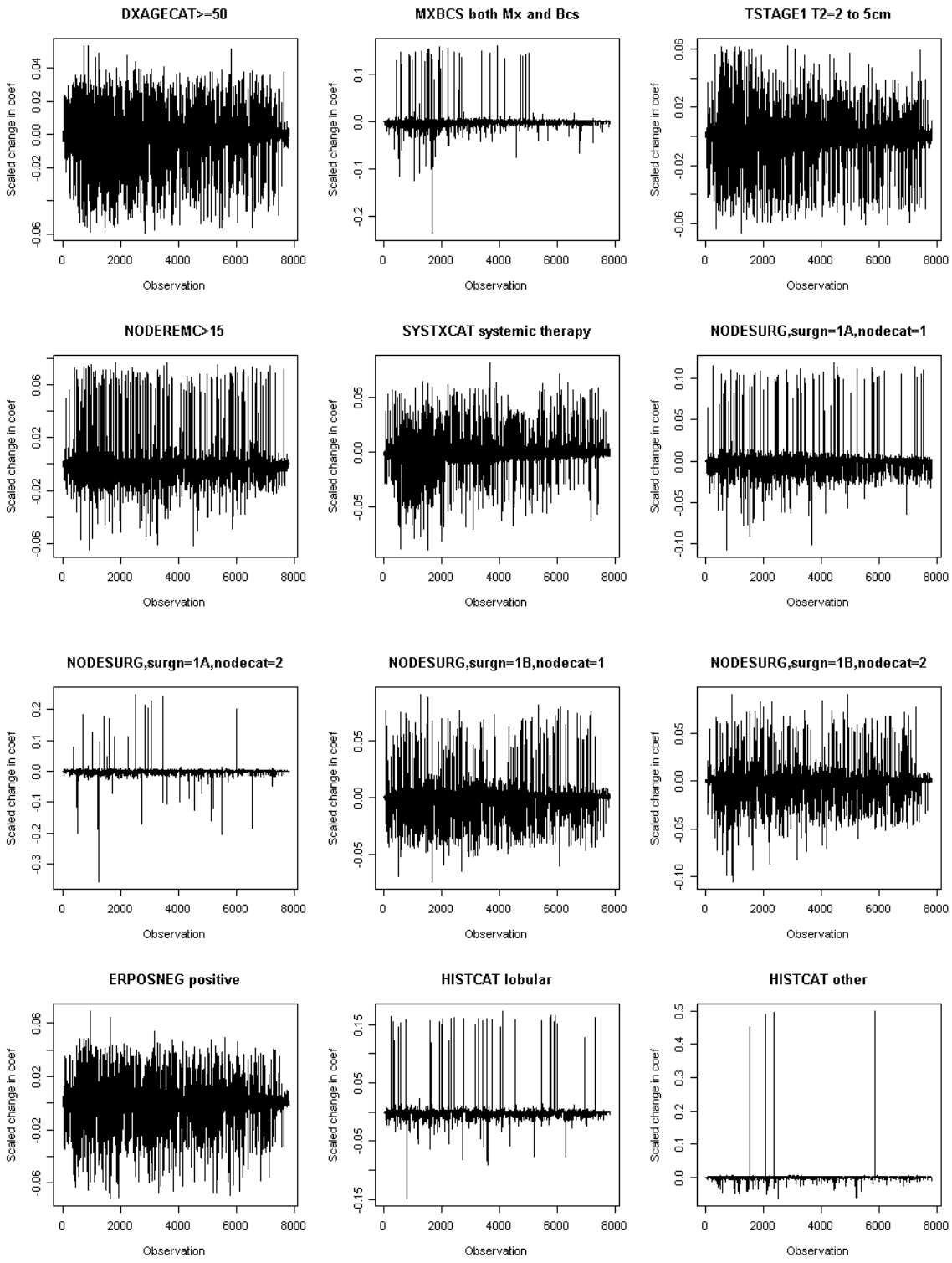


Figure 6.3.2: The deviance residuals plot for the BCSS model with number of positive nodes.


DXAGECAT>=50
MXBCS both Mx and Bcs
TSTAGE1 T2=2 to 5cm
NODEREMC>15
SYSTXCAT systemic therapy
NODESURG,surgn=1A,nodecat=1
NODESURG,surgn=1A,nodecat=2
NODESURG,surgn=1B,nodecat=1
NODESURG,surgn=1B,nodecat=2
ERPOSNEG positive
HISTCAT lobular
HISTCAT other
Scaled change in coef
Observation

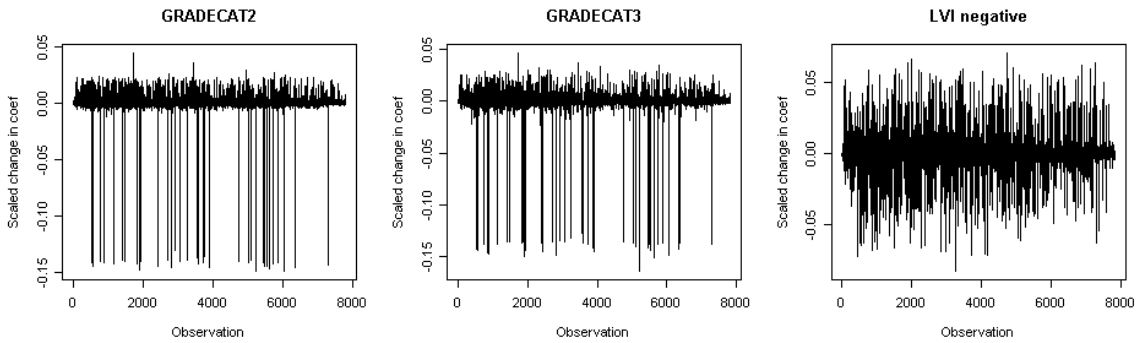
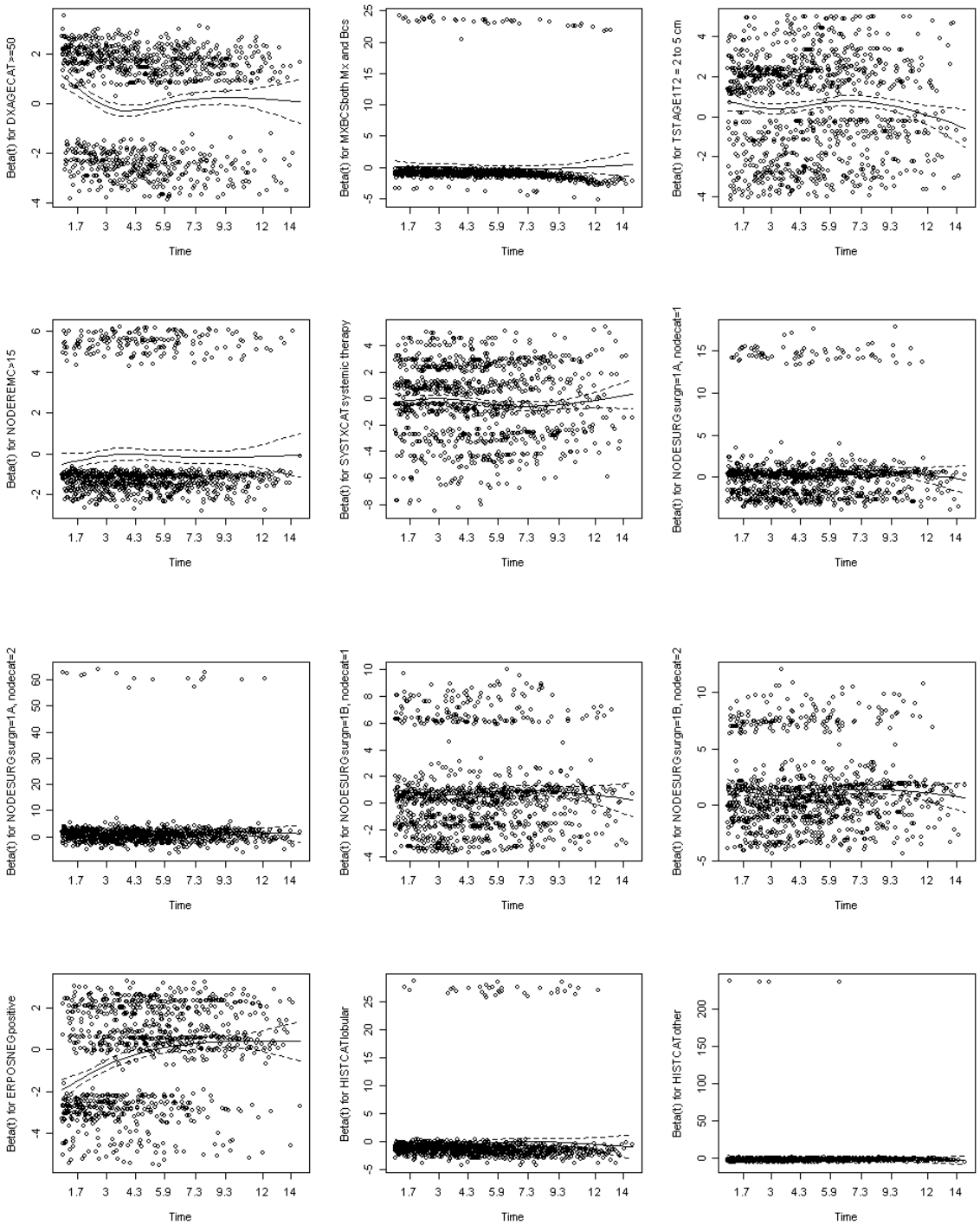


Figure 6.3.3: The influence plots for the fifteen important predictors for the BCSS model with number of positive nodes.


Beta(t) for DXAGECAT>=50
Beta(t) for MXBCSboth Mx and Bcs
Beta(t) for TSTAGE1T2 = 2 to 5 cm
Beta(t) for NODEREMC>15
Beta(t) for SYSTXCAT systemic therapy
Beta(t) for NODESURGsurgn=1A, nodecat=1
Beta(t) for NODESURGsurgn=1A, nodecat=2
Beta(t) for NODESURGsurgn=1B, nodecat=1
Beta(t) for NODESURGsurgn=1B, nodecat=2
Beta(t) for ERPOSNEGpositive
Beta(t) for HISTCATlobular
Beta(t) for HISTCATother
Time
1.7 3 4.3 5.9 7.3 9.3 12 14

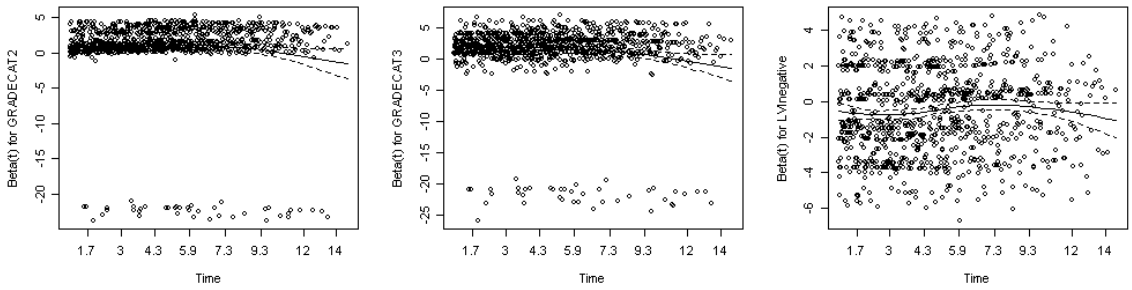


Figure 6.3.4: The rescaled Schoenfeld residuals plots for the BCSS model with number of positive nodes.

Table 6.3.1: Chi-square tests for significant slope in the rescaled Schoenfeld residuals plots in Figure 6.3.4.

	rho	chisq	p
DXAGECAT>=50	-0.035	1.15e+00	0.283
MXBCSboth Mx and Bcs	-0.017	2.76e-01	0.600
TSTAGE1T2 = 2 to 5 cm	-0.014	2.22e-01	0.638
NODEREMC>15	0.016	2.44e-01	0.622
SYSTXCATsystemic therapy	-0.040	1.66e+00	0.197
NODESURGsurgn=1A, nodecat=1	-0.003	8.33e-03	0.927
NODESURGsurgn=1A, nodecat=2	0.000	9.72e-05	0.992
NODESURGsurgn=1B, nodecat=1	0.003	9.13e-03	0.924
NODESURGsurgn=1B, nodecat=2	0.011	1.27e-01	0.721
ERPOSNEGpositive	0.275	7.41e+01	0.000
HISTCATlobular	0.046	2.00e+00	0.157
HISTCATother	-0.030	8.62e-01	0.353
GRADECAT2	-0.041	1.59e+00	0.208
GRADECAT3	-0.071	4.80e+00	0.028
LVInegative	0.051	2.65e+00	0.104
GLOBAL	NA	1.35e+02	0.000

6.4 Time-dependent Covariate

In most cases, the covariates for the Cox PH models are time-independent for each patient, but covariates that change over time can be encountered. A discrete time-dependent covariate usually has value 0 until some intermediate event occurs, then it becomes 1 [17]. A continuous time-dependent covariate takes on a series of measurements of certain explanatory characteristics over time, such as blood pressure and body mess index [17]. Including a time-dependent covariate provides new ways to explore potential associations, but also raises problems such as difficulties in the choice of covariate form (a function of time); potential for bias, erroneous inference and over-fitted modeling; lacking some of the properties of Cox models with fixed (time-independent) covariates; losing the power of estimating survival curves; and prediction not related to the usual hazard function due to strong association with the study unit [9].

Proportionality is one of the main assumptions of the Cox PH model. A significant time-dependent covariate suggests violation of the proportionality assumption for that specific predictor. The rescaled Schoenfeld residuals plot of ER positive from Section 6.3 suggests a dependency of residuals on time. Estrogen receptor (ER), a hormone activated receptor, has been used as an indicator of endocrine responsiveness, then as a prognostic factor for early recurrence in clinical breast cancer management [7]. ER is considered to be an important predictor of response to endocrine therapy, but its prognostic relevance is often contradicted [18]. Coradini et al. [7] conducted a study on 1,793 axillary lymph node negative breast cancer patients with a 10 year-follow up. Their results suggested “ER content failed to show a prognostic effect within the first years of follow-up; thereafter, a positive association with risk of relapse was observed”. More

specifically, ER content was positively associated with increased hazard ratio at 72 months.

There are various ways to deal with a time-dependent or non-proportional covariate. We can consider a parametric regression model (e.g. Weibull, log-logistic, log-normal) instead of the Cox PH model, include the non-proportional covariate as a function of time, or stratify on the time-dependent covariate [28]. In order to include the time-dependent covariate, let $Z(t) = [Z_1(t), Z_2, ..., Z_p]'$ represent a vector of covariates at time *t*. The hazard function for the Cox PH model is

$$h[t \mid Z(t)] = \exp[\beta_1 Z_1(t) + \beta_2 Z_2 \cdots + \beta_p Z_p] \cdot h_0(t) = \exp[\beta' Z(t)] \cdot h_0(t),$$

where $h_0(t)$ is an arbitrary baseline hazard rate that is the same for all individuals and $\beta = (\beta_1, ..., \beta_p)'$ is a parameter vector. The determination of the functional form of the $Z_1(t)$ usually requires biological understanding or biological hypothesis of the covariate.

In our study, the proportional hazard assumption holds for all covariates except estrogen receptor status. The rescaled Schoenfeld residuals plots for ER positive status in BCSS Cox PH model with number of positive nodes are shown in Figure 6.4.1. The rescaled Schoenfeld residual is defined as $S_{ij}(\beta) = Z_{ij}(t_i) - \overline{Z}(\beta, t_i)$. The residuals plots for the other three models (OS with number of positive nodes, BCSS and OS with LNR) are very similar, therefore they are not shown here. The slope of the smooth curves does not equal zero, indicating time-dependency or violation of the proportional assumption. The smooth curve of residuals increases from -2 to 0 in about the first 5.7 years and then remains constant. This is consistent with the clinical knowledge about ER positive patients that the benefit of being ER positive declined at about 5 years [24].

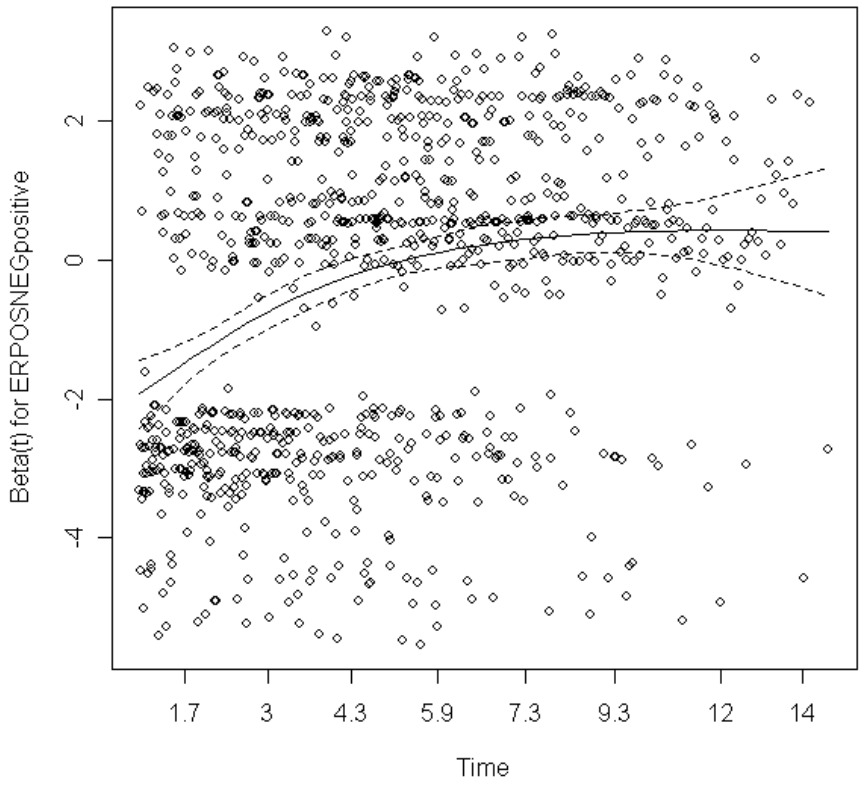


Figure 6.4.1: The rescaled Schoenfeld residuals plot for ER positive status in the BCSS Cox PH model with number of positive nodes.

To account for non-proportional ER, we stratified by ER status in all four Cox PH models for analysis. Hazard ratio estimates from the stratified models were consistent with results in Table 6.2.1 for unstratified models. As another check on the results, we fit log-normal regression models using the prognostic variables in Table 6.2.1. All parameter estimates were consistent with those given in Table 6.2.1. In the attempt to decide the functional form of the time-dependent ER status, we fit an adjusted normal cumulative distribution function to the residuals plot. Figure 6.4.2 is an example for BCSS Cox PH model with number of positive nodes, where the solid line is a smoother using locally-

weighted polynomial regression and the dotted line takes the form of $y = 4\Phi\left(\frac{2x}{7}\right) - 4$.

We compared the four Cox PH models with and without adding the time-dependent covariate in SAS and the estimates for all parameters were consistent. Treating missing values as a separate category did not change the results from the analyses above. We concluded that the covariate effects were robust to model specification.

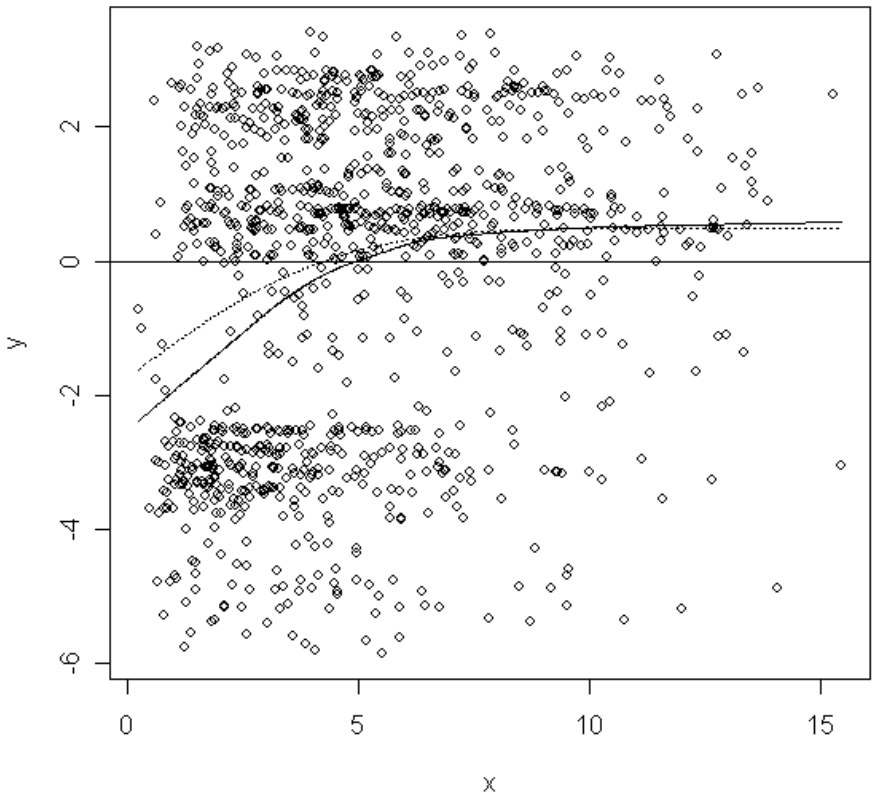


Figure 6.4.2: The rescaled Schoenfeld residuals plot for ER positive status in the BCSS Cox PH model with number of positive nodes. The solid line is a smoother and the dotted line takes the form of $y = 4\Phi\left(\frac{2x}{7}\right) - 4$.

6.5 Comparison of Unadjusted and Adjusted Survival

In this section, we compare unadjusted and adjusted survival curves. The two main methods to generate adjusted survival curves from the Cox proportional hazards models are the mean of covariates method and the corrected group prognosis method [25]. In the first method, the mean values of all covariates are computed and a survival curve is estimated for those values of the covariates [11]. This method is considered problematic by some [11]. For instance, if the sex of patients is coded as 0 and 1, and 50% of patients are female; the average covariate value 0.5 would be meaningless at the individual level [25].

The corrected group prognosis method was considered to produce more intuitive results. In this method, survival curves for each of the unique covariate combinations are first calculated based on the coefficients of covariates of interest from a single PH model generated from the entire data set [11]. Then, a weighted average of these individual survival curves is computed in which weights are proportional to the number of individuals at each level of the covariates in the entire sample [11]. Ghali et al. [11] found that the adjusted survival curves for patients undergoing cardiac catheterization with and without diabetes generated by the corrected group prognosis methods were more appropriately positioned between the unadjusted curves compared to the adjusted survival curves generated using the mean of covariates method. The mean of covariates method averages the PH survival function's exponents while the corrected group prognosis method averages the actual survival curves [11]. If only one covariate is controlled at a time, the distortion of the mean of covariates method was the largest when the controlled covariate was prevalent in the data set and its associated hazard ratio(s) was large [11]. In

our study, adjusted KM survival curves were computed using an S_PLUS/R routine available at http://stat.ubc.ca/~rollin/ based on the four Cox PH models used in previous sections.

Unadjusted survival curves for BCSS and OS by nodal status with number of positive nodes or LNR are shown in Figure 6.5.1 to 6.5.4. Survival curves were plotted against number of years from diagnosis to death due to breast cancer or October 31st 2004 if alive for BCSS and to death of any causes or October 31st 2004 if alive for OS. Adjusted survival curves for BCSS and OS with and without cases with missing values removed by nodal status were shown in Figures 6.5.5 to 6.5.12. When cases with missing values were retained, the adjusted survival curves for all models were similar to those for which cases with missing values were removed.

