Mega-dose Vitamins and Minerals for the Treatment of Breast Cancer: A Comparison Study of Treated Nonmetastatic Patients Versus Two Control Groups

By

Yang Zhao

B.Sc., Nankai University, 1992

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We accept this thesis as conforming

to the required standard.

[redacted]

Dr. M. Lesperance, Supervisor (Department of Mathematics & Statistics)

[redacted]

Dr. M. Tsao, Co-supervisor (Department of Mathematics & Statistics)

[redacted]

Dr. J. Thomas Buckley, Outside Member (Department of Biochemistry & Microbiology)

[redacted]

Dr. J. D. Kalbfleisch, External Examiner (University of Waterloo)

University of Victoria

Supervisor: Dr. M. Lesperance

ABSTRACT

Dr. Hoffer has been a practising psychiatrist in Victoria since 1967. During this time he has treated more than 200 female breast cancer patients with niacin, beta-carotene, selenium, vitamin C, Q10 and zinc. We compared the outcomes of Dr. Hoffer's nonmetastatic patients who were diagnosed from 1989 to 1996 with those from two sets of controls. The first set of controls was obtained as matched controls. The second set of controls includes all the female nonmetastatic breast cancer patients diagnosed at the Victoria office of the BC Cancer Agency between 1989 and 1996 who did not take large doses of vitamins as far as can be known. Using the Cox proportional hazards model and the aligned rank test we found that breast cancer patients' life times for Dr. Hoffer's patients are shorter than those of the control patients.

Examiners:

[redacted]

Dr. M. Lesperance, Supervisor (Department of Mathematics & Statistics)

[redacted]

Dr. M. Tsao, Co-supervisor (Department of Mathematics & Statistics)

[redacted]

Dr. J. Thomas Buckley, Outside Member (Department of Biochemistry & Microbiology)

[redacted]

Dr. J. D. Kalbfleisch, External Examiner (University of Waterloo)

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Mary Lesperance and Dr. Min Tsao for their encouragement, directions and comments.

Also, I want to thank Dr. J. D. Kalbfleisch and Dr. J. Thomas Buckley for examining my thesis and providing suggestions and comments.

Further thanks go to Hecht Memorial Fundation for funding to complete this project, and Miss Donna Mates, Miss Caroline Speers and Dr. Abraham Hoffer for providing me the data sets.

Finaly I thank my mother, my husband and my son for their support.

Chapter 1

Introduction

Many orthomolecular physicians prescribe large doses of vitamins, especially antioxidants, for their cancer patients. This is done in the belief that these essential nutrients enhance the functioning of the immune system. Such doctors also typically subscribe to the view that certain vitamins and minerals increase the efficacy of both irradiation and chemotherapy, while simultaneously decreasing the toxicity of these conventional treatments to normal body tissues. If this is correct and megavitamin treatment increases the efficacy of conventional cancer treatment, it is to be expected that patients receiving it, together with conventional therapy, will live longer than those patients who receive conventional therapy alone.

This research project endeavours to shed light on the question of whether breast cancer patients' life expectancies are increased by the addition of an orthomolecular program, consisting of mega-doses of beta-carotene, and vitamins C and E, to their conventional treatment protocol.

Dr. Abraham Hoffer has been a practising psychiatrist in Victoria since 1967. During this time he has treated more than 200 women breast cancer patients with niacin, beta-carotene, selenium, vitamin C, coenzyme Q10 and zinc. We compared the outcomes of his patients, who were diagnosed from 1989 to 1996, with those from two sets of controls.

The first set of controls was obtained as matched controls where the matching was based on year of diagnosis, age at diagnosis, TNM stage (M stage, T stage and N stage), conventional therapy (surgery, radiation, chemotherapy and hormone therapy), tumor grade, estrogen receptor status and lymphatic, vascular or neural invasion. The data set that contains Dr. Hoffer's patients and the matched controls is called Data set I.

The second set of controls includes all women nonmetastatic breast cancer patients diagnosed at the Victoria office of the BC Cancer Agency between 1989 and 1996 who did not take large doses of vitamins as far as can be known, which together with Dr. Hoffer's patients are called Data set II. We analyze the pair-matched data set first and use Data set II to test the consistency of the analysis.

Chapter 2 provides a very brief review of some literature on the use of vitamins, including beta-carotene and vitamins C and E, for the treatment of breast cancer. In Chapter 3 the pair-matched Breast Cancer data set is introduced. Chapter 4 contains information about data checking and the preliminary analysis which includes paired-t tests and the Pearson Chi-square tests for evaluating the matching within each pair, base line homogeneity tests and the plots of the Kaplan-Meier survivor functions. Chapter 5 introduces the Cox proportional hazards model with a time-dependent treatment effect, the parameter estimates, tests of the proportional hazards assumption, and plots of martingale

residuals. In Chapter 6, we use the aligned rank test to compare the pair-matched survival data. Chapter 7 contains the Cox model and the results for Data set II. The time to distant relapse is analyzed in Chapter 8 for Data set I and Data set II separately. Conclusions and comments are given in Chapter 9.