Patients in the pN1a and pN1b subgroups with smaller number of positive nodes and LNR had worse survival than patients in the pN0 subgroups and better survival than patients in the pN1a and pN1b subgroups with larger number of positive nodes and LNR for all survival plots. In the case of unadjusted OS with LNR, the survival curve for pN1a with smaller LNR values was very close to the curve for pN0. The survival curves for pN1a and pN1b subgroups with larger number of positive nodes and LNR were overlapped for all four unadjusted models.

Among patients with smaller positive number of nodes and LNR, patients in the pN1a subgroup had better survival than patients in the pN1b subgroup in all plots. For adjusted BCSS with LNR with and without missing values in particular, the survival curves for pN1a and pN1b subgroups with smaller LNR were similar.

Among all adjusted survival curves with larger number of positive nodes and LNR, the survival curves for pN1b were above pN1a. However, they were indistinguishable except for BCSS with number of positive nodes with and without cases with missing values (Figure 6.5.5 and Figure 6.5.9). In these two plots, patients in the pN1b subgroup with larger number of positive nodes had better survival than pN1a patients with larger number of positive nodes.

Table 6.5.1 contains adjusted five and ten-year survival with and without missing values treated as a separate category and unadjusted five and ten-year survival with cases with missing values removed for all four models. Unadjusted survival with missing values retained was not considered because there were no missing values in *nodesurg* and *lnr25surg*. Adjusted survival estimates with and without missing values treated as a separate category were similar. Comparing ten-year survival estimates between adjusted and unadjusted models, survival estimates for the pN0 subgroup were similar for all four models. Unadjusted and adjusted ten-year BCSS were similar only for the pN1a subgroup with smaller number of positive nodes and LNR. Adjusted and unadjusted ten-year OS were similar for all categories expect pN1b with larger number of positive nodes and LNR.

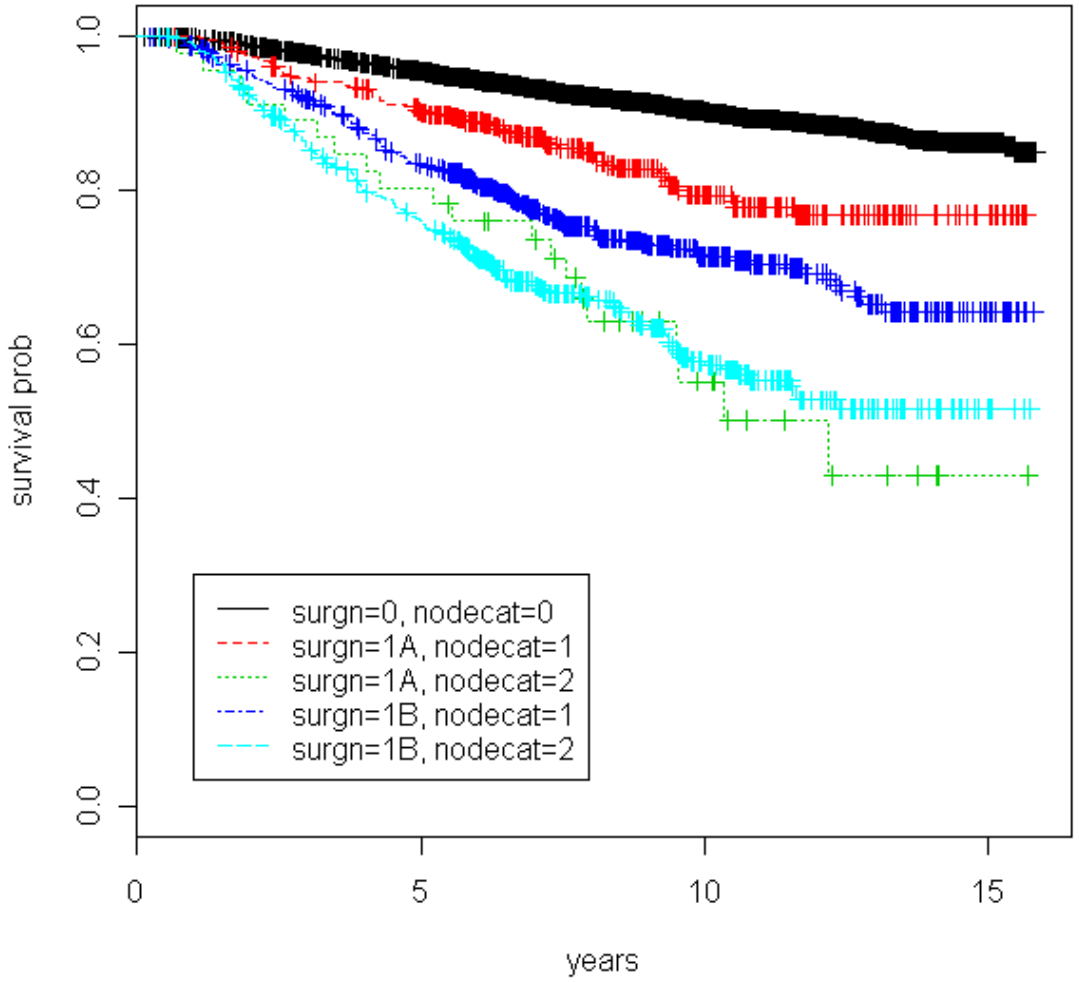


Figure 6.5.1: Unadjusted KM BCSS curves for the nodal subgroups with number of positive nodes. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively. *Nodecat* equals 0, 1 and 2 represent 0, 1-4 and ≥4 number of positive nodes respectively.

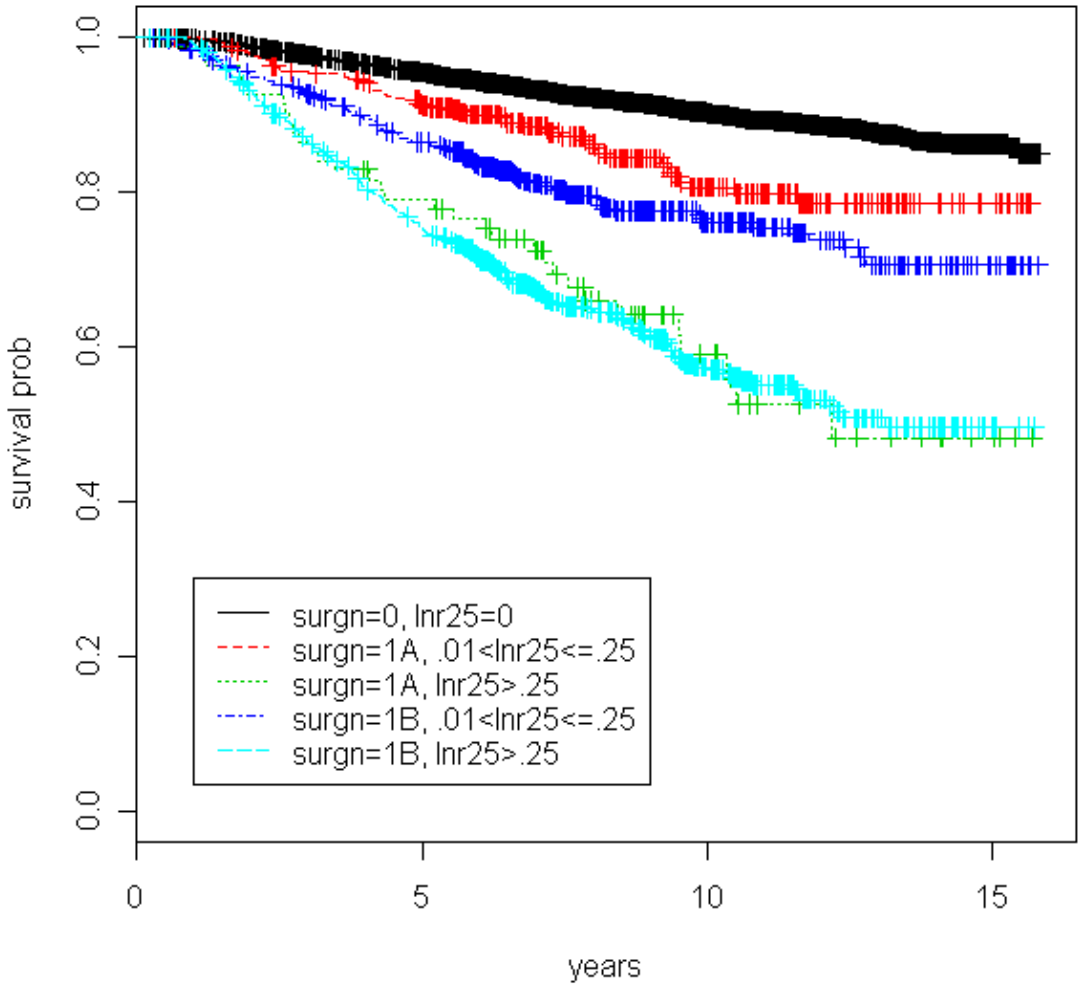


Figure 6.5.2: Unadjusted KM BCSS curves for the nodal subgroups with LNR. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

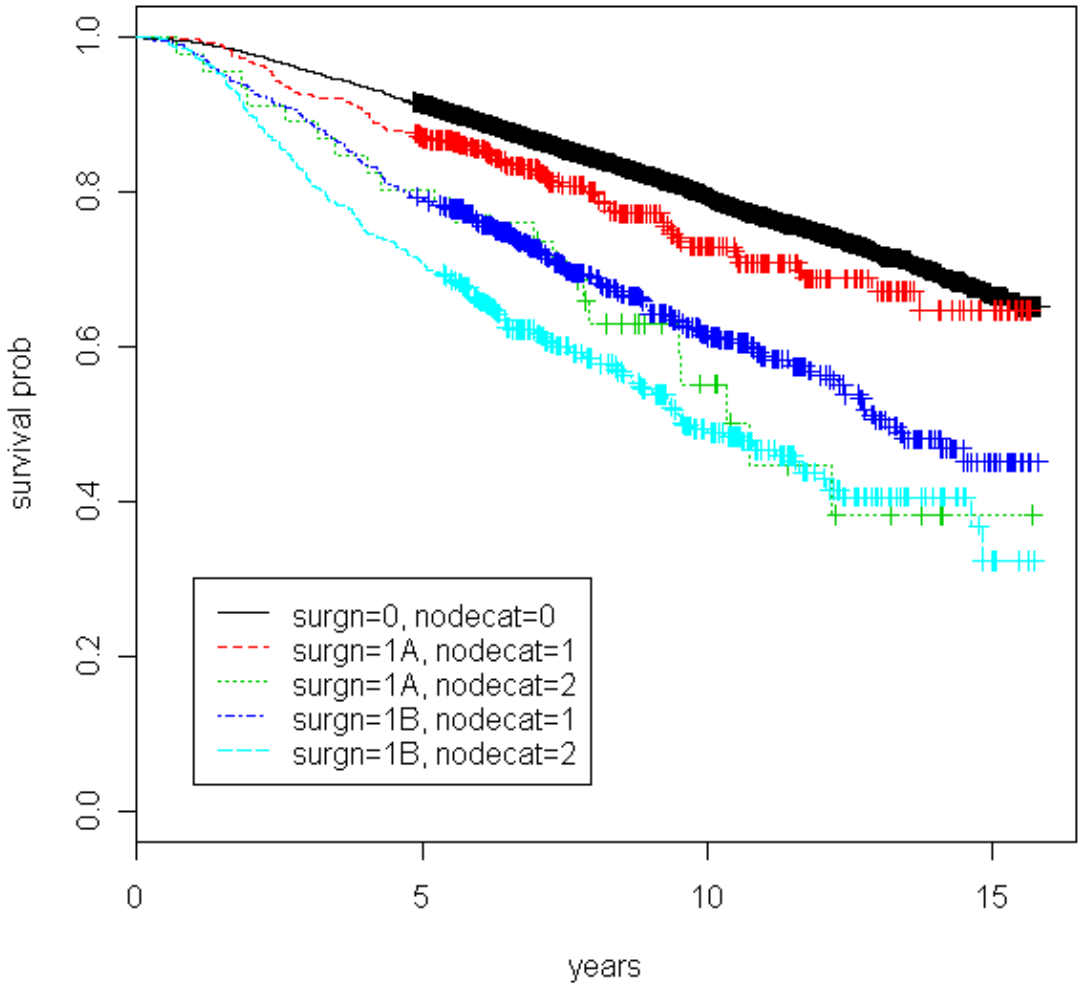


Figure 6.5.3: Unadjusted KM OS curves for the nodal subgroups with number of positive nodes. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively. *Nodecat* equals 0, 1 and 2 represent 0, 1-4 and ≥4 number of positive nodes respectively.

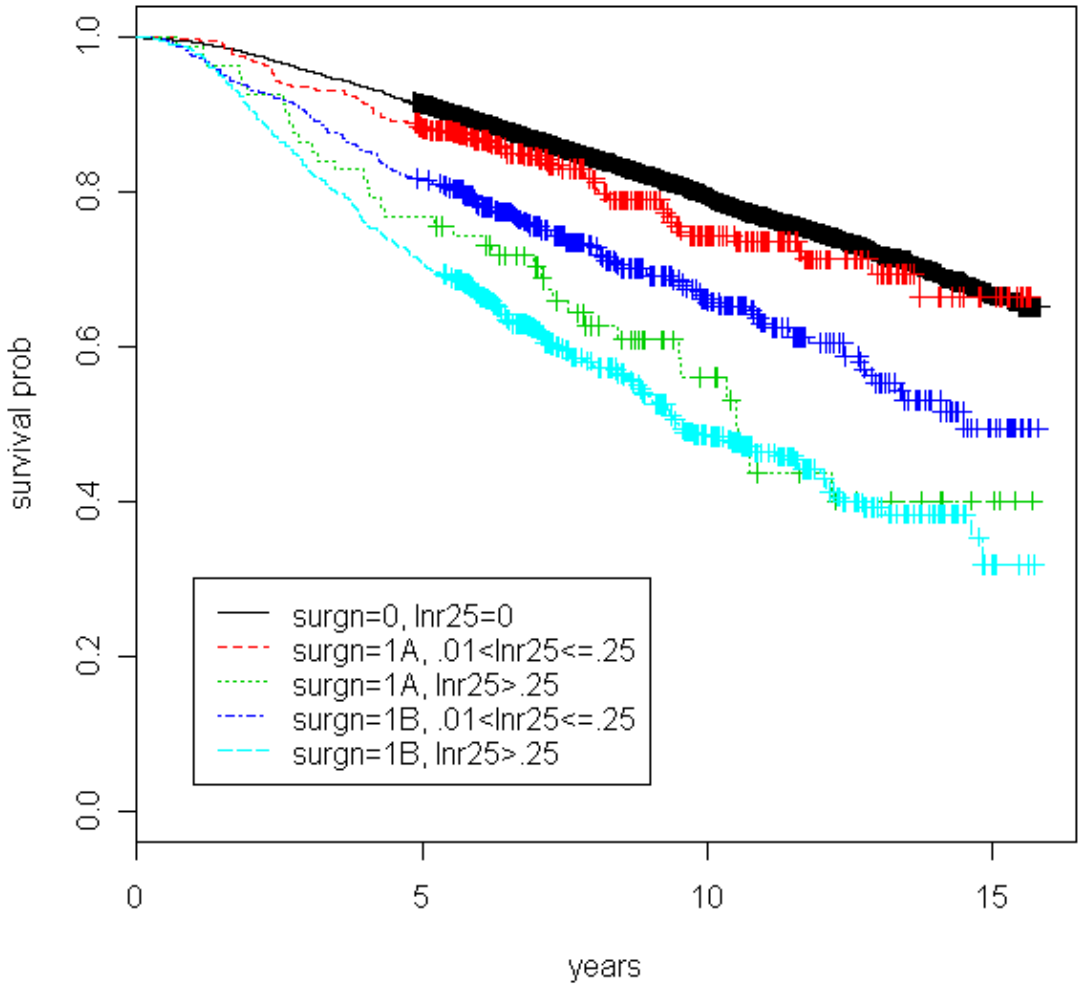


Figure 6.5.4: Unadjusted KM OS curves for the nodal subgroups with LNR. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

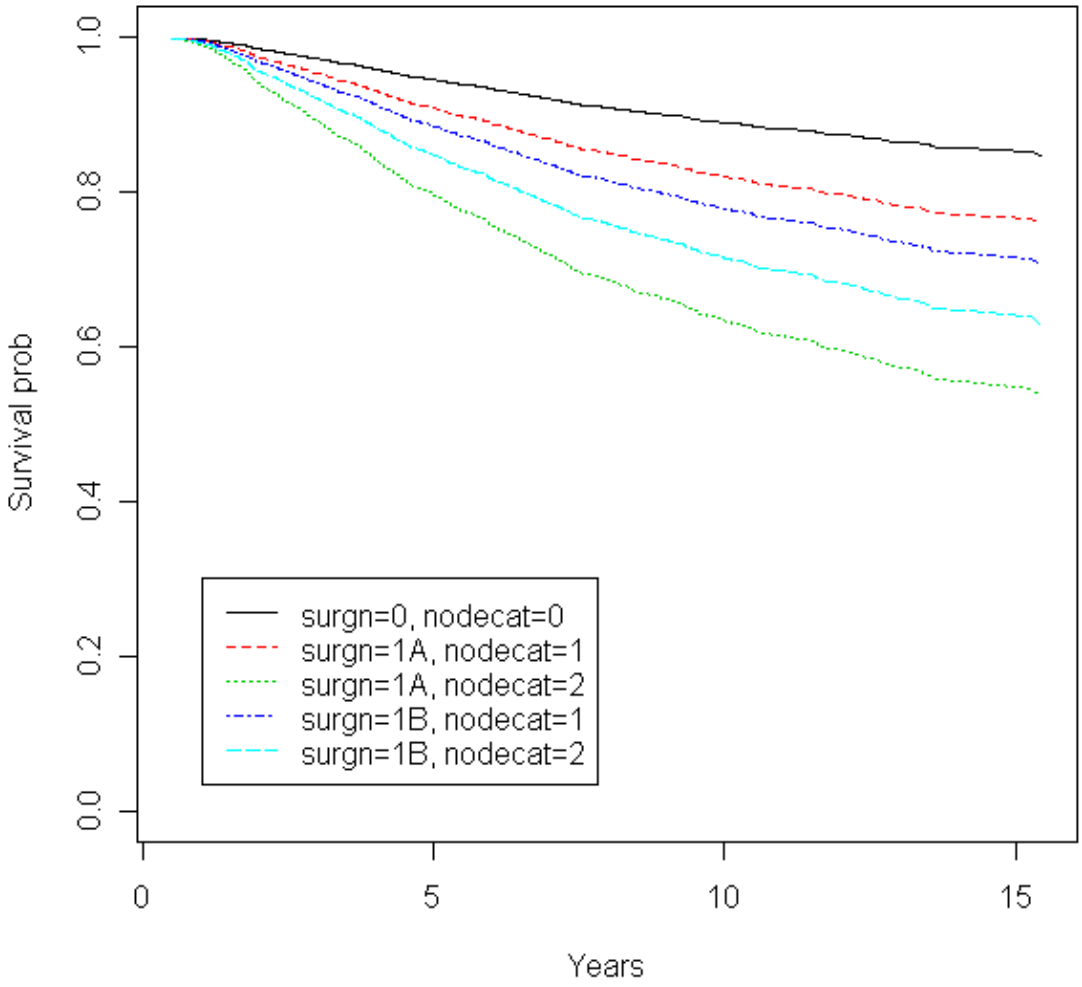


Figure 6.5.5: BCSS curves for the nodal subgroups with number of positive nodes adjusted for other prognostic variables with missing values removed. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively. *Nodecat* equals 0, 1 and 2 represent 0, 1-4 and ≥4 number of positive nodes respectively.

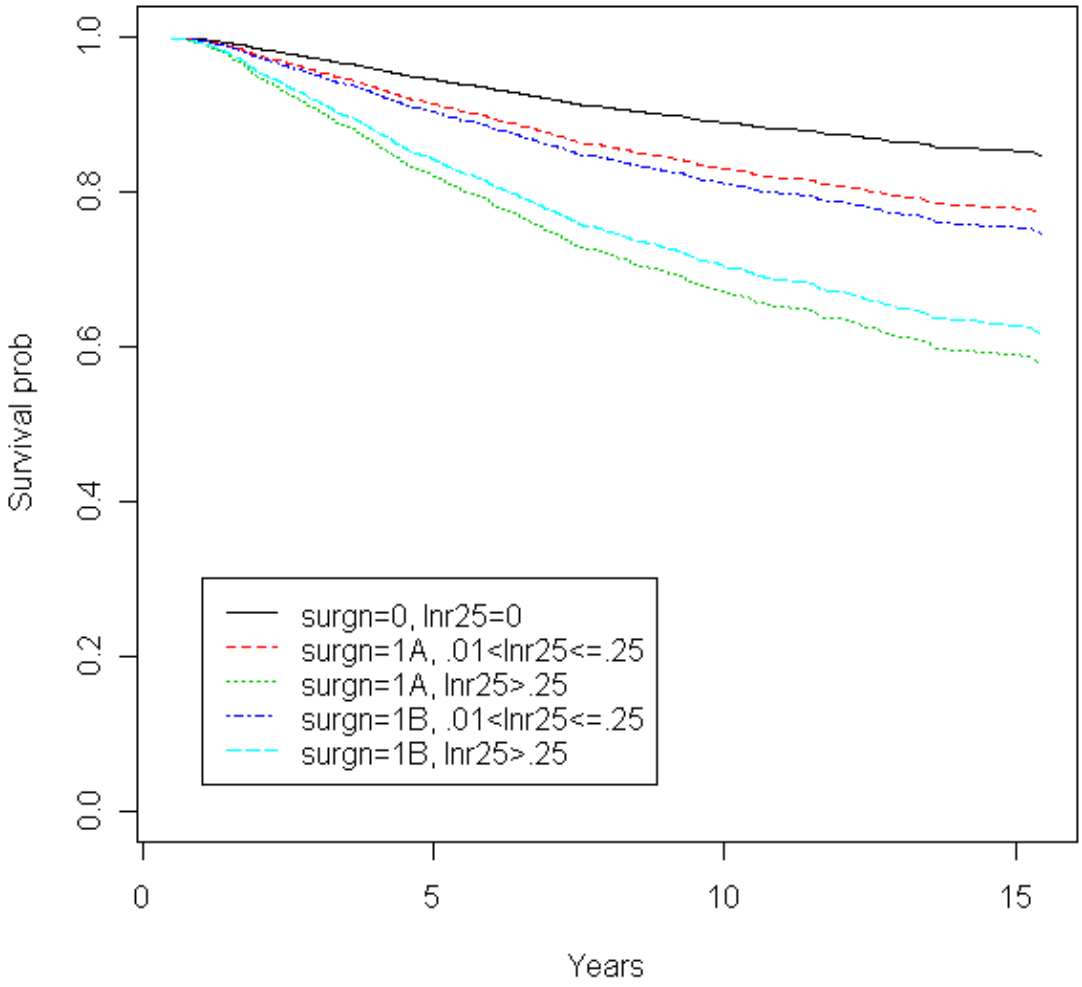


Figure 6.5.6: BCSS curves for the nodal subgroups with LNR adjusted for other prognostic variables with missing values removed. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

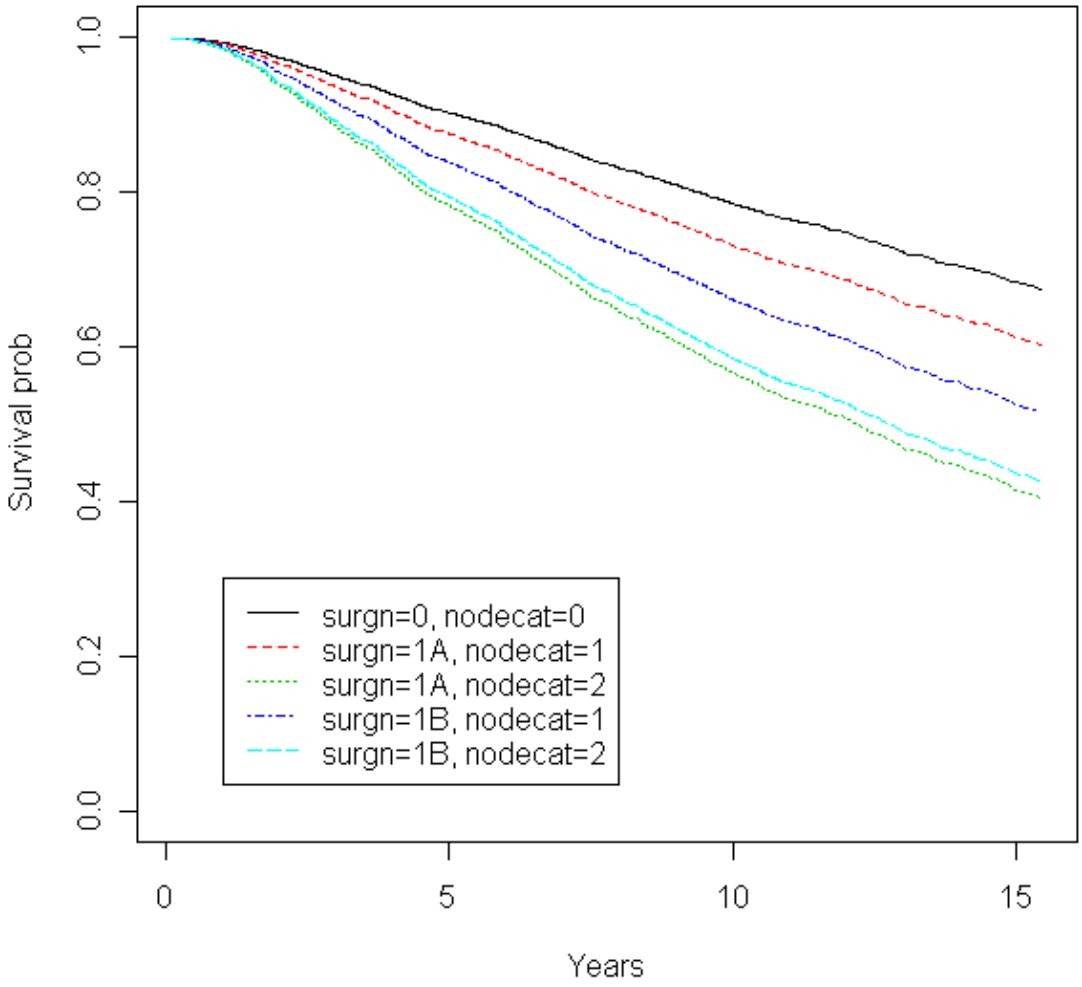


Figure 6.5.7: OS curves for the nodal subgroups with number of positive nodes adjusted for other prognostic variables with missing values removed. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively. *Nodecat* equals 0, 1 and 2 represent 0, 1-4 and ≥4 number of positive nodes respectively.

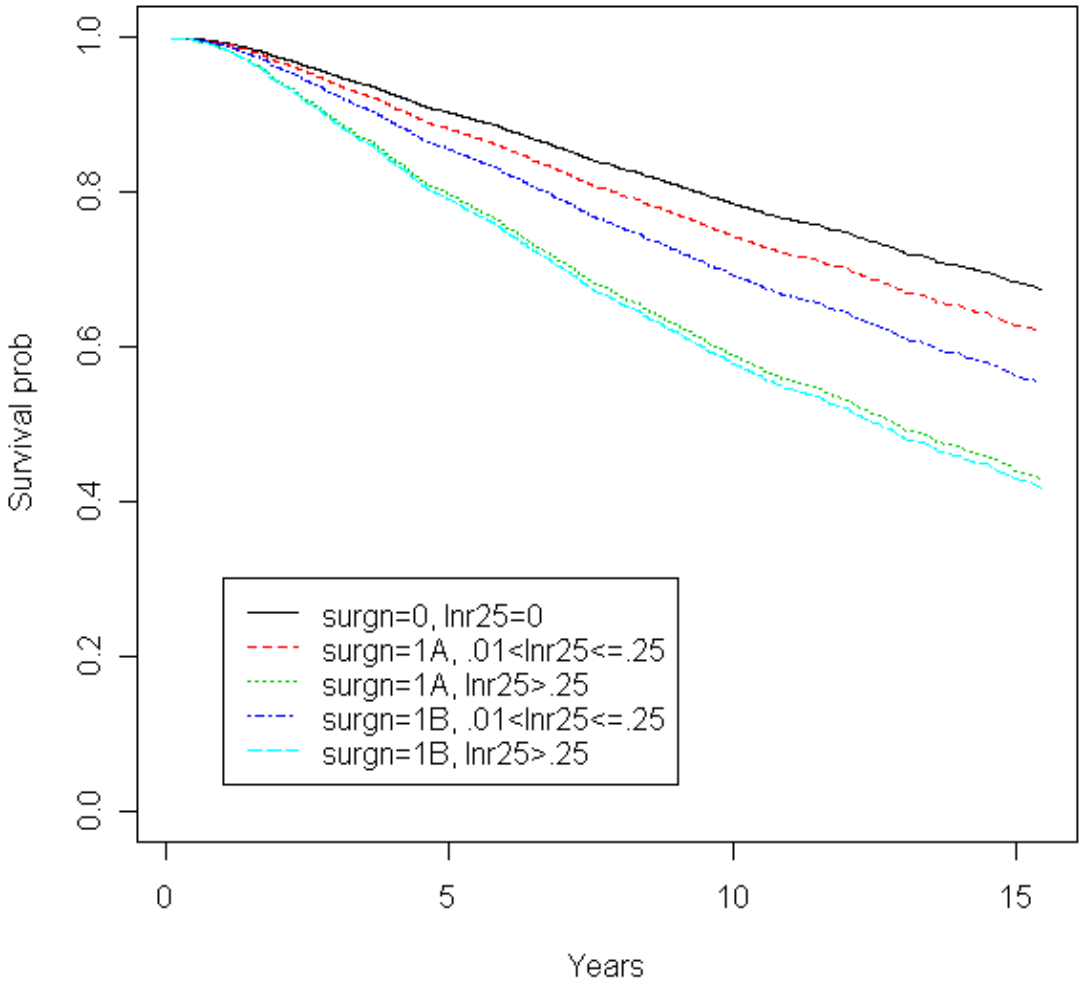


Figure 6.5.8: OS curves for the nodal subgroups with LNR adjusted for other prognostic variables with missing values removed. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

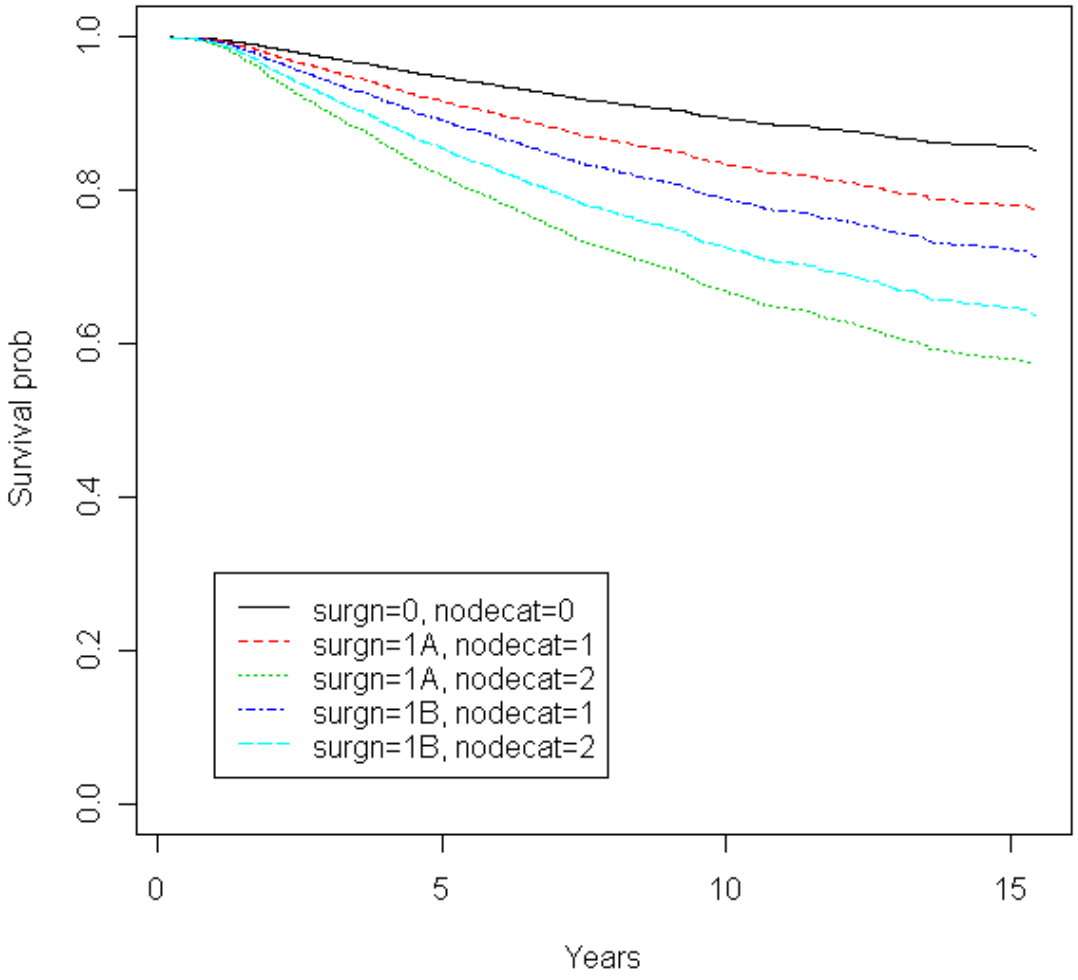


Figure 6.5.9: BCSS curves for the nodal subgroups with number of positive nodes adjusted for other prognostic variables with missing values retained. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively. *Nodecat* equals 0, 1 and 2 represent 0, 1-4 and ≥4 number of positive nodes respectively.

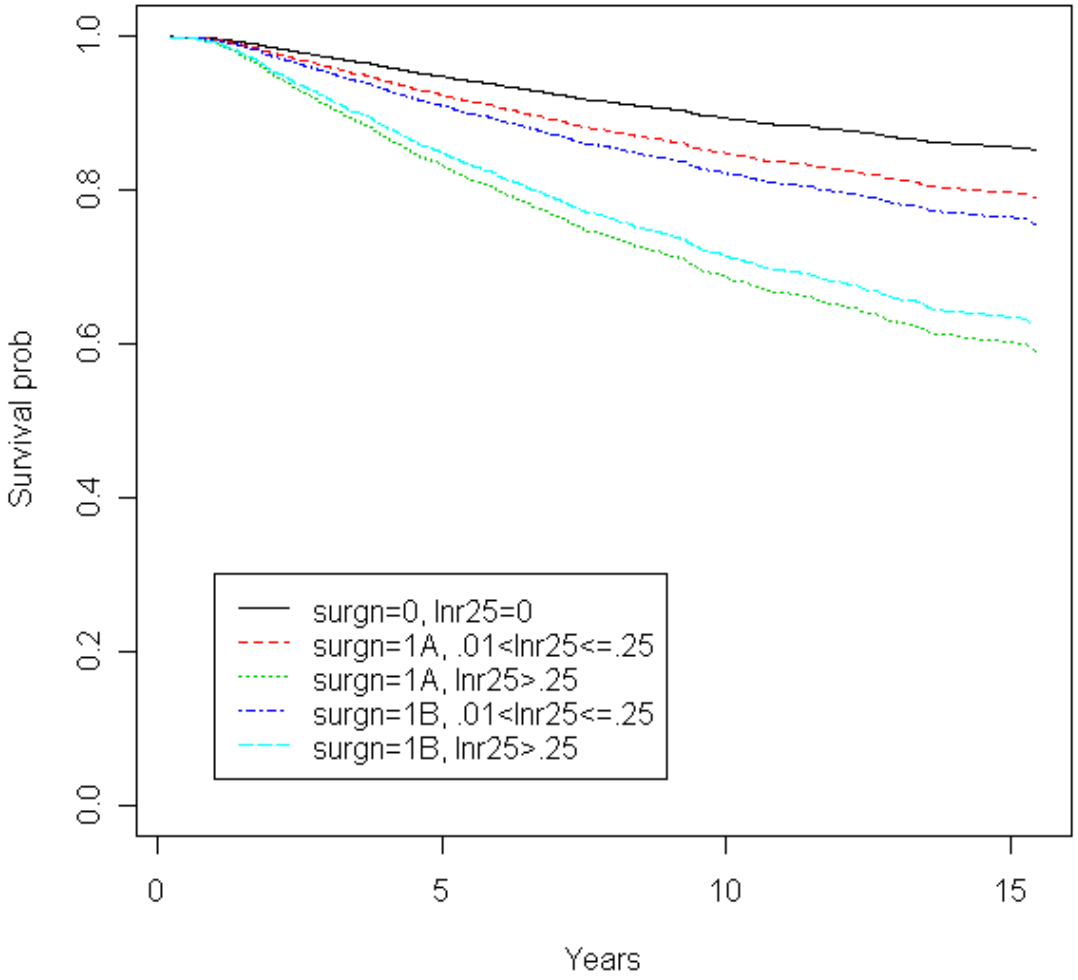


Figure 6.5.10: BCSS curves for the nodal subgroups with LNR adjusted for other prognostic variables with missing values retained. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

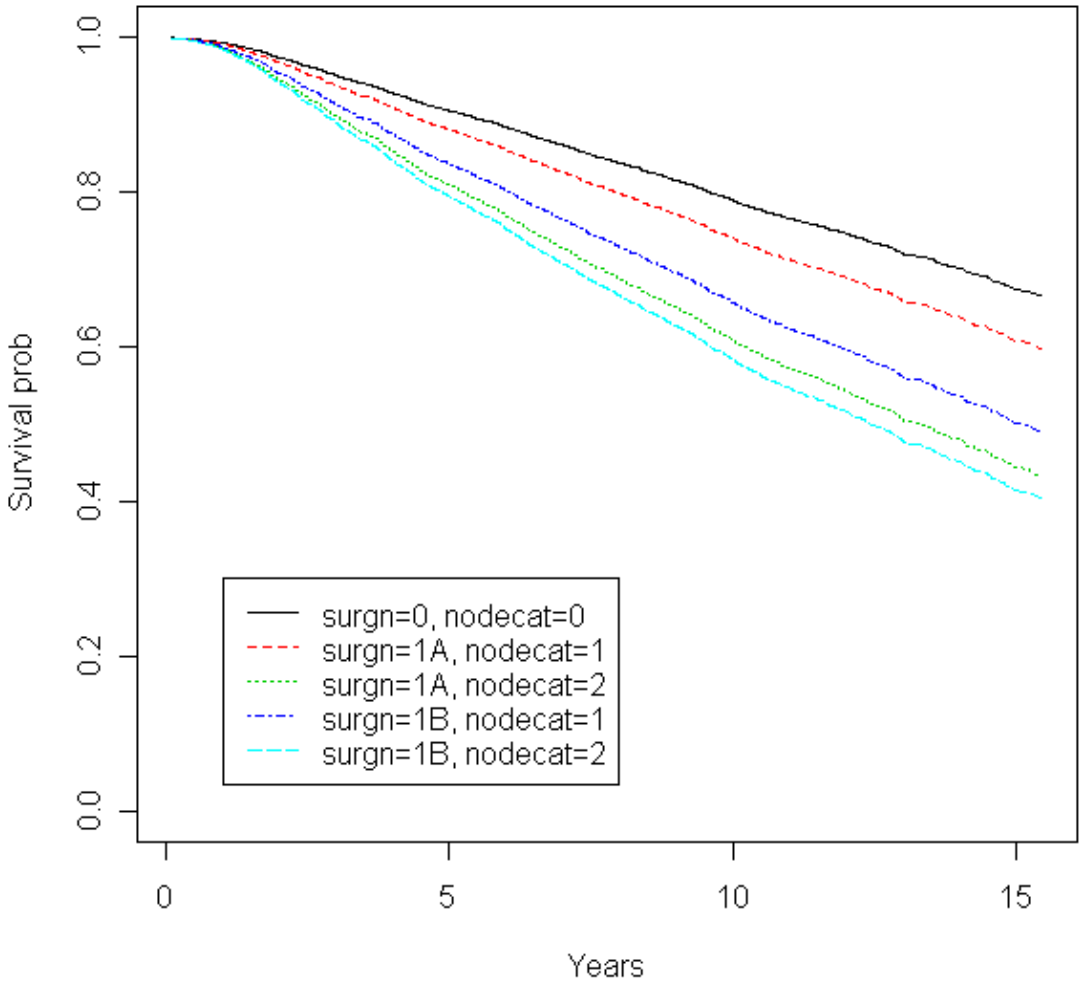


Figure 6.5.11: OS curves for the nodal subgroups with number of positive nodes adjusted for other prognostic variables with missing values retained. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

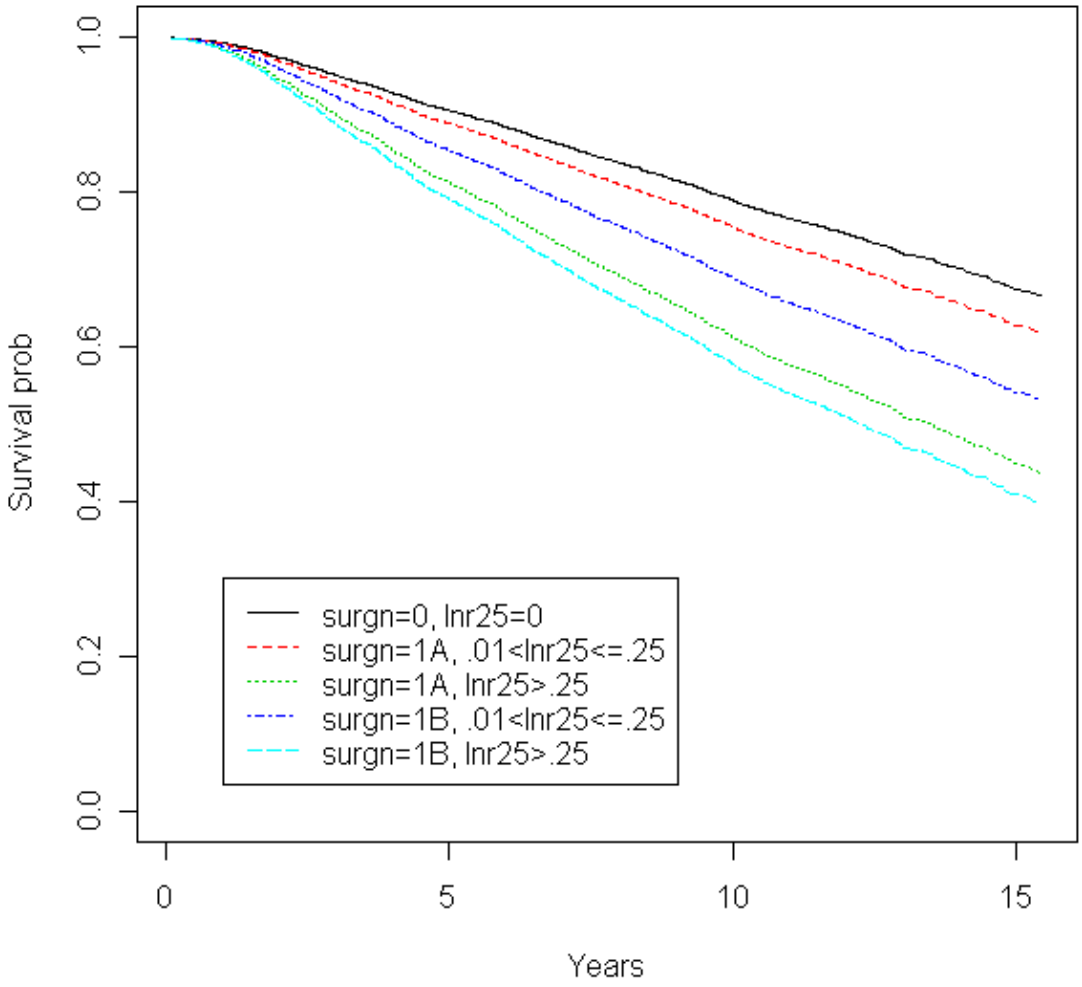


Figure 6.5.12: OS curves for the nodal subgroups with LNR adjusted for other prognostic variables with missing values retained. *Surgn* equals 0, 1A and 1B represent the pN0, pN1a and pN1b subgroups respectively.

Table 6.5.1: Adjusted five and ten-year survival with and without missing values removed and unadjusted KM five and ten-year survival.

Models	Levels	Adjusted				Unadjusted KM	
		With missing values removed		With missing values retained		With missing values removed	
		5-year Survival	10-year Survival	5-year Survival	10-year Survival	5-year Survival	10-year Survival
BCSS with # of positive nodes	surgn=0, nodecat=0	.9459	.8900	.9484	.8942	.9554	.9028
	surgn=1A, nodecat=1	.9088	.8203	.9174	.8350	.9014	.7942
	surgn=1A, nodecat=2	.7967	.6338	.8203	.6682	.8043	.5516
	surgn=1B, nodecat=1	.8856	.7789	.8922	.7890	.8335	.7139
	surgn=1B, nodecat=2	.8481	.7151	.8555	.7256	.7601	.5785
BCSS with LNR	surgn=0, lnr25=0	.9460	.8900	.9485	.8942	.9554	.9028
	surgn=1A, .01<lnr25<=.25	.9145	.8304	.9247	.8484	.9126	.8065
	surgn=1A, lnr25>.25	.8210	.6706	.8327	.6873	.7916	.5899
	surgn=1B, .01<lnr25<=.25	.9040	.8113	.9108	.8224	.8649	.7613
	surgn=1B, lnr25>.25	.8418	.7042	.8494	.7148	.7535	.5740
OS with # of positive nodes	surgn=0, nodecat=0	.9031	.7856	.9060	.7900	.9143	.7943
	surgn=1A, nodecat=1	.8759	.7317	.8815	.7408	.8695	.7298
	surgn=1A, nodecat=2	.7830	.5669	.8099	.6093	.8043	.5516
	surgn=1B, nodecat=1	.8385	.6620	.8375	.6579	.7896	.6160
	surgn=1B, nodecat=2	.7944	.5856	.7951	.5842	.7091	.4934
OS with LNR	surgn=0, lnr25=0	.9032	.7855	.9061	.7900	.9143	.7943
	surgn=1A, .01<lnr25<=.25	.8823	.7439	.8892	.7559	.8824	.7435
	surgn=1A, lnr25>.25	.7971	.5896	.8125	.6134	.7683	.5606
	surgn=1B, .01<lnr25<=.25	.8557	.6931	.8548	.6893	.8156	.6627
	surgn=1B, lnr25>.25	.7908	.5792	.7917	.5782	.7089	.4868

6.6 Survival Trees

Tree-based modeling is an exploratory technique used in both classification and regression problems to uncover data structure. The models are easy to interpret and useful for screening variables, devising prediction rules, assessing the adequacy of linear models, and summarizing large multivariable data sets [27]. In the regression tree-based model, there is a response variable (y) and a set of predictor variables (x). The regression rules for prediction are determined by recursive partitioning. A predictor variable category is selected at each split to divide the data set into two partitions.