Chapter 2

Literature review on the use of vitamins for the treatment of cancer

In this chapter, I provide a very brief review of some literature on the use of vitamins for the treatment of cancer.

Vitamin A

Moriguchi and Kishino (1990) and Frey et al (1991) reported that several natural and synthetic vitamin A derivatives (retinoids) inhibit tumour development. Similar reactions have also been shown in some clinical studies (Lippman et al, 1992). The anticancer impacts of retinoids may perhaps be due to their ability to reduce the rate at which cancer cells grow and multiply. Retinoids also may promote cell differentiation, by changing the structure and function of oncogeneses. Animal research has also shown the ability

of vitamin A and retinoids to enhance immune response, slow the growth of tumours and even decrease their size (Bendick, 1991; Bollage and Holdener, 1992; Rustin, 1992; Ip and Ip, 1981). However, retinoids appear to differ in their impacts on specific tumours, which also vary with the animal species being studied. Although in human cell lines and patient studies, retinoids appear to have the greatest impact on cancers of the upper-bowel, bladder, cervix and airways, fenretinide shows promise in both breast cancer prevention and treatment (Cobleigh, 1994).

In a study to determine the efficacy of fenretinide, a vitamin A analogue, in preventing a second breast malignancy in women with breast cancer, 2972 women, aged 30-70 years, with surgically removed stage I breast cancer or ductal carcinoma in situ were randomly assigned to 5 years either fenretinide orally (200 mg/day) or no treatment. There were no statistically significant differences in the occurrence of contralateral breast cancer or ipsilateral breast cancer between the two groups at a median observation time of 97 months. And in the report the authors concluded that fenretinide treatment of women with breast cancer for 5 years appears to have no statistical significant effect on the incidence of second breast malignancies overall, although a possible benefit was detected in premenopausal women (Veronesi et al, 1999).

Vitamin C

In 1972, Stone conducted a detailed literature review which led him to conclude that vitamin C was of value in the treatment of a variety of cancers. This viewpoint was supported by Cameron and Pauling (1979) and, as a result, vitamin C has been widely used by orthomolecular physicians to treat breast cancer since this time. Cameron and

Pauling (1993) have published clinical evidence to support this position. Some physicians are now using intravenous ascorbic acid, often in doses in excess of 100 grams, (Jackson and Riordan, 1995; Riordan et al, 1990) to treat a variety of cancers, including that of the breast. This treatment is still highly controversial (Richards, 1991).

In September 1990, at a meeting co-sponsored by the US National Cancer Institute, held in Bethesda, Maryland, four significant properties of vitamin C that might be of value in the treatment of breast cancer were identified. Firstly, ascorbic acid was shown to protect plasma lipids against oxidative damage because it scavenges free radicals. Secondly, it decreases the toxicity against normal tissues of drugs, such as adriamycin, used in chemotherapy, (Shimpo et al, 1991) and reduces the toxicity of radiation, without protecting tumours. Thirdly, ascorbic acid has anti-cancer properties. To illustrate, it can protect cells against transformation into cancer by methylcholanthrene. Fourthly, 33 out of 44 epidemiological studies on the topic have supported the position that vitamin C intake decreases cancer incidence, high serum levels being apparently protective. This is interesting since, as early 1948, Goth and Litmann argued that the cancer only began to develop in organs whose ascorbic acid levels were depressed.

Vitamin E

Epidemiological research and clinical observations suggest that there may be an inverse relationship between vitamin E levels in the blood and cancer development (Longnecker et al 1992; Knekt, 1991; Palan et al, 1991 and Knekt et al, 1991). Research attempting to establish such links for breast cancer is inconsistent, but the results of cellular studies examining the effects of vitamin E succinate on mammary cancers have been

promising (Kline, 1994; Gopalakrishna, 1994 and Turley et al, 1997). Supplementation of 800 mg/day of alpha-tocopherol also appears to reduce the adverse side effects of radiation in breast cancer patients (Canadian Breast Cancer Research Initiative, 1996).

Combinations

In 1996, a study was designed to evaluate the modifying effects of vitamin A, E and selenium serum concentrations and glutathione peroxidase activity on preoperative radio- and chemotherapy of breast cancer by Hartmann et al. They concluded that serum concentrations of antioxidants had no modifying influence on tumor response in breast cancer patients following induction radio-chemotherapy and subsequent surgery in their study.