A regression tree is constructed by a collection of rules displayed as a binary tree. A root and a leaf represent the top node and a terminal node of a tree. A split is a rule for creating new branches or creating two daughter nodes from one node. In order to grow a tree, data are split at each node recursively by the binary partitioning algorithm until the node is homogeneous or contains too few observations [27]. The number of allowable splits is one less than the number of its distinctly observed values for an ordinal or continuous variable and $2^{k-1}-1$ for a nominal variable with k levels [31]. The goodness of each allowable split for each variable in terms of homogeneity or purity can be calculated for each selection decision. Gordon and Olshen's Rule suggests that a node is considered pure if all observations are censored or all failures in the leaf occur at the same time followed by censored times or not censored observations [31]. Likelihood-based selection rules choose the split that maximizes the sum of the log likelihoods from the two daughter nodes [31]. The log-rank test is commonly used to test the significance of the difference between the survival times of two groups.

Ten-year survival trees for BCSS and OS with all the variables used in the Cox PH model for cases with missing values retained are shown in Figure 6.6.1 and 6.6.3. The R function "rpart" (Recursive Partitioning and Regression Trees) was used to produce the survival trees. The algorithm "complexity parameter (cp)" was selected for our study, meaning "any split that does not decrease the overall lack of fit by a factor of cp is not attempted" [10]. This option saves computing time by pruning off splits that are not worthwhile. We used cp = 0.02 in our study. Therefore, any split that did not improve the fit by 0.02 will be pruned off by cross-validation.

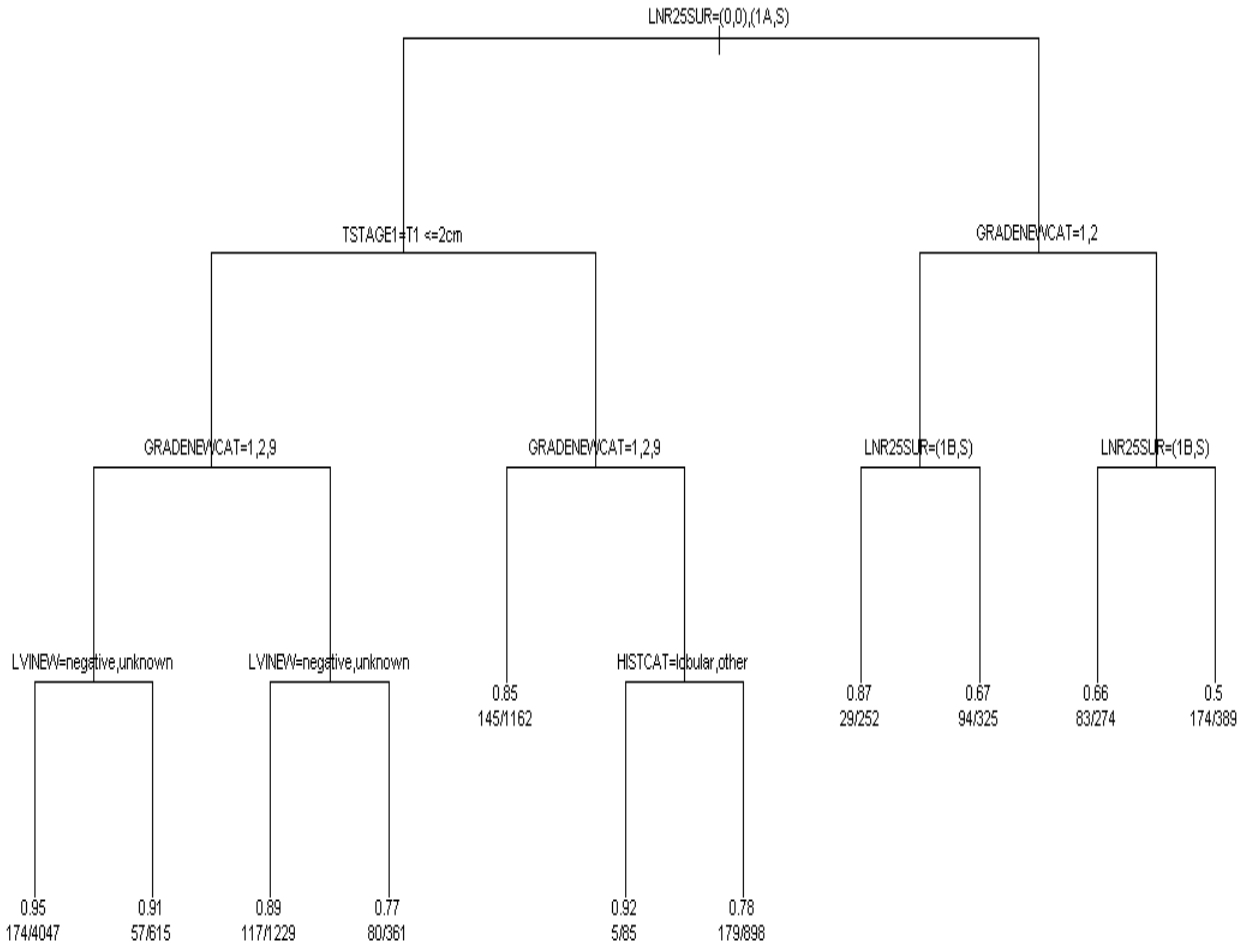
Both the BCSS tree and OS tree selected the variable combining the nodal subgroups and LNR with 0.25 cut point as the first split. In the first separation, the pN0 subgroup and the pN1a subgroup with smaller LNR were grouped together while the pN1a subgroup with lager LNR and the pN1b subgroup with both smaller and larger LNR were grouped together. The following two paragraphs interpret ten-year survival rates along different paths among patients in the pN0 subgroup and pN1a subgroup with smaller LNR, which are shown on the left hand side of both survival trees after the first split.

Based on the BCSS survival tree, patients with T stage 1, grade 1, 2, and unknown, LVI negative and unknown had the highest ten-year BCSS that was 0.95. Ten-year BCSS for patients with T stage 1, grade 1, 2, and unknown, LVI positive was 0.91; for patients with T stage 1, grade 3, LVI negative and unknown was 0.89; for patients with T stage 1, grade 3, LVI positive was 0.77; for patients with T stage 2, grade 1, 2, and unknown was 0.85; for patients with T stage 2, grade 3, histology type lobular and other was 0.92; for patients with T stage 2 cancer, grade 3, histology type ductal was 0.78.

The highest ten-year OS obtained from the group with age less than 50, grade 1 and 2, T stage 1 was 0.94. The ten-year OS for patients with age less than 50, grade 1 and 2, T stage 2 was 0.85; for patients with age less than 50, grade 3 and unknown was 0.81; for patients with age greater or equal to 50, T stage 1, LVI negative and unknown was 0.8; for patients with age greater or equal to 50, T stage 1, LVI positive was 0.71; for patients with age greater or equal to 50, T stage 2 was 0.66.

Among patients in the pN1a subgroup with larger LNR and the pN1b subgroup with both smaller and larger LNR, ten-year BCSS for patients with grade 1 and 2, in the pN1b subgroup with smaller LNR was 0.87; for patients with grade 1 and 2, in the pN1a and the pN1b subgroups with larger LNR was 0.67; for patients with grade 3, in the pN1b subgroup with smaller LNR was 0.66; for patients with grade 3, in the pN1a and the pN1b subgroups with larger LNR was 0.5. The ten-year OS for patients with T stage 1 was 0.68; for patients with T stage 2, age less than 50 was 0.6; for patients with T stage 2, age greater or equal to 50 was 0.43.

The BCSS and OS Kaplan-Meier curves for the corresponding groups defined by the survival trees are shown in Figure 6.6.2 and 6.6.4. The survival trees suggested that patients with micrometastatic disease and smaller LNR had similar survival to patients with node negative metastases. Grouping results on age, T stage, tumor grade, LVI status, and histology category agreed with the multivariable Cox PH analyses in Section 6.2. The BCSS tree suggested that patients in the pN1a with larger LNR and pN1b subgroups with larger LNR had similar survival outcomes, which agreed with results from unadjusted and adjusted KM curves in Section 6.5. Results from the log rank test are $\chi^2 = 1083$, *d.f.* $= 10$, $p = 0$ for BCSS and $\chi^2 = 738$, *d.f.* $= 8$, $p = 0$ for OS.



Group 1 2 3 4 5 6 7 8 9 10 11

LNR25SUR: Surgical N staging & Categorized lymph node ratio (LNR)
(0,0): The pN0 subgroup, LNR=0
(1A,S): The pN1a subgroup, 0.01<LNR<=0.25
(1B,S): The pN1b subgroup, 0.01<LNR<=0.25

Figure 6.6.1: BCSS survival tree showing ten-year survival rates, cases with missing values retained. The group numbers correspond with BCSS KM curves groups in Figure 6.6.2.

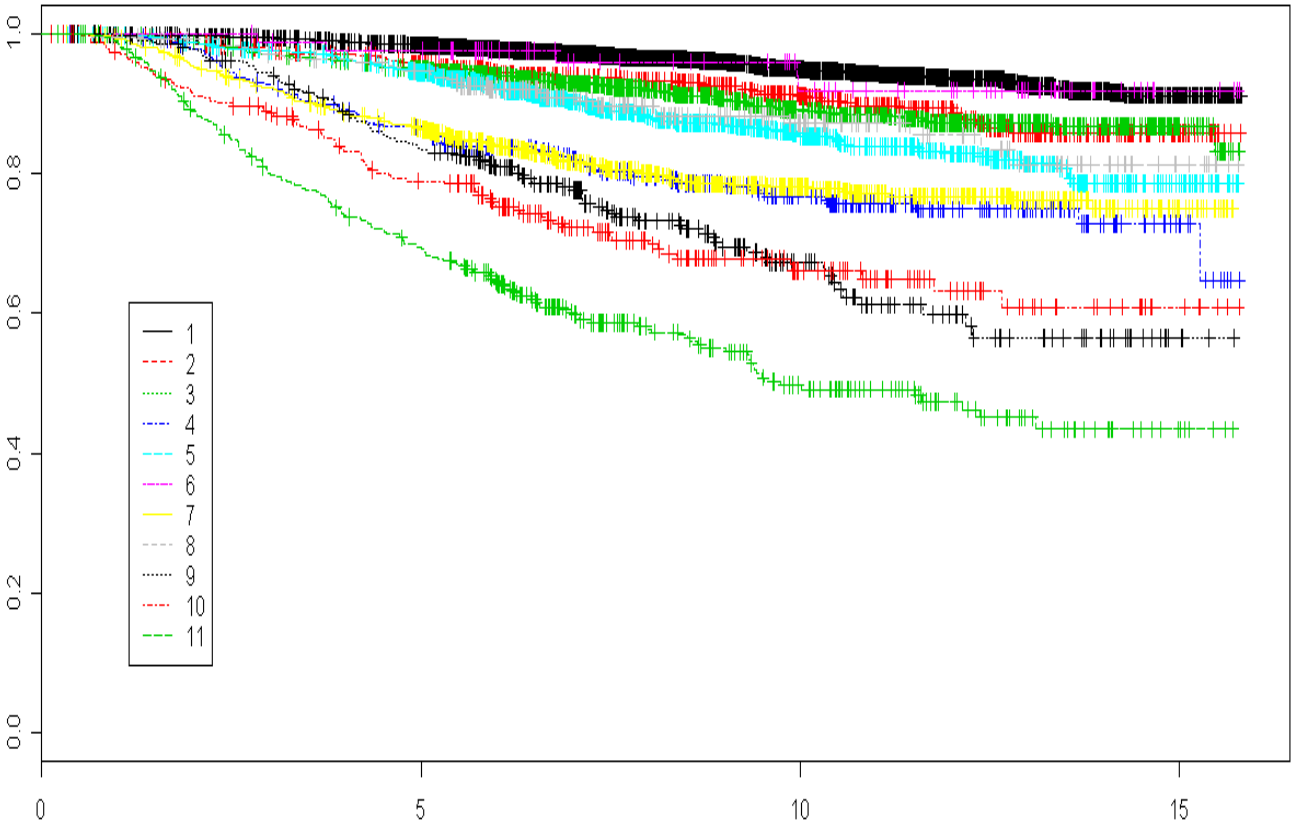
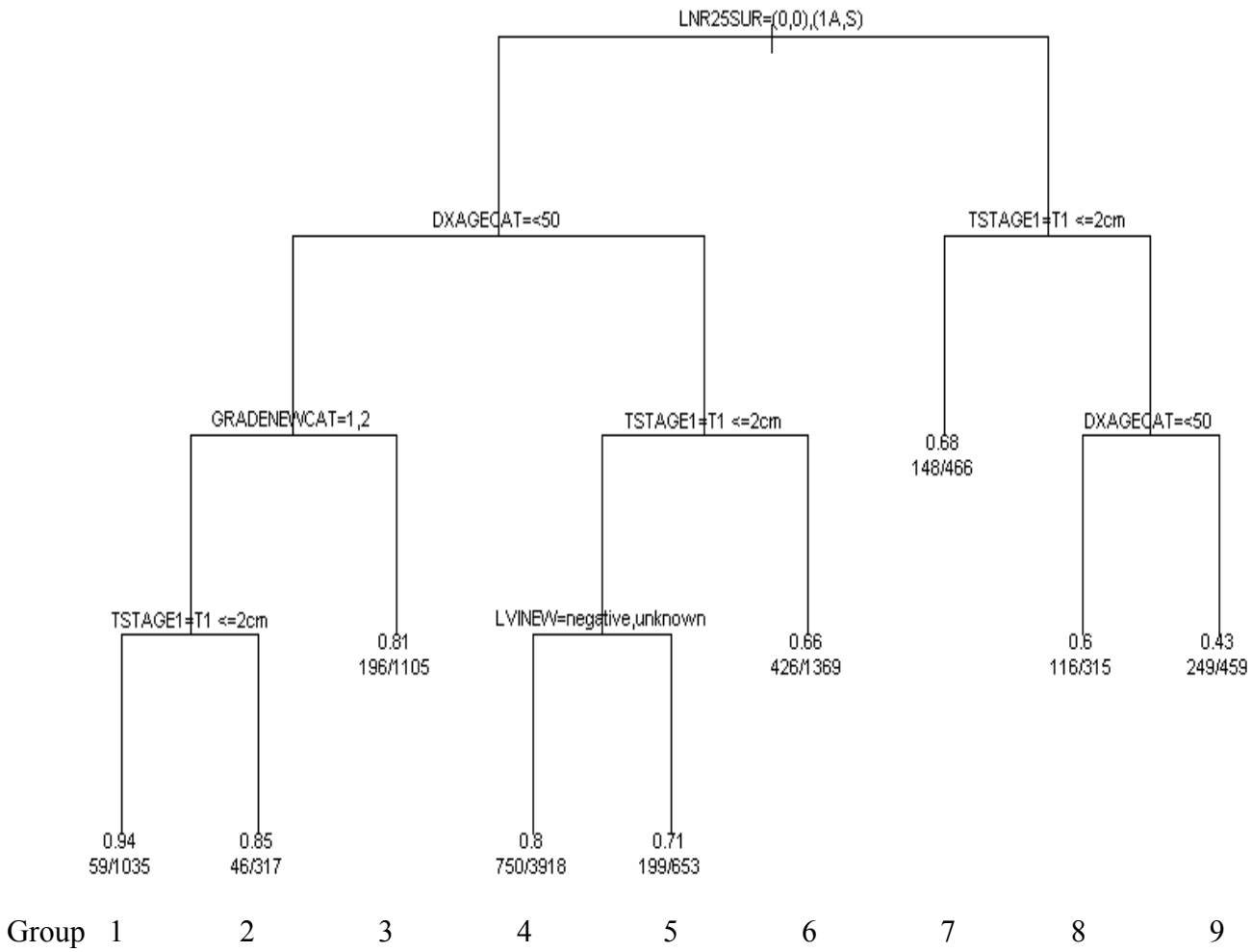


Figure 6.6.2: KM BCSS survival curves for groups defined by the survival tree.



LNR25SUR: Surgical N staging & Categorized lymph node ratio (LNR)
(0,0): The pN0 subgroup, LNR=0
(1A,S): The pN1a subgroup, 0.01<LNR<=0.25

Figure 6.6.3: OS survival tree showing ten-year survival rates, cases with missing values retained. The group numbers correspond with OS KM curves groups in Figure 6.6.4.

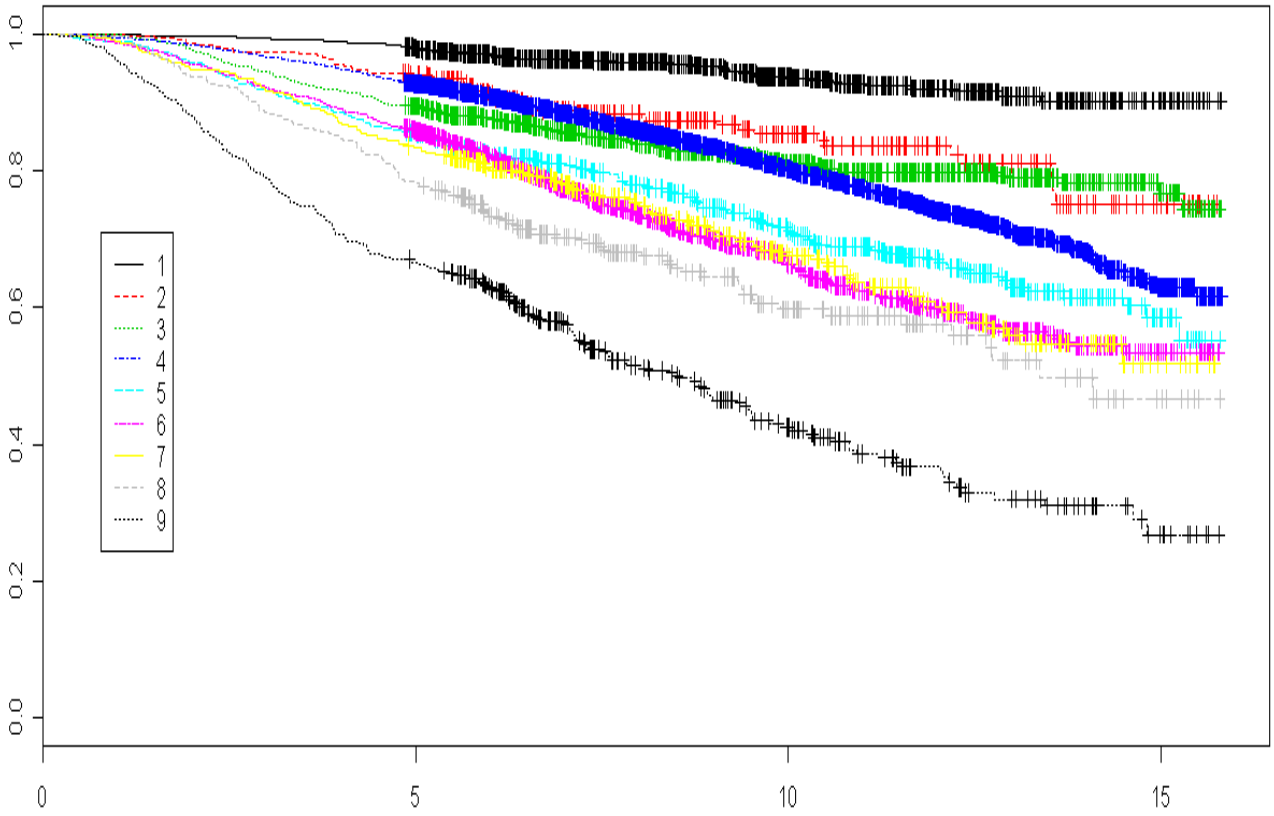


Figure 6.6.4: KM OS survival curves for groups defined by the survival tree.

6.7 Nomograms

A nomogram is a graphic calculating device that predicts the probability of a certain clinical outcome for an individual patient using specific clinic variables [8]. The first medical nomogram was developed in 1928 by L.J. Henderson [2]. A nomogram is a two-dimensional graph that translates results from regression analyses. It provides disease-related risk estimation for each individual patient and is used by physicians for patient consultation. In contrast to prediction using the traditional grouping approach that groups patients with similar characteristics, the nomogram makes personalized prediction based on individual patient characteristics [8]. Patient groups selected from cutoff values for nomograms are more homogeneous since a standard variable, the predicted probability, is used. Many studies have used Cox PH models with categorical covariates to construct nomograms for breast cancer survival estimation [12, 13, 26]. There are online nomogram tools, such as the Memorial Sloan-Kettering Cancer Center (www.mskcc.org), which allows patients, physicians, or researchers to enter specific patient characteristics for prognosis prediction [12]. The accuracy is independent of the numbers of covariates in a prediction model but restricted by data quality [2]. A nomogram should provide predictions that are close to the actual outcomes and consistent results when applied to different datasets.

In our study, survival nomograms constructed by the four Cox PH models introduced in Section 6.2, Table 6.2.1 were generated using Harrell's Design library in R 2.4.1. Nomograms illustrating the results of the Cox PH model estimates are provided in Figure 6.5.1-6.5.4 for ten-year survival. To estimate ten-year survival for a given patient, read off the points using the top scale for each covariate. Add up all points and use the

"Total Points" scale to line up with ten-year survival. For example, from Figure 6.7.1, a patient with age less than 50, mastectomy, T2 stage, less than 15 nodes removed, no initial systemic therapy, in pN1a subgroup with 1-3 positive nodes, ER positive, histology lobular, grade 2 primary tumor, LVI positive has a total points of 251 and a ten-year survival probability of 0.8.

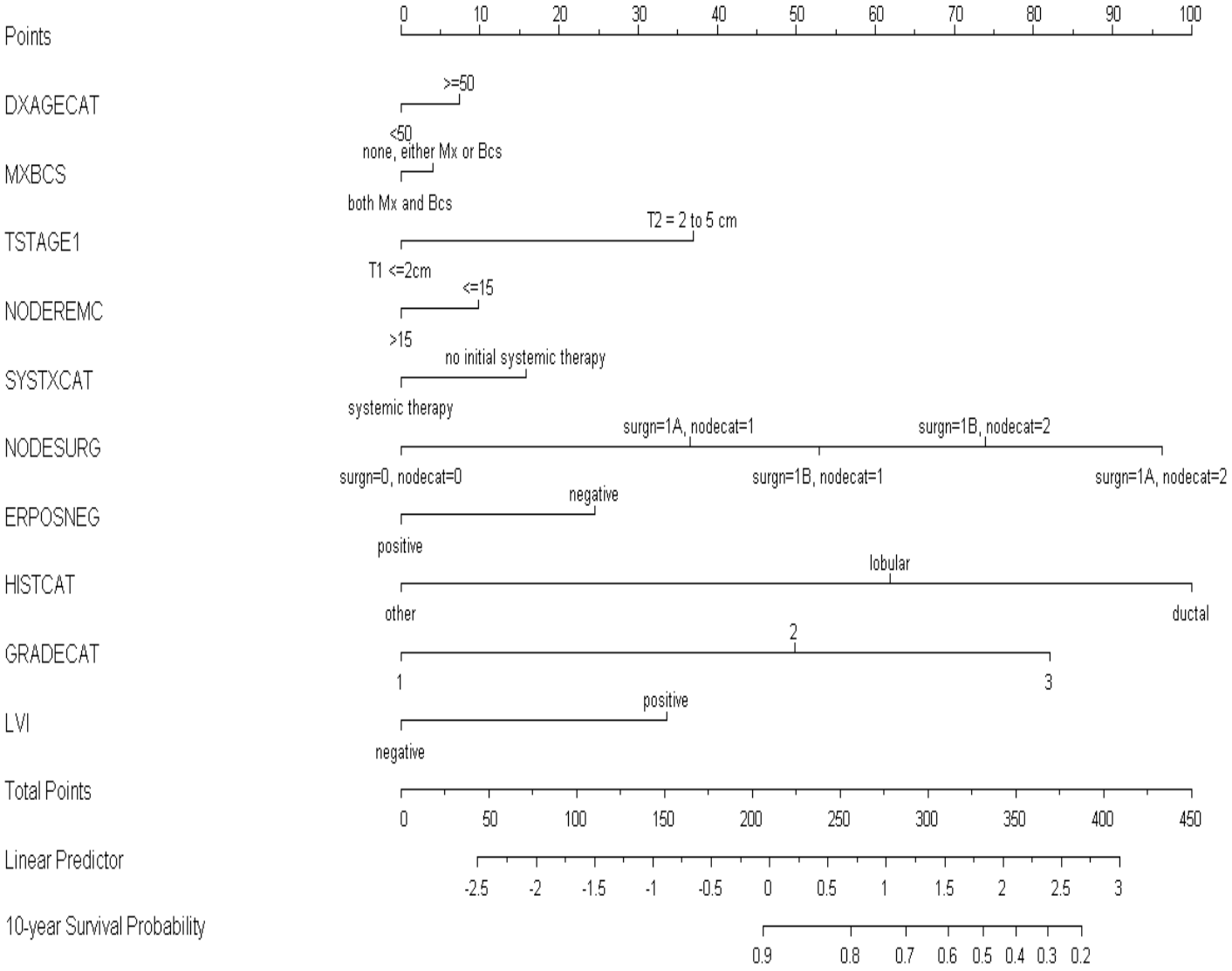


Figure 6.7.1: Nomogram illustrating the results of the BCSS with number of positive nodes model for ten-year survival.

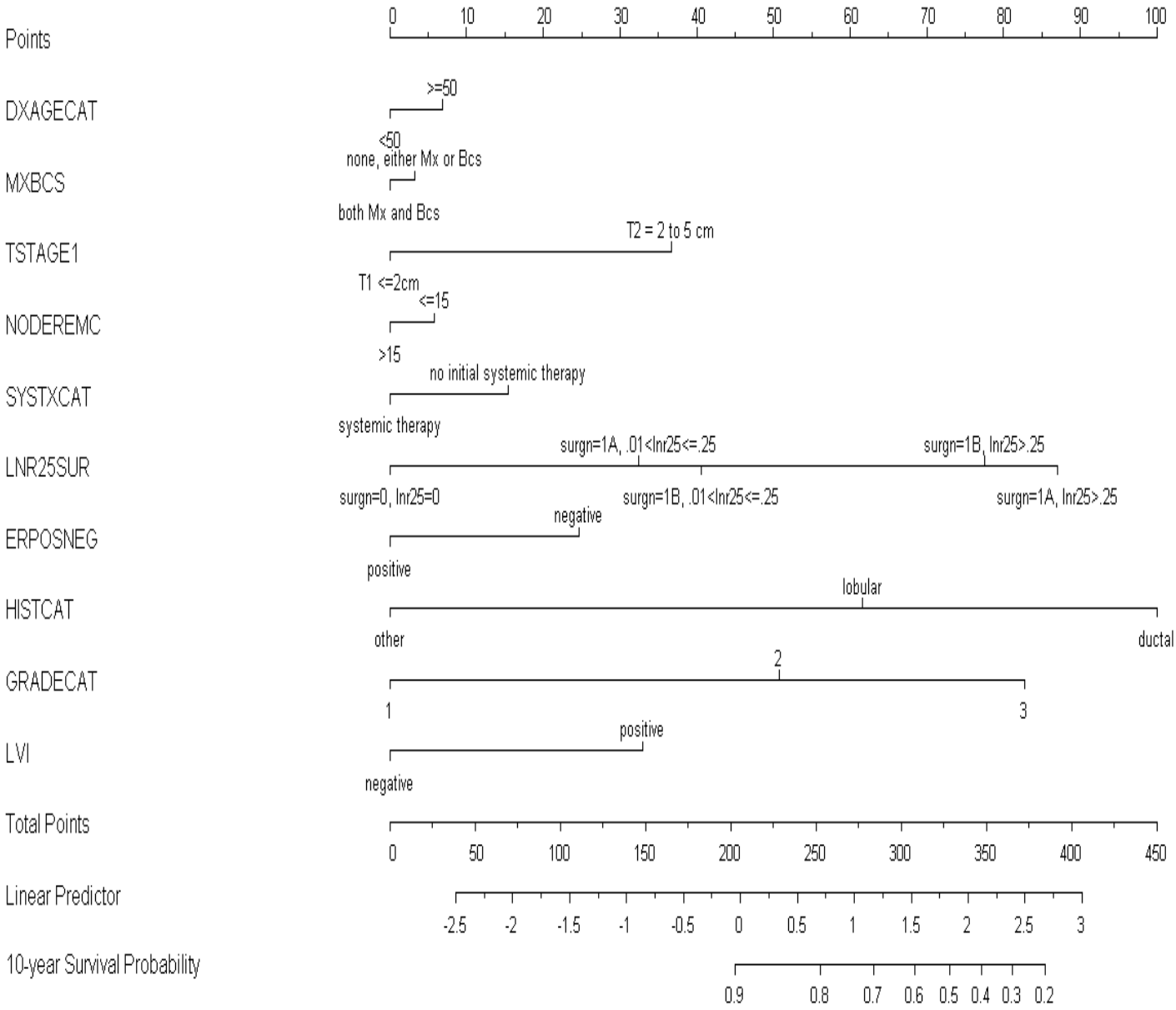


Figure 6.7.2: Nomogram illustrating the results of the BCSS with LNR model for ten-year survival.

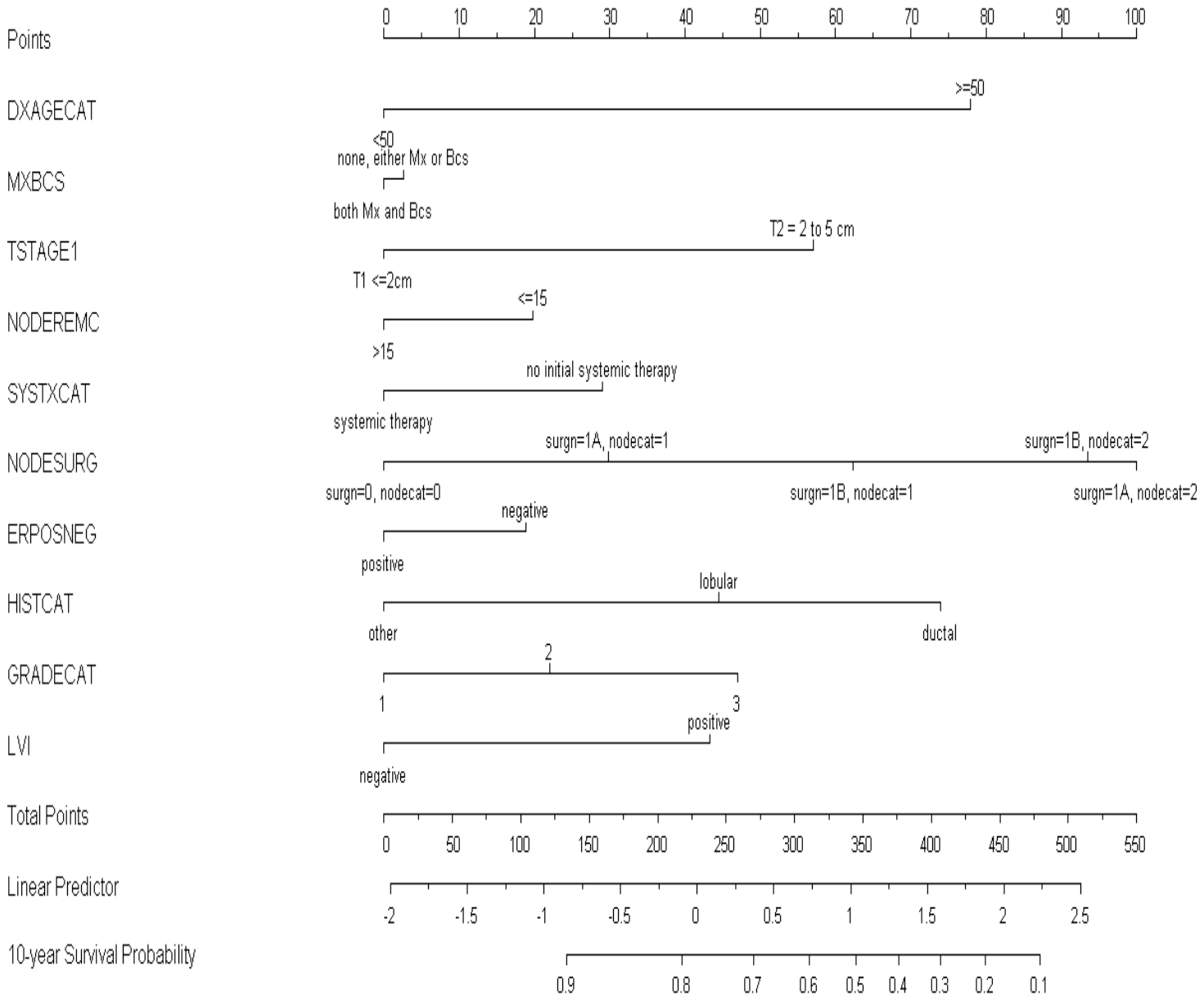


Figure 6.7.3: Nomogram illustrating the results of the OS with number of positive nodes model for ten-year survival.

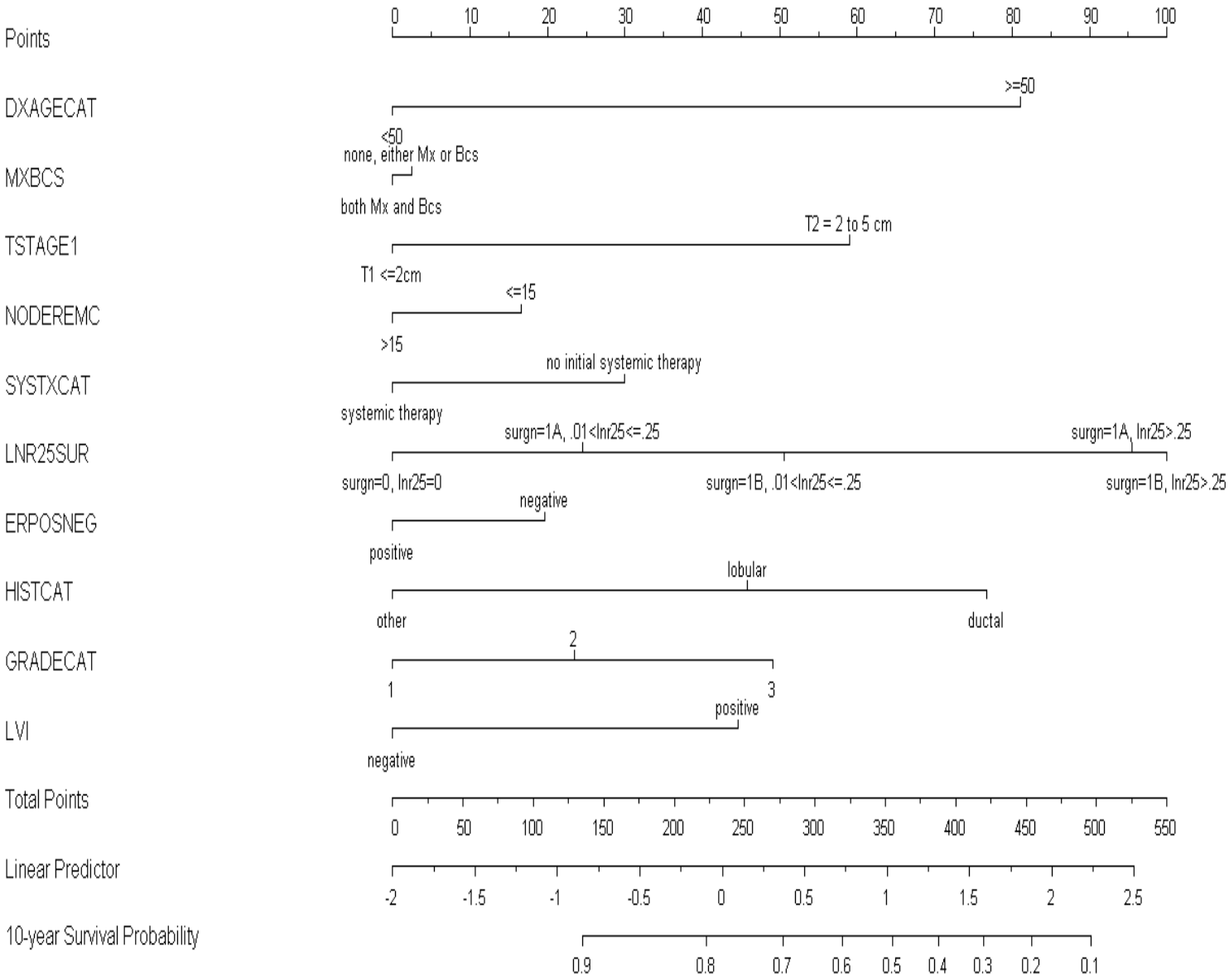


Figure 6.7.4: Nomogram illustrating the results of the OS with LNR model for ten-year survival.

Chapter 7

7 Conclusions

The survival and therapeutic implication of patients with micrometastatic node-positive cancer currently remains unclear. Knowledge of survival outcomes of the pN1a subgroup in comparison to the pN0 and pN1b subgroups and the prognostic impact of the number of positive nodes and lymph node ratio can provide insight into breast cancer management and improve clinicians' ability to appraise risks in women.

Despite the relatively small number of micrometastatic patients, this study demonstrated that patients with micrometastatic node-positive breast cancer had worse survival outcomes than node-negative patients, but better survival outcomes compared to macrometastatic node-positive patients. The prognostic impact of the absolute number of positive nodes and the ratio of positive versus excised nodes on survival were significant. Ten-year breast cancer specific survival and overall survival estimated using the Kaplan-Meier method suggested that increasing number of positive nodes and larger values of lymph node ratio were associated with worse survival in both pN1a and pN1b subgroups. The multivariable Cox PH model analyses indicated that both the number of positive nodes and lymph node ratio were strong prognostic indicators for survival in the entire cohort and in all three subgroups. Results from a population based analysis, conducted by Dr. P. Truong et al. [30], on 62,551 women identified by the Surveillance Epidemiology and the End Results (SEER) database agreed with our findings. The median follow-up was 7.3 years and the number of patients in pN0, pN1a, and pN1b subgroups were 57,980, 1,818, and 2,753 respectively.

The confounding effect of the number of excised nodes on determining number of positive nodes and its impact on survival outcomes and disease management remain unclear. The lymph node ratio that standardized the traditional staging using absolute number of positive node to the number of excised nodes may be a more comprehensive choice for prognosis estimation [30]. Results confirmed that LNR, as well as the number of positive nodes, were significant prognostic indicators.

Determining the functional form of the time-dependent covariate ER status was challenging. Different functions and other parametric survival models may be fit to further investigate the time-dependent property. We did not illustrate all the estimators from stratified models, log-normal regression models, or time-dependent models in SAS, because results were consistent. A further comparison among the goodness-of-fit from different methods may be carried out.

Other methods such as survival trees and nomograms provided exciting ways to explore survival prediction. Survival trees with different algorithms are worth examining. We discovered some interesting findings from the adjusted survival curves. Among patients with larger number of positive nodes and LNR, pN1b subgroup had better survival outcomes compared to the pN1a group, although the differences were only noticeable in BCSS with number of positive nodes. Future studies can take one step further to investigate the reasons behind this.

Bibliography

[1] Attiyeh FF, Jensen M, Huvos AG, Fracchia A. Axillary micrometastasis and macrometastasis in carcinoma of the breast. *Surg Gynecol Obstet.* 1977; 144 (6): 839-42.

[2] Bianco FJ Jr. Nomograms and Medicine. *Eur Urol.* 2006; 50 (5): 884-6.

[3] Bonadonna G, Hortobagyi GN, Balagussa P. *Textbook of Breast Cancer: A Clinical Guide to Therapy* (Ed 3). London and New York: Taylor & Francis Group, 2006.

[4] Chambers JM, Hastie TJ. (Eds.) *Statistical Models in S.* California: Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[5] Chen SL, Hoehne FM, Giuliano AE. The prognostic significance of micrometastases in breast cancer: a SEER population-based analysis. *Ann Surg Oncol.* 2007; 14(12): 3378-84.

[6] Clayton F, Hopkins CL. Pathologic correlates of prognosis in lymph node-positive breast carcinomas. *Cancer.* 1993; 71 (5): 1780-90.

[7] Coradini D, Daidone MG, Boracchi P, Biganzoli E, Oriana S, Bresciani G, Pellizzaro C, Tomasic G, Fronzo GD, Marubini E. Time-dependent Relevance of steroid receptors in breast cancer. *J Clin Oncol.* 2000;18(14):2702-9.

[8] Diblasio CJ, Kattan MW. Use of nomograms to predict the risk of disease recurrence after definitive local therapy for prostate cancer. *Urology*. 2003; 62 (Suppl 6B): 9-18.

[9] Fisher LD, Lin DY. Time-dependent covariates in the Cox Proportional-Hazards regression model. Annu. *Rev. Public Health.* 1999; 20: 145-57.

[10] *Getting Started with S-PLUS 6 for Windows.* Washington: Insightful Corporation, 2001.

[11] Ghali MA, Quan H, Brant R, van Melle G, Norris CM, Faris PD, Galbraith PD, Knudtson M. Comparison of 2 Methods for Calculating Adjusted Survival Curves From Proportional Hazards Models. *JAMA.* 2001; 286(12): 1494-7.

[12] Gur AS, Unal B, Johnson R, Ahrendt G, Bonaventura M, Gordon P, Soran A. Predictive probability of four different breast cancer nomograms for nonsentinel axillary lymph node metastasis in positive sentinel node biopsy. *J Am Coll Surg.* 2009; 208(2): 299-35.

[13] Hanrahan EO, Gonzalez-Angulo AM, Giordano SH, Rouzier R, Broglio KR, Hortobagyi GN, Valero V. Overall survival and cause-specific mortality of patients with stage T1a,bN0M0 breast carcinoma. *J Clin Oncol*. 2007; 25(31): 4952-60.

[14] Harrell F. *Regression Modelling Strategies*. New York, Springer-Velag, 2001.

[15] Hartveit F, Lilleng PK. Breast cancer: two micrometastatic variants in the axilla that differ in prognosis. *Histopathology.* 1996; 28(3): 241-6.

[16] Klauber-DeMore N, Ollila DW, Moore DT, Livasy C, Calvo BF, Kim HJ, Dees EC, Sartor CI, Sawyer LR, Graham M, Carey LA. Size of residual lymph node metastasis after neoadjuvant chemotherapy in locally advanced breast cancer patients is prognostic. *Ann Surg Oncol.* 2006; 13(5): 685-91.

[17] Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data* (Ed 2). New York: Springer, 2003.

[18] Knight WA, Osborne CK, Yochmowitz MG, McGuire WL. Steroid hormone receptors in the management of human breast cancer. *Ann Clin Res*. 1980; 12(5): 202-7.

[19] Kuijt GP, van de Poll-Franse LV, Roumen RMH, van Beek MWPM, Voogd AC. The significance of one positive axillary node. *Eur J Surg Oncol*. 2006; 32(2): 139-42.

[20] Kuijt GP, Voogd AC, van de Poll-Franse LV, Scheijmans LJEE, van Beek MWPM, Roumen RMH. The prognostic significance of axillary lymph-node micrometastases in breast cancer patients. *Eur J Surg Oncol.* 2005; 31(5): 500-5.

[21] Kurosumi M, Suemasu K, Tabei T, Inoue K, Matsumoto H, Sugamata N, Higashi Y. Relationship between existence of lymphatic invasion in peritumoral breast tissue and presence of axillary lymph node metastasis in invasive ductal carcinoma of the breast. *Oncol Rep.* 2001; 8(5): 1051-5.

[22] Lilleng PK, Maehle BO, Hartveit F. The size of a micrometastasis in the axilla in breast cancer: a study of nodal tumour-load related to prognosis. *Eur J Gynaecol Oncol.* 1998; 19(3): 220-4.

[23] Olivotto I, Gelmon K, Kuusk U. *Intelligent Patient Guide to Breast Cancer.* Vancouver: Intelligent Patient Guide Ltd. 1996.

[24] Osborne CK. Steroid hormone receptors in breast cancer management. *Breast Cancer Res Treat.* 1998; 51(3): 227-38.

[25] Nieto FJ, Coresh J. Adjusting Survival Curves for Confounders: A Review and a New Method. *Am. J. Epidemiol.* 1996; 143(10): 1059-68.

[26] Rouzier R, Pusztai L, Delaloge S, Gonzalez-Angulo AM, Andre F, Hess KR, Buzdar AU, Garbay JR, Spielmann M, Mathieu MC, Symmans WF, Wagner P, Atallah D, Valero V, Berry DA, Hortobagyi GN. Nomograms to predict pathologic complete

response and metastasis-free survival after preoperative chemotherapy for breast cancer. *J Clin Oncol*. 2005; 23(33): 8831-9.

[27] S-*PLUS 6 for Windows: Guide to Statistics, Volume 2*. Washington: Insightful Corporation, 2001.

[28] Statistical Computing, UCLA Academic Technology Services. http://www.ats.ucla.edu/stat/

[29] Tableman M, Kim JS. *Survival Analysis Using S: Analysis of Time-to-Event Data.* Florida: Chapman & Hall/CRC, 2004.

[30] Truong PT, Cserni G, Woodward W, Janni W, Tai P, Vlastos G, Vinh-Hung V. The prognostic impact of the axillary lymph node ratio in patients with micrometastatic node-positive breast cancer. Poster #3008, 28th San Antonio Breast Cancer Symposium, December 2005.

[31] Zhang H, Singer B. *Statistics for Biology and Health: Recursive Partitioning in the Health Sciences.* New York: Springer-Verlag, 1999.

Appendix A

A.1 Residual Results for BCSS Model with LNR

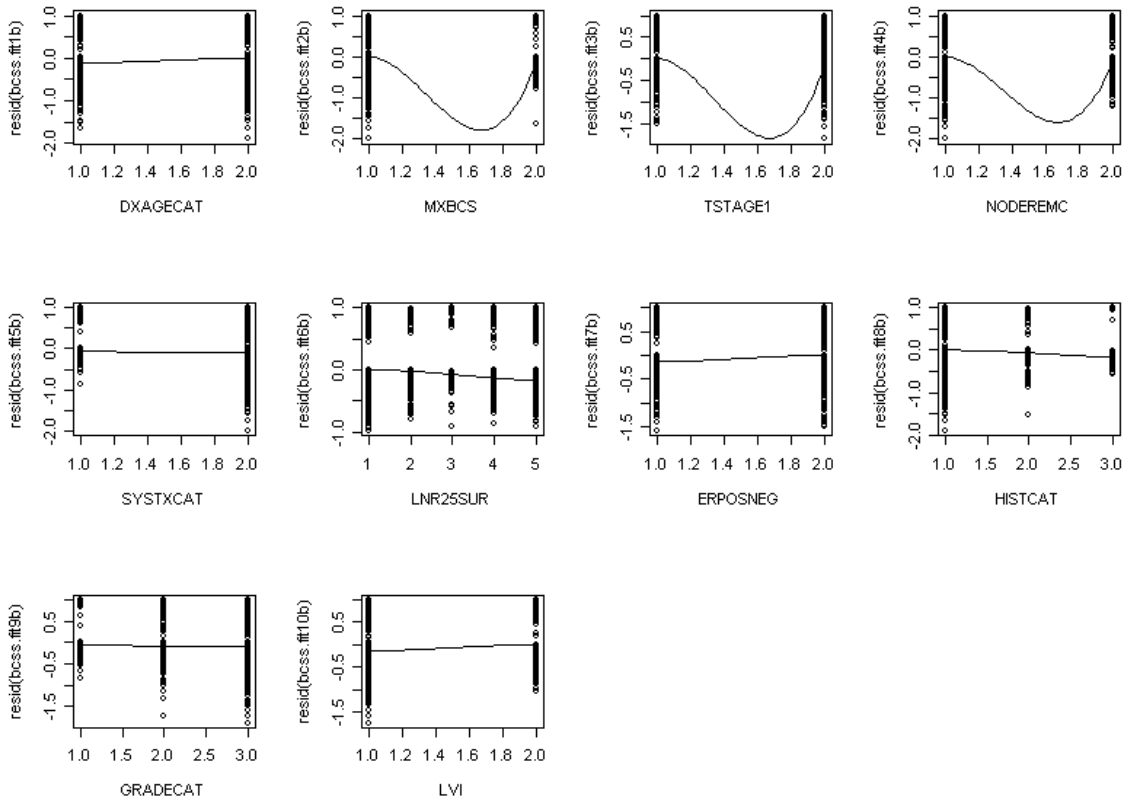


Figure A.1.1: The martingale residuals plots for the BCSS model with LNR.