In summary, the literature on the value of beta-carotene, vitamins C and E for the treatment of breast cancer is not conclusive. However, it is suggestive and appears to justify a detailed statistical comparison of the case histories of those women who have chosen this form of complementary treatment in addition to conventional treatment with those who have not.

Chapter 3

Description of the breast cancer Data set I

3.1 Data collection and description of the variables

This data set was provided by the BC Cancer Agency's Breast Cancer Outcome Unit in Vancouver and Dr. Abram Hoffer. Dr. Hoffer treated his breast cancer patients with large doses of niacin, beta-carotene, selenium, vitamin C, coenzyme Q10 and zinc. He provided data on 271 of his patients who had seen him for treatment for their breast cancer. His file contains the patients' health number ($id2$), name, date of birth, sex, organ (breast or not), symptoms onset, date first seen, the treatment variables which include surgery, radiation, chemotherapy and coenzyme Q10 with values Yes or No, doses of vitamin C, niacin, selenium, beta-carotene and zinc separately, and date of death if applicable.

All of the people in Dr. Hoffer's file were also patients of the BC Cancer Agency

and had undergone conventional therapy. These patients formed the treated group. To investigate whether breast cancer patients' life expectancies are increased by the addition of the orthomolecular treatment, a matched control group was generated from the BC Cancer Agency Information System (CAIS) by Donna Mates. For each of the patients in the treated group, a control patient who did not take large doses of vitamins as far as can be known was selected from the BC Cancer Agency Information System database based on the matching of the following variables. The variables are listed in the order of importance for the matching when they were matched:

- *agedx* is the age at diagnosis. Calculated in the warehouse tables by subtracting birth date from diagnosis date (divide by 365). If the day part of the birthday has been estimated, then subtract 15 days from the total days to estimate the middle of the month. If the month of the birthday has been estimated for the birthday, then subtract 182 days from the total days, to estimate the middle of the year.

- *dxyear* is the year from the diagnosis date. Diagnosis date is defined as: the earliest date of the following sources: class IV/ V cytology or positive Pathology or Autopsy Report. If none of the above confirms the malignancy/condition, the date of the first positive lab is used. If there is no positive lab, the date diagnosis was confirmed by a physician clinically was used. The date of positive radiology is used for non-referred cases when no other information is available.

- *stagem* is TNM pathological M stage. If pathological is missing or cannot assess then the clinical M stage is entered in this variable. A 1 denotes distant metastasis, 0 denotes no distant metastasis.

- *staget* is coded from TNM pathological stage, if missing then coded from size of lesion, if still missing then coded from TNM clinical T stage. A -1 denotes carcinoma in-situ: intraductal carcinoma, or lobular carcinoma in-situ, or Paget's disease of the nipple with no tumour (Paget's disease associated with tumour is classified according to the size of the tumour), 0 denotes no evidence of primary tumour, 1 to 3 gives the categorical tumour size in increasing order (2 cm or less, more than 2 cm and but not more than 5 cm, and more than 5 cm), and 4 denotes tumour of any size with direct extension to chest wall or skin.

- *stagepn* is TNM surgical N stage. A code which corresponds to the absence or presence and extent of regional lymph node metastasis. Pathological Classification based on evidence acquired before treatment supplemented or modified by evidence acquired from surgery and pathological examination. A -1 denotes no axillary dissection or cannot assess, 0 denote no evidence of invasion of regional nodes, 1 denotes evidence of invasion of movable homolateral axillary lymph nodes, 2 denotes evidence of invasion of homolateral axillary lymph nodes fixed to one another or to other structures, and 3 denotes evidence of invasion of homolateral internal mammary lymph nodes.

- *bccasr* with N or Y indicates whether or not the patient had surgery. Surgical treatments that relate directly to treatment of the cancer the patient was referred for are recorded according to Site/Tumour Group specifications.

- *bccard* with N or Y indicates whether or not the patient received radiation therapy treatment. This may refer to either initial treatment or subsequent treatment.

- *bccach* with N or Y indicates whether or not the patient had chemotherapy treatment, 1 denotes the initial chemotherapy given as part of the initial treatment plan, 2 denotes subsequent chemotherapy is given for residual disease or at relapse for recurrent or metastatic disease, and 3 denotes chemotherapy given as both initial and further subsequent treatment. Note: Cancer Registry: Chemotherapy includes Hormone Therapy.