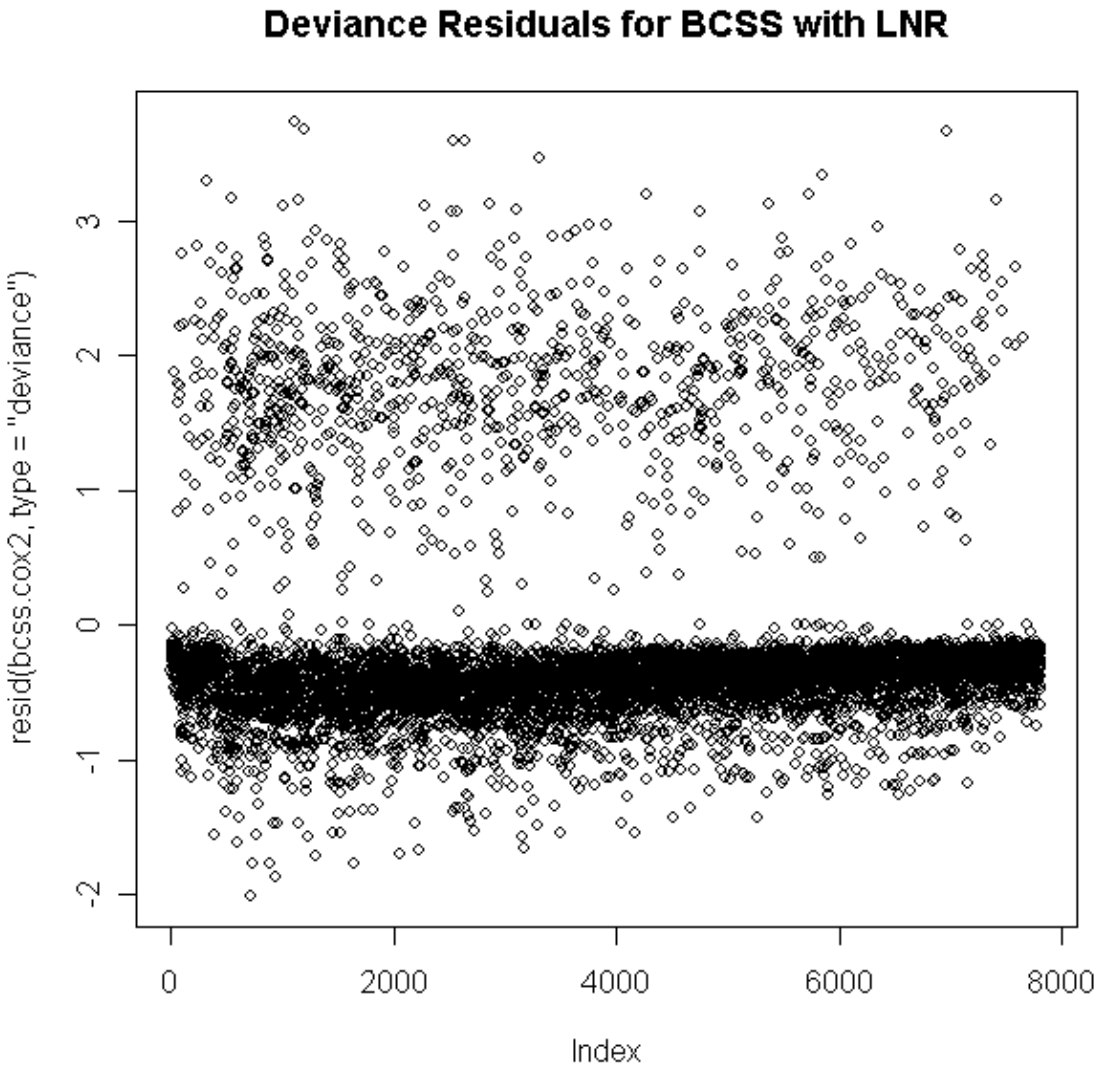


Figure A.1.2: The deviance residuals plot for the BCSS model with LNR.

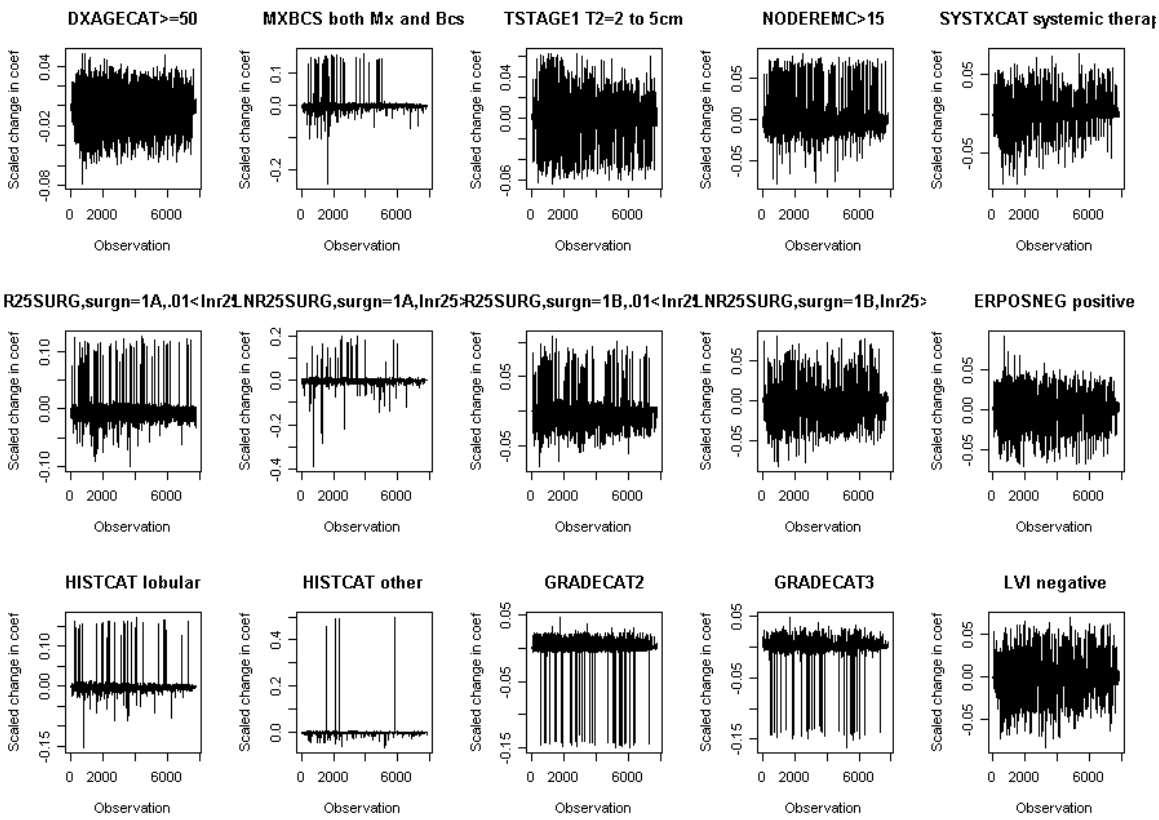


Figure A.1.3: The influence plots for the fifteen important predictors for the BCSS model with LNR.

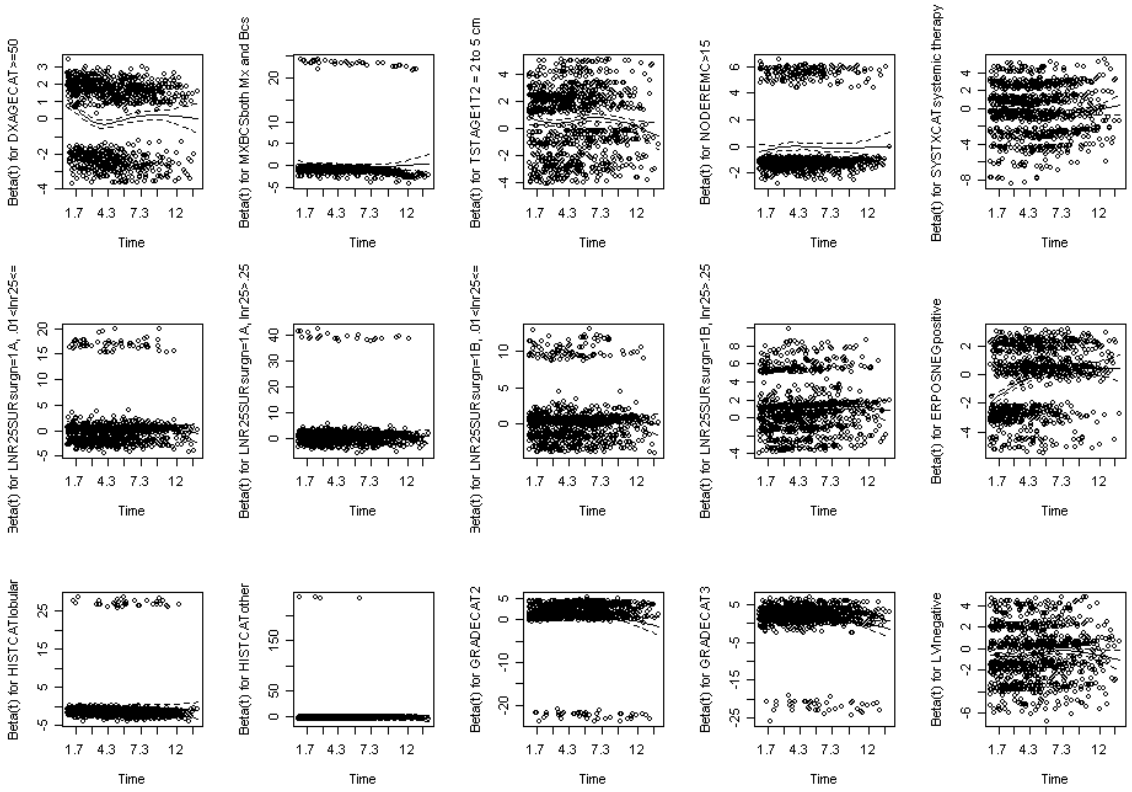


Figure A.1.4: The rescaled Schoenfeld residuals plots for the BCSS model with LNR.

Table A.1.1: Statistical tests for significant slope in the rescaled Schoenfeld residuals plots in Figure A.1.4.

	rho	chisq	p
DXAGECAT>=50	-0.03502	1.17e+00	0.2791
MXBCSboth Mx and Bcs	-0.01499	2.15e-01	0.6428
TSTAGE1T2 = 2 to 5 cm	-0.01329	1.92e-01	0.6612
NODEREMC>15	0.01497	2.12e-01	0.6452
SYSTXCATsystemic therapy	-0.04037	1.67e+00	0.1956
LNR25SURsurgn=1A, .01<lnr25<=.25	0.00268	6.89e-03	0.9339
LNR25SURsurgn=1A, lnr25>.25	-0.00692	4.57e-02	0.8308
LNR25SURsurgn=1B, .01<lnr25<=.25	-0.00422	1.73e-02	0.8954
LNR25SURsurgn=1B, lnr25>.25	0.01648	2.68e-01	0.6045
ERPOSNEGpositive	0.27678	7.51e+01	0.0000
HISTCATlobular	0.04603	1.97e+00	0.1607
HISTCATother	-0.03016	8.61e-01	0.3534
GRADECAT2	-0.04023	1.53e+00	0.2167
GRADECAT3	-0.07087	4.79e+00	0.0287
LVInegative	0.04797	2.38e+00	0.1227
GLOBAL	NA	1.36e+02	0.0000

A.2 Residual Results for OS Model with Number of Positive Nodes

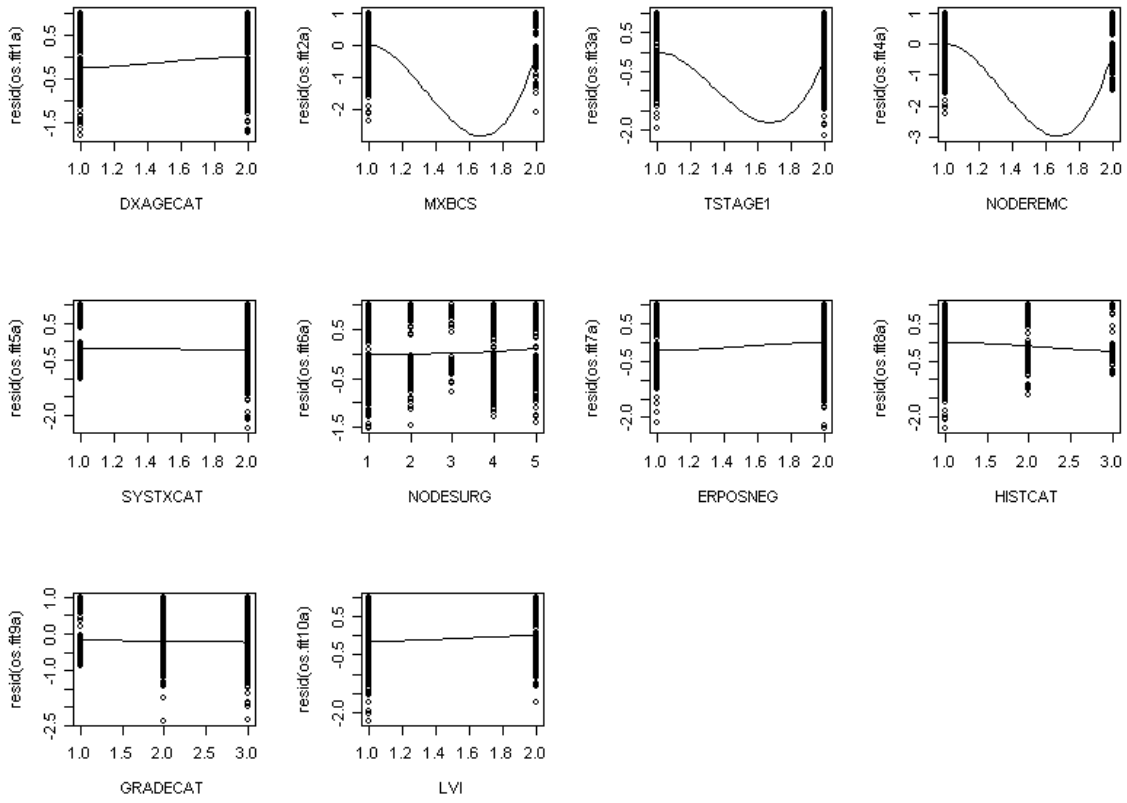


Figure A.2.1: The martingale residuals plots for the OS model with number of positive nodes.

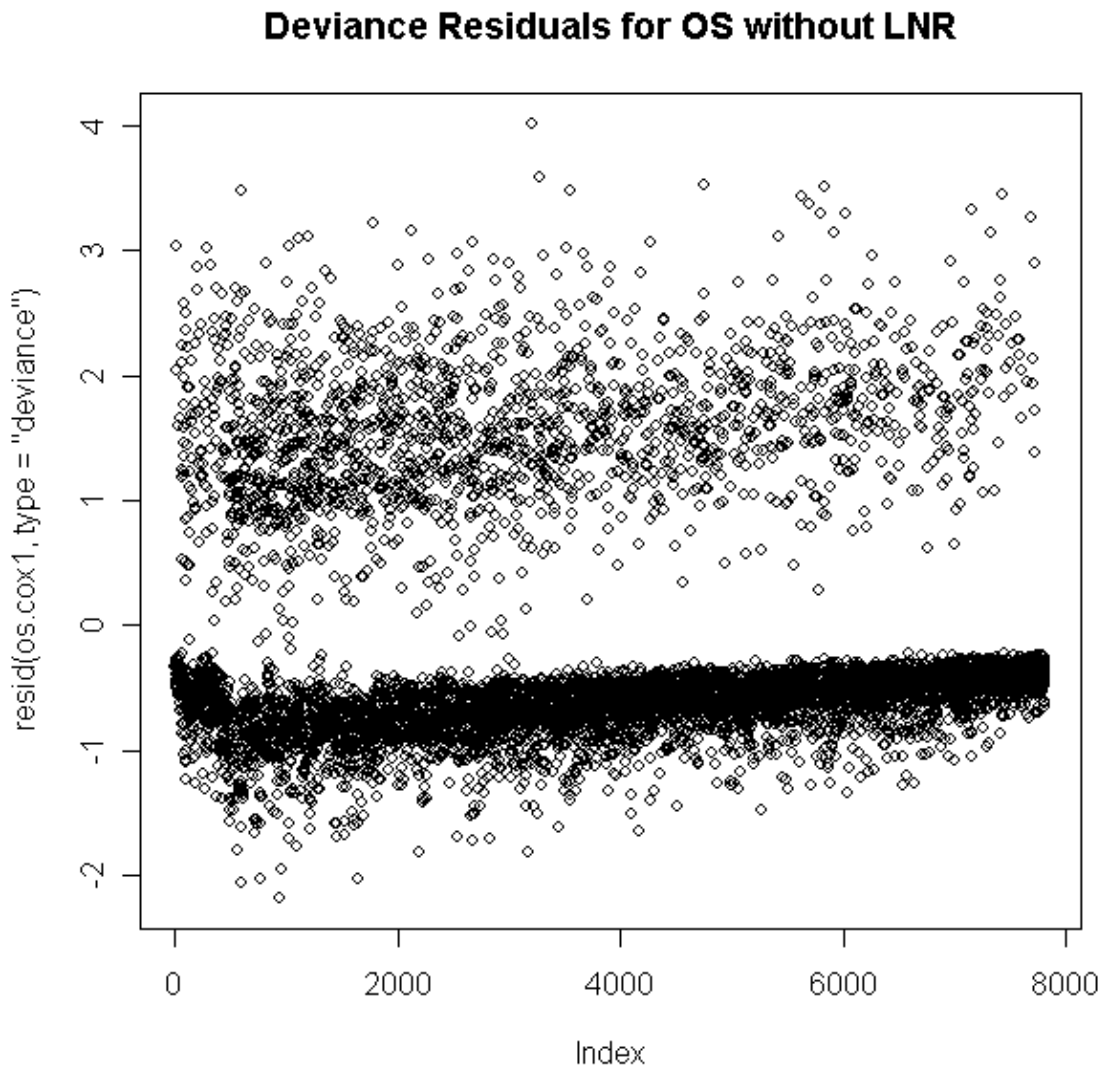


Figure A.2.2: The deviance residuals plot for the OS model with number of positive nodes.

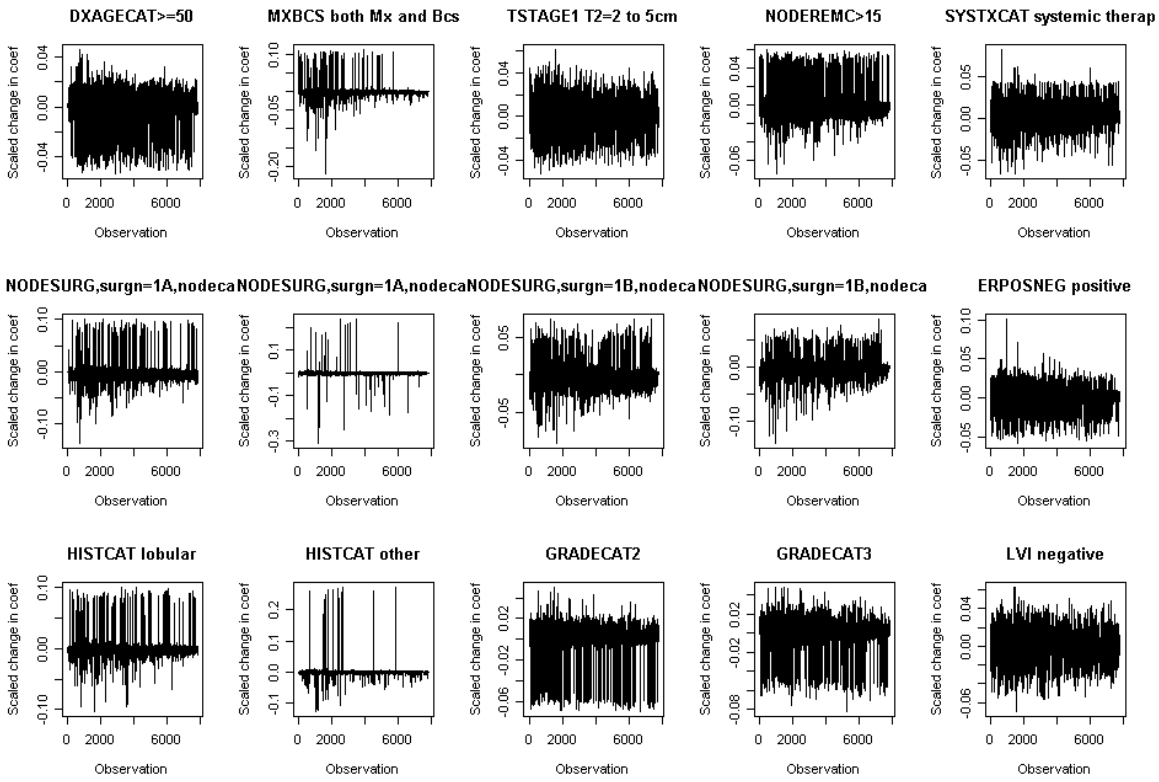


Figure A.2.3: The influence plots for the fifteen important predictors for the OS model with number of positive nodes.

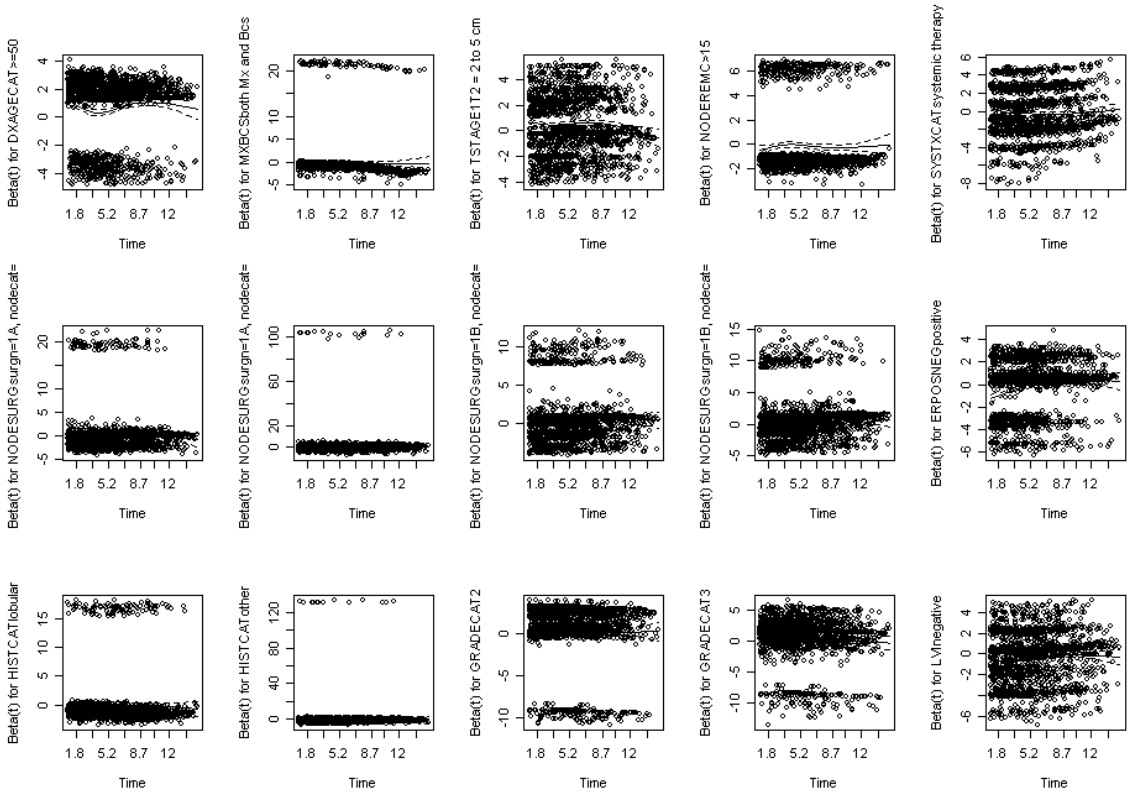


Figure A.2.4: The rescaled Schoenfeld residuals plots for the OS model with number of positive nodes.

Table A.2.1: Statistical tests for significant slope in the rescaled Schoenfeld residuals plots in Figure A.2.4.

	rho	chisq	p
DXAGECAT>=50	0.037790	2.36e+00	1.25e-01
MXBCSboth Mx and Bcs	-0.033601	1.98e+00	1.60e-01
TSTAGE1T2 = 2 to 5 cm	-0.036024	2.41e+00	1.21e-01
NODEREMC>15	0.000274	1.29e-04	9.91e-01
SYSTXCATsystemic therapy	0.013808	3.53e-01	5.52e-01
NODESURGsurgn=1A, nodecat=1	-0.024596	1.04e+00	3.09e-01
NODESURGsurgn=1A, nodecat=2	0.006903	8.27e-02	7.74e-01
NODESURGsurgn=1B, nodecat=1	-0.005434	5.15e-02	8.20e-01
NODESURGsurgn=1B, nodecat=2	-0.004323	3.24e-02	8.57e-01
ERPOSNEGpositive	0.186629	5.90e+01	1.55e-14
HISTCATlobular	0.022947	8.99e-01	3.43e-01
HISTCATother	0.004625	3.69e-02	8.48e-01
GRADECAT2	-0.017148	5.04e-01	4.78e-01
GRADECAT3	-0.067013	7.61e+00	5.81e-03
LVInegative	0.073610	9.71e+00	1.83e-03
GLOBAL	NA	1.61e+02	0.00e+00

A.3 Residual Results for OS Model with LNR

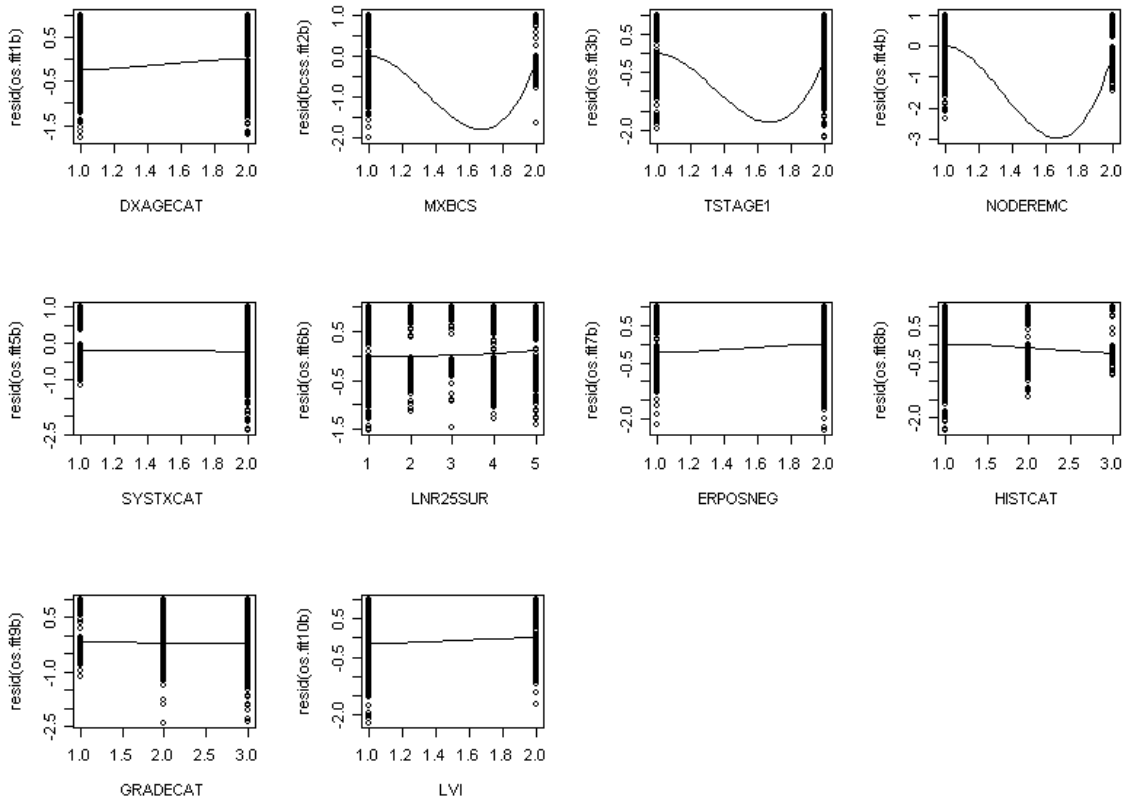


Figure A.3.1: The martingale residuals plots for the OS model with LNR.

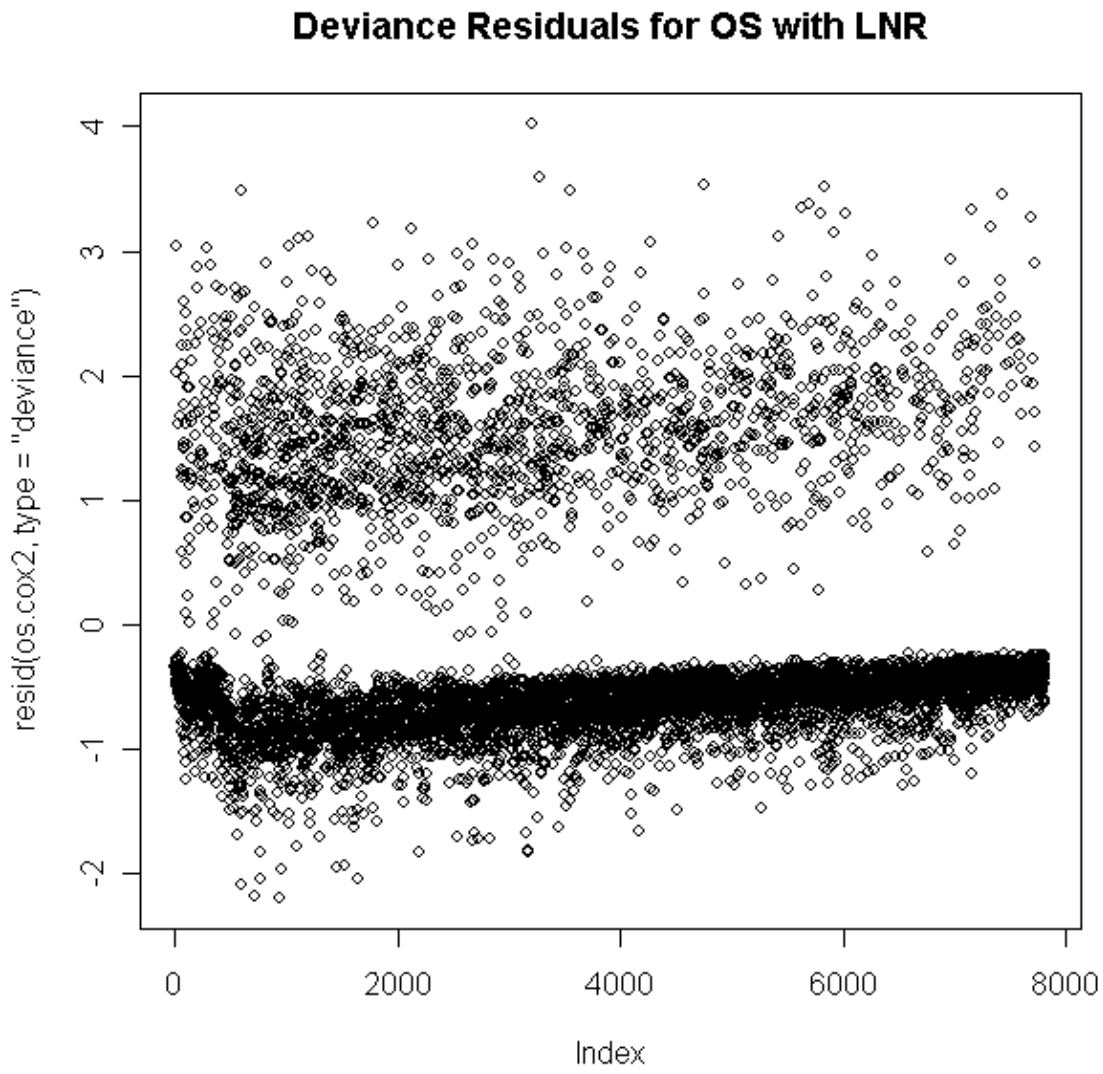


Figure A.3.2: The deviance residuals plot for the OS model with LNR.

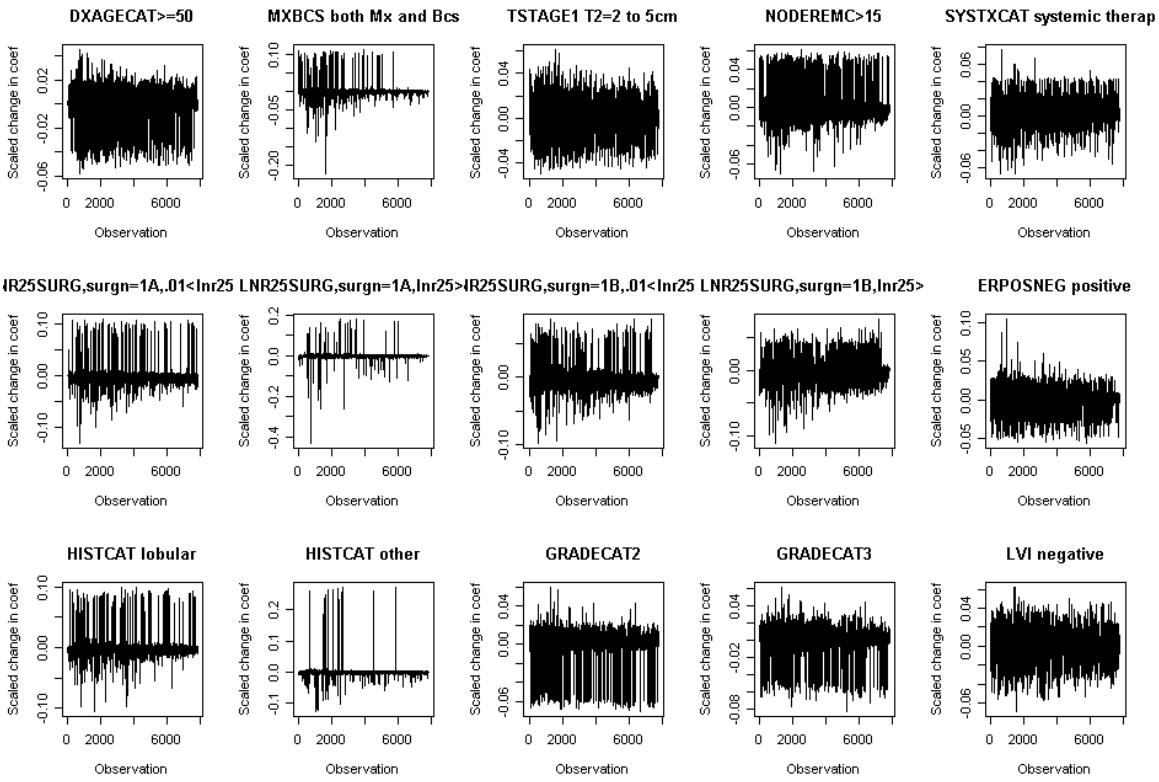


Figure A.3.3: The influence plots for the fifteen important predictors for the OS model with LNR.

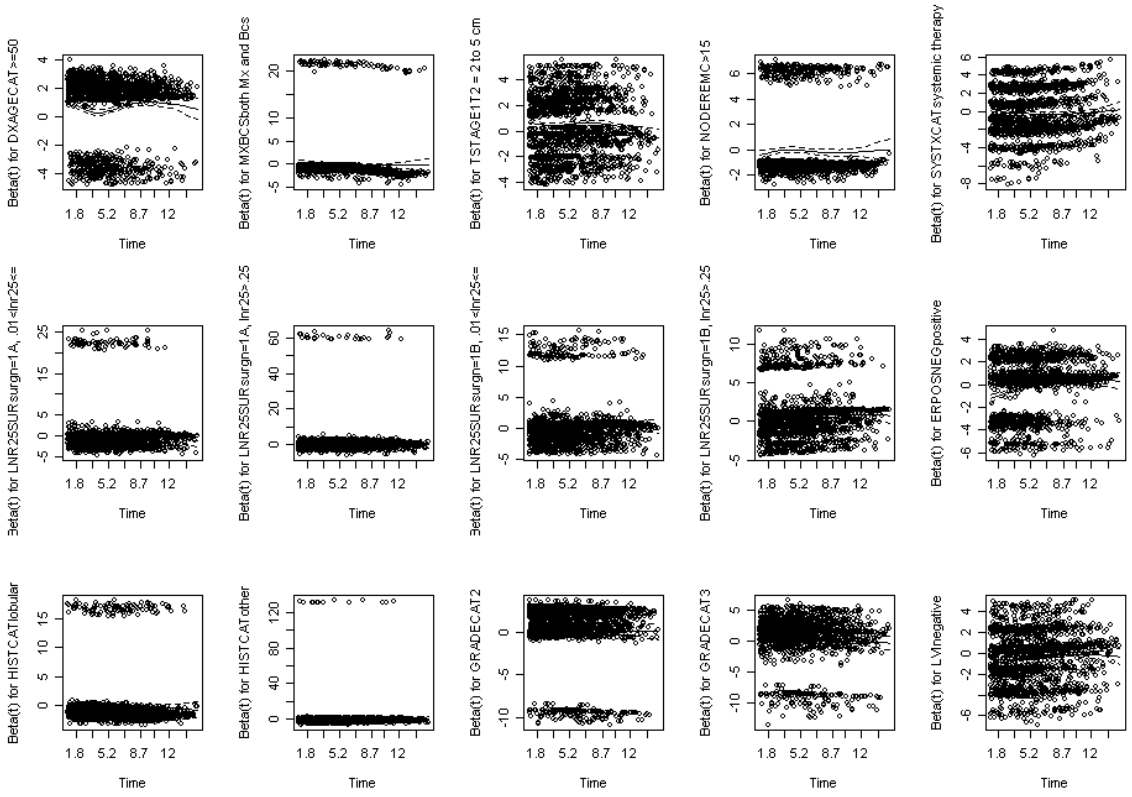


Figure A.3.4: The rescaled Schoenfeld residuals plots for the OS model with LNR.

Table A.3.1: Statistical tests for significant slope in the rescaled Schoenfeld residuals plots in Figure A.3.4.

	rho	chisq	p
DXAGECAT>=50	0.03717	2.27e+00	1.32e-01
MXBCSboth Mx and Bcs	-0.03351	1.97e+00	1.60e-01
TSTAGE1T2 = 2 to 5 cm	-0.03575	2.37e+00	1.24e-01
NODEREMC>15	-0.00060	6.18e-04	9.80e-01
SYSTXCATsystemic therapy	0.01325	3.26e-01	5.68e-01
LNR25SURsurgn=1A, .01<lnr25<=.25	-0.02357	9.53e-01	3.29e-01
LNR25SURsurgn=1A, lnr25>.25	-0.00211	7.73e-03	9.30e-01
LNR25SURsurgn=1B, .01<lnr25<=.25	-0.00267	1.25e-02	9.11e-01
LNR25SURsurgn=1B, lnr25>.25	-0.00461	3.68e-02	8.48e-01
ERPOSNEGpositive	0.18833	6.01e+01	9.10e-15
HISTCATlobular	0.02308	9.07e-01	3.41e-01
HISTCATother	0.00481	4.00e-02	8.41e-01
GRADECAT2	-0.01624	4.52e-01	5.01e-01
GRADECAT3	-0.06706	7.61e+00	5.79e-03
LVInegative	0.07132	9.12e+00	2.53e-03
GLOBAL	NA	1.61e+02	0.00e+00

Appendix B

B.1 R code for the Multivariable Cox PH Models Analyses

```
library('survival')
library('foreign')

brdata5<-data.frame(brdata3)
levels(brdata5$LVI)[3]<-NA
levels(brdata5$ERPOSNEG)[3]<-NA
brdata5$GRADECAT<-factor(brdata5$GRADE)

bcss.cox1<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)
bcss.cox2<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)
os.cox1<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)
os.cox2<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)

summary(bcss.cox1)
summary(bcss.cox2)
summary(os.cox1)
summary(os.cox2)

coef(bcss.cox1)
coef(bcss.cox2)
coef(os.cox1)
coef(os.cox2)
```

B.2 R code for Residual Analysis

B.2.1 The Martingale Residual Plots

```
nbrdata5<-
na.exclude(brdata5[,c("SURVYRS","BRDEATH2","DXAGECAT","MXBCS","TSTAGE1"
,"NODEREMC","SYSTXCAT","NODESURG","LNR25SUR","ERPOSNEG","HISTCAT","GRAD
ECAT","LVI")])

par(mfrow=c(3,4))
attach(nbrdata5)
bcss.fit1a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+NODESUR
G+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(DXAGECAT,resid(bcss.fit1a))
bcss.fit2a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+TSTAGE1+NODEREMC+SYSTXCAT+NODE
SURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(MXBCS,resid(bcss.fit2a))
bcss.fit3a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+NODEREMC+SYSTXCAT+NODESU
RG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(TSTAGE1,resid(bcss.fit3a))
bcss.fit4a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+SYSTXCAT+NODESUR
G+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(NODEREMC,resid(bcss.fit4a))
bcss.fit5a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+NODESUR
G+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(SYSTXCAT,resid(bcss.fit5a))
bcss.fit6a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(NODESURG,resid(bcss.fit6a))
bcss.fit7a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(ERPOSNEG,resid(bcss.fit7a))
bcss.fit8a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(HISTCAT,resid(bcss.fit8a))
bcss.fit9a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+LVI,data=nbrdata5)
scatter.smooth(GRADECAT,resid(bcss.fit9a))
bcss.fit10a<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+GRADECAT,data=nbrdata5)
scatter.smooth(LVI,resid(bcss.fit10a))

par(mfrow=c(3,4))
attach(nbrdata5)
```

```
bcss.fit1b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+LNR25SU
R+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(DXAGECAT,resid(bcss.fit1b))
bcss.fit2b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+TSTAGE1+NODEREMC+SYSTXCAT+LNR2
5SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(MXBCS,resid(bcss.fit2b))
bcss.fit3b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+NODEREMC+SYSTXCAT+LNR25S
UR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(TSTAGE1,resid(bcss.fit3b))
bcss.fit4b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+SYSTXCAT+LNR25SU
R+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(NODEREMC,resid(bcss.fit4b))
bcss.fit5b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+LNR25SU
R+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(SYSTXCAT,resid(bcss.fit5b))
bcss.fit6b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(LNR25SUR,resid(bcss.fit6b))
bcss.fit7b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(ERPOSNEG,resid(bcss.fit7b))
bcss.fit8b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(HISTCAT,resid(bcss.fit8b))
bcss.fit9b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+LVI,data=nbrdata5)
scatter.smooth(GRADECAT,resid(bcss.fit9b))
bcss.fit10b<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT,data=nbrdata5)
scatter.smooth(LVI,resid(bcss.fit10b))


par(mfrow=c(3,4))
attach(nbrdata5)
os.fit1a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+NODESUR
G+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(DXAGECAT,resid(os.fit1a))
os.fit2a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+TSTAGE1+NODEREMC+SYSTXCAT+NODE
SURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(MXBCS,resid(os.fit2a))
os.fit3a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+NODEREMC+SYSTXCAT+NODESU
RG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(TSTAGE1,resid(os.fit3a))
```

```
os.fit4a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+SYSTXCAT+NODESUR
G+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(NODEREMC,resid(os.fit4a))
os.fit5a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+NODESUR
G+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(SYSTXCAT,resid(os.fit5a))
os.fit6a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(NODESURG,resid(os.fit6a))
os.fit7a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(ERPOSNEG,resid(os.fit7a))
os.fit8a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(HISTCAT,resid(os.fit8a))
os.fit9a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+LVI,data=nbrdata5)
scatter.smooth(GRADECAT,resid(os.fit9a))
os.fit10a<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+GRADECAT,data=nbrdata5)
scatter.smooth(LVI,resid(os.fit10a))

par(mfrow=c(3,4))
attach(nbrdata5)
os.fit1b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+LNR25SU
R+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(DXAGECAT,resid(os.fit1b))
os.fit2b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+TSTAGE1+NODEREMC+SYSTXCAT+LNR2
5SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(MXBCS,resid(bcss.fit2b))
os.fit3b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+NODEREMC+SYSTXCAT+LNR25S
UR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(TSTAGE1,resid(os.fit3b))
os.fit4b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+SYSTXCAT+LNR25SU
R+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(NODEREMC,resid(os.fit4b))
os.fit5b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+LNR25SU
R+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(SYSTXCAT,resid(os.fit5b))
os.fit6b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(LNR25SUR,resid(os.fit6b))
```

```
os.fit7b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+HISTCAT+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(ERPOSNEG,resid(os.fit7b))
os.fit8b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+GRADECAT+LVI,data=nbrdata5)
scatter.smooth(HISTCAT,resid(os.fit8b))
os.fit9b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+LVI,data=nbrdata5)
scatter.smooth(GRADECAT,resid(os.fit9b))
os.fit10b<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT,data=nbrdata5)
scatter.smooth(LVI,resid(os.fit10b))
```

B.2.2 The Deviance Residuals Plots

```
plot(resid(bcss.cox1,type="deviance"))
title("Deviance Residuals for BCSS with number of positive nodes")
plot(resid(bcss.cox2,type="deviance"))
title("Deviance Residuals for BCSS with LNR")
plot(resid(os.cox1,type="deviance"))
title("Deviance Residuals for OS without LNR")
plot(resid(os.cox2,type="deviance"))
title("Deviance Residuals for OS with LNR")
```

B.2.3 The Influence Plots

```
par(mfrow=c(3,5))
bresid<-resid(bcss.cox1,type="dfbetas")
plot(1:7814,bresid[,1],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("DXAGECAT>=50")
plot(1:7814,bresid[,2],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("MXBCS both Mx and Bcs")
plot(1:7814,bresid[,3],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("TSTAGE1 T2=2 to 5cm")
plot(1:7814,bresid[,4],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODEREMC>15")
plot(1:7814,bresid[,5],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("SYSTXCAT systemic therapy")
plot(1:7814,bresid[,6],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1A,nodecat=1")
```

```
plot(1:7814,bresid[,7],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1A,nodecat=2")
plot(1:7814,bresid[,8],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1B,nodecat=1")
plot(1:7814,bresid[,9],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1B,nodecat=2")
plot(1:7814,bresid[,10],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("ERPOSNEG positive")
plot(1:7814,bresid[,11],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT lobular")
plot(1:7814,bresid[,12],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT other")
plot(1:7814,bresid[,13],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT2")
plot(1:7814,bresid[,14],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT3")
plot(1:7814,bresid[,15],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LVI negative")

par(mfrow=c(3,5))
bresid<-resid(bcss.cox2,type="dfbetas")
plot(1:7814,bresid[,1],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("DXAGECAT>=50")
plot(1:7814,bresid[,2],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("MXBCS both Mx and Bcs")
plot(1:7814,bresid[,3],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("TSTAGE1 T2=2 to 5cm")
plot(1:7814,bresid[,4],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODEREMC>15")
plot(1:7814,bresid[,5],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("SYSTXCAT systemic therapy")
plot(1:7814,bresid[,6],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1A,.01<lnr25<=.25")
plot(1:7814,bresid[,7],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1A,lnr25>.25")
plot(1:7814,bresid[,8],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1B,.01<lnr25<=.25")
plot(1:7814,bresid[,9],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1B,lnr25>.25")
```

```
plot(1:7814,bresid[,10],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("ERPOSNEG positive")
plot(1:7814,bresid[,11],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT lobular")
plot(1:7814,bresid[,12],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT other")
plot(1:7814,bresid[,13],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT2")
plot(1:7814,bresid[,14],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT3")
plot(1:7814,bresid[,15],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LVI negative")

par(mfrow=c(3,5))
bresid<-resid(os.cox1,type="dfbetas")
plot(1:7814,bresid[,1],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("DXAGECAT>=50")
plot(1:7814,bresid[,2],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("MXBCS both Mx and Bcs")
plot(1:7814,bresid[,3],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("TSTAGE1 T2=2 to 5cm")
plot(1:7814,bresid[,4],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODEREMC>15")
plot(1:7814,bresid[,5],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("SYSTXCAT systemic therapy")
plot(1:7814,bresid[,6],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1A,nodecat=1")
plot(1:7814,bresid[,7],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1A,nodecat=2")
plot(1:7814,bresid[,8],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1B,nodecat=1")
plot(1:7814,bresid[,9],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODESURG,surgn=1B,nodecat=2")
plot(1:7814,bresid[,10],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("ERPOSNEG positive")
plot(1:7814,bresid[,11],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT lobular")
plot(1:7814,bresid[,12],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT other")
```

```
plot(1:7814,bresid[,13],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT2")
plot(1:7814,bresid[,14],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT3")
plot(1:7814,bresid[,15],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LVI negative")

par(mfrow=c(3,5))
bresid<-resid(os.cox2,type="dfbetas")
plot(1:7814,bresid[,1],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("DXAGECAT>=50")
plot(1:7814,bresid[,2],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("MXBCS both Mx and Bcs")
plot(1:7814,bresid[,3],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("TSTAGE1 T2=2 to 5cm")
plot(1:7814,bresid[,4],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("NODEREMC>15")
plot(1:7814,bresid[,5],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("SYSTXCAT systemic therapy")
plot(1:7814,bresid[,6],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1A,.01<lnr25<=.25")
plot(1:7814,bresid[,7],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1A,lnr25>.25")
plot(1:7814,bresid[,8],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1B,.01<lnr25<=.25")
plot(1:7814,bresid[,9],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LNR25SURG,surgn=1B,lnr25>.25")
plot(1:7814,bresid[,10],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("ERPOSNEG positive")
plot(1:7814,bresid[,11],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT lobular")
plot(1:7814,bresid[,12],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("HISTCAT other")
plot(1:7814,bresid[,13],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT2")
plot(1:7814,bresid[,14],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("GRADECAT3")
plot(1:7814,bresid[,15],type="h",ylab="Scaled change in
coef",xlab="Observation")
title("LVI negative")
```

B.2.4 The Rescaled Schoenfeld Residuals Plots and Corresponding Tests

```
par(mfrow=c(3,5))
plot(cox.zph(bcss.cox1))
cox.zph(bcss.cox1)

par(mfrow=c(3,5))
plot(cox.zph(bcss.cox2))
cox.zph(bcss.cox2)

par(mfrow=c(3,5))
plot(cox.zph(os.cox1))
cox.zph(os.cox1)

par(mfrow=c(3,5))
plot(cox.zph(os.cox2))
cox.zph(os.cox2)
```

B.3 R code for Adjusted and Unadjusted Survival Analysis

B.3.1 Unadjusted KM Survival

```
km1<-survfit(Surv(SURVYRS,BRDEATH2==1)~NODESURG)
km2<-survfit(Surv(SURVYRS,BRDEATH2==1)~LNR25SUR)
km3<-survfit(Surv(SURVYRS,BRDEATH2!=3)~NODESURG)
km4<-survfit(Surv(SURVYRS,BRDEATH2!=3)~LNR25SUR)

nstrat<-length(km1$ntimes.strata)
jj<-
tapply(km1$surv[km1$time<=5],rep(1:nstrat,km1$ntimes.strata)[km1$time<=
5],'min')
jj<-
tapply(km1$surv[km1$time<=10],rep(1:nstrat,km1$ntimes.strata)[km1$time<
=10],'min')
jj

se5<-
tapply(km1$std.err[km1$time<=5],rep(1:nstrat,km1$ntimes.strata)[km1$tim
e<=5],'min')
se5
se10<-
tapply(km1$std.err[km1$time<=10],rep(1:nstrat,km1$ntimes.strata)[km1$ti
me<=10],'min')
se10

nstrat<-length(km2$ntimes.strata)
jj<-
tapply(km2$surv[km2$time<=5],rep(1:nstrat,km2$ntimes.strata)[km2$time<=
5],'min')
```

```
jj<-
tapply(km2$surv[km2$time<=10],rep(1:nstrat,km2$ntimes.strata)[km2$time<
=10],'min')
jj
nstrat<-length(km3$ntimes.strata)
jj<-
tapply(km3$surv[km3$time<=5],rep(1:nstrat,km3$ntimes.strata)[km3$time<=
5],'min')
jj<-
tapply(km3$surv[km3$time<=10],rep(1:nstrat,km3$ntimes.strata)[km3$time<
=10],'min')
jj
nstrat<-length(km4$ntimes.strata)
jj<-
tapply(km4$surv[km4$time<=5],rep(1:nstrat,km4$ntimes.strata)[km4$time<=
5],'min')
jj<-
tapply(km4$surv[km4$time<=10],rep(1:nstrat,km4$ntimes.strata)[km4$time<
=10],'min')
jj

plot(km1,xlab="years",ylab="survival prob",col=1:5,lty=1:5)
title("Unadjusted KM BCSS with number of positive nodes")
legend(0.3,c("surgn=0, nodecat=0","surgn=1A, nodecat=1","surgn=1A,
nodecat=2","surgn=1B, nodecat=1","surgn=1B, nodecat=2"),col=1:5,
lty=1:5)

plot(km2,xlab="years",ylab="survival prob",col=1:5,lty=1:5)
title("Unadjusted KM BCSS with LNR")
legend(0.3,c("surgn=0, lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),col=1:5,
lty=1:5)

plot(km3,xlab="years",ylab="survival prob",col=1:5,lty=1:5)
title("Unadjusted KM OS with number of positive nodes")
legend(0.3,c("surgn=0, nodecat=0","surgn=1A, nodecat=1","surgn=1A,
nodecat=2","surgn=1B, nodecat=1","surgn=1B, nodecat=2"),col=1:5,
lty=1:5)

plot(km4,xlab="years",ylab="survival prob",col=1:5,lty=1:5)
title("Unadjusted KM OS with LNR")
legend(0.3,c("surgn=0, lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),col=1:5,
lty=1:5)
```

B.3.2 Adjusted KM Survival with Missing Values Removed

```
avg.surv <- function(cfit, var.name, var.values, data, weights)
{
#"Corrected group prognostic" method for adjusting survival curves
#Rollin Brant - <http://stat.ubc.ca/~rollin/>
#Nieto, F.J., Coresh, J. (1996), Adjusting survival curves for
#confounders: a review and a new method, \fIAmerican Journal of
#Epidemiology\fP, \fB143:10\fP, 1059-1068.
      if(missing(data)) {
```

```
            if(!is.null(cfit$model))
                  mframe <- cfit$model
            else mframe <- model.frame(cfit, sys.parent())
      }   else mframe <- model.frame(cfit, data)
      var.num <- match(var.name, names(mframe))
     data.patterns <- apply(data.matrix(mframe[,  - c(1, var.num)]), 1,
            paste, collapse = ",")
        data.patterns <-
factor(data.patterns,levels=unique(data.patterns))
      if(missing(weights))
            weights <- table(data.patterns)
      else weights <- tapply(weights, data.patterns, sum)
        kp <- !duplicated(data.patterns)
      mframe <- mframe[kp,]
        obs.var <- mframe[,var.num]
        lps <- (cfit$linear.predictor)[kp]
        tframe <- mframe[rep(1,length(var.values)),]
        tframe[,var.num] <- var.values
        xmat <- model.matrix(cfit,data=tframe)[,-1]
        tlps <- as.vector(xmat%*%cfit$coef)
        names(tlps) <- var.values
        obs.off <- tlps[as.character(obs.var)]
        explp.off <- exp(outer(lps - obs.off ,tlps,"+"))
      bfit <- survfit(cfit, se.fit = F)
        fits <- outer(bfit$surv,explp.off,function(x,y) x^y)
        avg.fits <-
           apply(sweep(fits,2,weights,"*"),c(1,3),sum)/sum(weights)
        dimnames(avg.fits) <- list(NULL,var.values)
        list(time=bfit$time,fits=avg.fits)
}

##adjusted for "NODESURG" and "LNR25SUR"#########

bcss.cox1<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)
bcss.cox2<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)
os.cox1<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)
os.cox2<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5)

fit1<-avg.surv(bcss.cox1,"NODESURG",c("surgn=0, nodecat=0","surgn=1A,
nodecat=1","surgn=1A, nodecat=2","surgn=1B, nodecat=1","surgn=1B,
nodecat=2"),brdata5)
matplot(fit1$time,fit1$fits,col=1:5,lty=1:5,type="l",xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted BCSS with number of positive nodes
(with missing values removed)")
legend(0.3,c("surgn=0, nodecat=0","surgn=1A, nodecat=1","surgn=1A,
nodecat=2","surgn=1B, nodecat=1","surgn=1B, nodecat=2"),col=1:5,
lty=1:5)
```

```
fit2<-avg.surv(bcss.cox2,"LNR25SUR",c("surgn=0,
lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),brdata5)
matplot(fit2$time,fit2$fits,col=1:5,lty=1:5,type='l',xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted BCSS with LNR
(with missing values removed)")
legend(0.3,c("surgn=0, lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),col=1:5,
lty=1:5)

fit3<-avg.surv(os.cox1,"NODESURG",c("surgn=0, nodecat=0","surgn=1A,
nodecat=1","surgn=1A, nodecat=2","surgn=1B, nodecat=1","surgn=1B,
nodecat=2"),brdata5)
matplot(fit3$time,fit3$fits,col=1:5,lty=1:5,type='l',xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted OS with number of positive nodes
(with missing values removed)")
legend(0.3,c("surgn=0, nodecat=0","surgn=1A, nodecat=1","surgn=1A,
nodecat=2","surgn=1B, nodecat=1","surgn=1B, nodecat=2"),col=1:5,
lty=1:5)

fit4<-avg.surv(os.cox2,"LNR25SUR",c("surgn=0,
lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),brdata5)
matplot(fit4$time,fit4$fits,col=1:5,lty=1:5,type='l',xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted OS with LNR
(with missing values removed)")
legend(0.3,c("surgn=0, lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),col=1:5,
lty=1:5)

summary(fit1)

#5-yr BCSS
fit1$fits[443]
fit1$fits[1263]
fit1$fits[2083]
fit1$fits[2903]
fit1$fits[3723]

fit2$fits[443]
fit2$fits[1263]
fit2$fits[2083]
fit2$fits[2903]
fit2$fits[3723]

#10-year BCSS
fit1$fits[755]
fit1$fits[1575]
fit1$fits[2395]
fit1$fits[3215]
fit1$fits[4035]

fit2$fits[755]
fit2$fits[1575]
```

```
fit2$fits[2395]
fit2$fits[3215]
fit2$fits[4035]

#5-yr BCSS
fit3$fits[672]
fit3$fits[2072]
fit3$fits[3472]
fit3$fits[4872]
fit3$fits[6272]

fit4$fits[672]
fit4$fits[2072]
fit4$fits[3472]
fit4$fits[4872]
fit4$fits[6272]

#10-year BCSS
fit3$fits[1228]
fit3$fits[2628]
fit3$fits[4028]
fit3$fits[5428]
fit3$fits[6828]

fit4$fits[1228]
fit4$fits[2628]
fit4$fits[4028]
fit4$fits[5428]
fit4$fits[6828]
```

B.3.3 Adjusted KM Survival with Missing Values Retained

```
##adjusted for "NODESURG" and "LNR25SUR"#########

bcss.cox1<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSN_A+HISTCAT+GRADECATNEW+LVINEW,data=brdata5)
bcss.cox2<-
coxph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSN_A+HISTCAT+GRADECATNEW+LVINEW,data=brdata5)
os.cox1<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+ERPOSN_A+HISTCAT+GRADECATNEW+LVINEW,data=brdata5)
os.cox2<-
coxph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+LNR25SUR+ERPOSN_A+HISTCAT+GRADECATNEW+LVINEW,data=brdata5)

summary(bcss.cox1)
summary(bcss.cox2)
summary(os.cox1)
summary(os.cox2)

fit1<-avg.surv(bcss.cox1,"NODESURG",c("surgn=0, nodecat=0","surgn=1A,
nodecat=1","surgn=1A, nodecat=2","surgn=1B, nodecat=1","surgn=1B,
nodecat=2"),brdata5)
```

```
matplot(fit1$time,fit1$fits,col=1:5,lty=1:5,type="l",xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted BCSS with number of positive nodes
(with missing values retained)")
legend(0.3,c("surgn=0, nodecat=0","surgn=1A, nodecat=1","surgn=1A,
nodecat=2","surgn=1B, nodecat=1","surgn=1B, nodecat=2"),col=1:5,
lty=1:5)

fit2<-avg.surv(bcss.cox2,"LNR25SUR",c("surgn=0,
lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),brdata5)
matplot(fit2$time,fit2$fits,col=1:5,lty=1:5,type='l',xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted BCSS with LNR
(with missing values retained)")
legend(0.3,c("surgn=0, lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),col=1:5,
lty=1:5)

fit3<-avg.surv(os.cox1,"NODESURG",c("surgn=0, nodecat=0","surgn=1A,
nodecat=1","surgn=1A, nodecat=2","surgn=1B, nodecat=1","surgn=1B,
nodecat=2"),brdata5)
matplot(fit3$time,fit3$fits,col=1:5,lty=1:5,type='l',xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted OS with number of positive nodes
(with missing values retained)")
legend(0.3,c("surgn=0, nodecat=0","surgn=1A, nodecat=1","surgn=1A,
nodecat=2","surgn=1B, nodecat=1","surgn=1B, nodecat=2"),col=1:5,
lty=1:5)

fit4<-avg.surv(os.cox2,"LNR25SUR",c("surgn=0,
lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),brdata5)
matplot(fit4$time,fit4$fits,col=1:5,lty=1:5,type='l',xlab='Years',ylab=
'Survival prob',ylim=c(0,1))
title("Adjusted OS with LNR
(with missing values retained)")
legend(0.3,c("surgn=0, lnr25=0","surgn=1A, .01<lnr25<=.25","surgn=1A,
lnr25>.25","surgn=1B, .01<lnr25<=.25","surgn=1B, lnr25>.25"),col=1:5,
lty=1:5)

#5-yr BCSS
fit1$fits[512]
fit1$fits[512+979]
fit1$fits[512+979+979]
fit1$fits[512+979+979+979]
fit1$fits[512+979+979+979+979]

fit2$fits[512]
fit2$fits[512+979]
fit2$fits[512+979+979]
fit2$fits[512+979+979+979]
fit2$fits[512+979+979+979+979]

#10-year BCSS
fit1$fits[887]
fit1$fits[887+979]
```

```
fit1$fits[887+979+979]
fit1$fits[887+979+979+979]
fit1$fits[887+979+979+979+979]

fit2$fits[887]
fit2$fits[887+979]
fit2$fits[887+979+979]
fit2$fits[887+979+979+979]
fit2$fits[887+979+979+979+979]

#5-yr BCSS
fit3$fits[774]
fit3$fits[774+1718]
fit3$fits[774+1718+1718]
fit3$fits[774+1718+1718+1718]
fit3$fits[774+1718+1718+1718+1718]

fit4$fits[774]
fit4$fits[774+1718]
fit4$fits[774+1718+1718]
fit4$fits[774+1718+1718+1718]
fit4$fits[774+1718+1718+1718+1718]

#10-year BCSS
fit3$fits[1457]
fit3$fits[1457+1718]
fit3$fits[1457+1718+1718]
fit3$fits[1457+1718+1718+1718]
fit3$fits[1457+1718+1718+1718+1718]

fit4$fits[1457]
fit4$fits[1457+1718]
fit4$fits[1457+1718+1718]
fit4$fits[1457+1718+1718+1718]
fit4$fits[1457+1718+1718+1718+1718]
```

B.4 R code for Survival Trees

B.4.1 BCSS Tree

```
library('rpart')
levels(brdata5$LNR25SUR)<-c("(0,0)","(1A,S)","(1A,L)","(1B,S)","(1B,L)")

#BCSS
m.tree.cp004.b<-
rpart(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+LNR25SUR+

ERPOSN_A+HISTCAT+GRADENEWCAT+LVINEW,data=brdata5,control=rpart.control(
cp=0.002))
m.tree.cp004.b
```

```
# To plot:
plot(m.tree.cp004.b,uniform=TRUE,branch=1,compress=FALSE,margin=0.05);
text(m.tree.cp004.b,splits=TRUE,pretty=0,use.n=TRUE,fancy=FALSE,
cex=.7);


#Plot KM curves for groupings defined by tree leaves
gpecnm.tree.km.b<-
survfit(Surv(SURVYRS,BRDEATH2==1)~m.tree.cp004.b$where,
   data=brdata5)
nstrat.b<-length(gpecnm.tree.km.b$ntimes.strata)
plot(gpecnm.tree.km.b, lty=1:nstrat.b, col=1:nstrat.b)
title("BCSS KM curves for groups defined by tree")
legend(locator(1),paste(1:length(unique(m.tree.cp004.b$where))),lty=1:n
strat.b
      ,col=1:nstrat.b)

#log rank test

survdiff(Surv(SURVYRS,BRDEATH2==1)~m.tree.cp004.b$where,data=brdata5)


#Compute 10-year survival
jkm.b<-gpecnm.tree.km.b
nstrat.b<-length(jkm.b$ntimes.strata)
jj<-
tapply(jkm.b$surv[jkm.b$time<=10],rep(1:nstrat.b,jkm.b$ntimes.strata)[j
km.b$time<=10],'min')
jj
j<-m.tree.cp004.b
j$frame$yval2[j$frame$var=='<leaf>',1]<-round(jj,2)
plot(j,uniform=TRUE,branch=1,compress=FALSE, margin=.05);
text(j,splits=TRUE,pretty=0,use.n=TRUE,fancy=FALSE, cex=.7);
title('BCSS tree, all variables, 10yr survival, #events/#pts')
```

B.4.2 OS Tree

```
#OS
m.tree.cp004.o<-
rpart(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCA
T+NODESURG+LNR25SUR+

ERPOSN_A+HISTCAT+GRADENEWCAT+LVINEW,data=brdata5,control=rpart.control(
cp=0.002))
m.tree.cp004.o

# To plot:
plot(m.tree.cp004.o,uniform=TRUE,branch=1,compress=FALSE,margin=0.05);
text(m.tree.cp004.o,splits=TRUE,pretty=0,use.n=TRUE,fancy=FALSE,
cex=.7);


#Plot KM curves for groupings defined by tree leaves
```

```
gpecnm.tree.km.o<-
survfit(Surv(SURVYRS,BRDEATH2!=3)~m.tree.cp004.o$where,
   data=brdata5)
nstrat.o<-length(gpecnm.tree.km.o$ntimes.strata)
plot(gpecnm.tree.km.o, lty=1:nstrat.o, col=1:nstrat.o)
title("OS KM curves for groups defined by tree")
legend(locator(1),paste(1:length(unique(m.tree.cp004.o$where))),lty=1:n
strat.o
      ,col=1:nstrat.o)

#log rank test

survdiff(Surv(SURVYRS,BRDEATH2!=3)~m.tree.cp004.o$where,data=brdata5)

#Compute 10-year survival
jkm.o<-gpecnm.tree.km.o
nstrat.o<-length(jkm.o$ntimes.strata)
jj<-
tapply(jkm.o$surv[jkm.o$time<=10],rep(1:nstrat.o,jkm.o$ntimes.strata)[j
km.o$time<=10],'min')
jj
j<-m.tree.cp004.o
j$frame$yval2[j$frame$var=='<leaf>',1]<-round(jj,2)
plot(j,uniform=TRUE,branch=1,compress=FALSE, margin=.05);
text(j,splits=TRUE,pretty=0,use.n=TRUE,fancy=FALSE, cex=.7);
title('OS tree, all variables, 10yr survival, #events/#pts')
```

B.5 R code for Nomograms

```
attach(brdata5)
ddist <-
datadist(DXAGECAT,MXBCS,TSTAGE1,NODEREMC,SYSTXCAT,NODESURG,LNR25SUR,ERP
OSNEG,HISTCAT,GRADECAT,LVI)
options(datadist='ddist')

fit1<-
cph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+
NODESURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5,surv=T)
surv1<-Survival(fit1)
nomogram(fit1, fun=list(function(x) surv1(10, x)),funlabel="10-year
Survival Probability", xfrac=.5)

fit2<-
cph(Surv(SURVYRS,BRDEATH2==1)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+
LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5,surv=T)
surv2<-Survival(fit2)
nomogram(fit2, fun=list(function(x) surv2(10, x)),funlabel="10-year
Survival Probability", xfrac=.5)

fit3<-
cph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+
NODESURG+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5,surv=T)
surv3<-Survival(fit3)
```

```
nomogram(fit3, fun=list(function(x) surv3(10, x)),funlabel="10-year
Survival Probability", xfrac=.5)

fit4<-
cph(Surv(SURVYRS,BRDEATH2!=3)~DXAGECAT+MXBCS+TSTAGE1+NODEREMC+SYSTXCAT+
LNR25SUR+ERPOSNEG+HISTCAT+GRADECAT+LVI,data=brdata5,surv=T)
surv4<-Survival(fit4)
nomogram(fit4, fun=list(function(x) surv4(10, x)),funlabel="10-year
Survival Probability", xfrac=.5)
```