- *bccahr* indicates whether or not the patient received hormone therapy. Hormone treatment includes: 1) hormones given to inhibit tumour growth; 2) drugs given that will decrease natural hormone production within a patient in order to inhibit the tumour growth. A 1 indicates hormone therapy given as part of initial treatment plan that includes continuous hormone therapy given when one treatment fails and another one is started, 2 indicates subsequent hormone therapy is given for residual disease or at relapse for recurrent/metastatic disease, 3 indicates hormone given as both initial and subsequent treatment, N indicates no hormone therapy given and Y indicates hormone therapy given but cannot differentiate between initial or subsequent treatment. Note: 1. Hormone Therapy is included with Chemotherapy for non-referred cases; 2. Hormones given for replacement purposes are not considered to be hormone treatment, e.g., pituitary, thyroid, gyne; 3. Hormones given for alleviation of symptoms are not considered hormone treatment e.g., steroids given in brain patients to reduce swelling of the brain; 4. NSABP Trial B21, B23, or B24 and NCIC MA 12 – double-blind study in which tamoxifen or placebo is given – coded as 1 (initial treatment) and note is entered on note variable.

- *dxgrade* designates the histopathological degree of differentiation of malignant neoplasms or the total number of histopathological features translated into a grade (e.g. Astrocytoma). Grade is determined from the following sources: i) Pathology Stamp on back of Pathology Review, then front of review; ii) Biopsy/Mastectomy Report. Values 1 to 3 denote the degree of differentiation of malignant neoplasms, 1 is well differentiated and 3 is poorly differentiated (the higher the grade, the faster the cancer tends to grow and spread)

  Notes: 1) Nuclear Grade takes precedence over histologic/cytologic grade. 2) Type of grade is reported in variable *dxgradet*. 3) The highest of the nuclear/histologic grade if multiple comments/specimens is coded. E.g. Grade 2-3 = Grade 3. 4) Invasive Grade is coded when there is a discrepancy between invasive and in-situ. 5) When grade is reported as 2/3, Grade 2 is entered. 6) If surgery was done after chemotherapy, hormone or radiation therapy, the Grade of the Fine Needle Aspirate is coded. 7) If a patient was diagnosed with breast cancer from another site i.e. Patient had a FNA of the axillary node done first and it is determined that patient has breast cancer, then the grade from the FNA is recorded. 8) If the front of the path review states moderately differentiated and the back of the review has a 2+ ; Grade = 2. 9) If the front of the path review states poorly differentiated and the back of the review has a 2+; Grade = 3.

- *dxer* indicates estrogen receptor status coded from Provincial ER Report or Immunohistochemical stain (er or immuno). Value 1 or 2 indicates estrogen receptor status is negative (1 thru 14 or "Negative") or positive (15+ or "low positive" or "moderately

positive" or "strongly positive"). Nov 23/93 – not recorded for benign cases.

- *dxlvn* is the status of lymphatics/veins/nerves of the breast tumour at diagnosis. Positive invasion for either lymphatics or veins or nerves or any combination results in positive value for this variable. Effective Nov 23/93 - not recorded for benign cases. Value 0 or 1 indicates no invasion or positive invasion (the worse case) of lymphatics or veins or nerves.

There were 55 patients of the 271 treated patients who were not matched: 27 of them had bilateral breast cancers for whom it would be nearly impossible to find a control with the same bilateral breast cancer characteristics and the remaining 28 unmatched are because of incomplete records in the CAIS database. 131 pairs of matched cases and controls with *dxyear* from 1989 to 1996 were available for analysis. Data recorded during this period of time are most accurate and complete (Mates, 1999).

Other important variables in the data set include:

- *dthcsurv* is the number of days from diagnosis date to death from breast cancer, or if no death from breast cancer then death any cause or last contact date with patient.
- *dsurv* is the number of days from diagnosis to first distant relapse, if no relapse then last contact date or death date (any cause).
- *dthcevt* is an event variable with 1 or 0 indicates the patient death or no death from breast cancer and -1 denotes lost to follow-up.
- *devent* is an event variable that indicates if site had a distant relapse. A -1 denotes lost to follow-up, 0 denotes no distant relapse and 1 denotes distant relapse.

- *treatment* with 1 or 0 indicates the patient had large doses of Beta-carotene, vitamin C and E treatment or not.

- *daytohof* is the number of days from diagnosis date to the date when Dr. Hoffer was first seen, which is deemed as the date on which the vitamin treatment started.

- *posnod* indicates the number of positive lymph nodes pathologically examined. Applicable to locally advanced and inflammatory cases. If discrepancy in count, count from path review is used. Not entered for 1992 (admit date); not recorded for benign cases Nov 23/93 onwards (admit date).

- *dxposnod* is a categorical variable coded from *posnod* to show if nodes are positive or not. If *posnod* is missing, then coded from *stagepn*. A -1 indicates no axillary dissection, or cannot assess, 0 indicates no positive nodes and 1 indicates positive nodes.

- *histcat* is a categorical variable for histology created from hist1, hist2 and hist3 (An International Classification of Diseases for Oncology, Second Edition (ICD-O) or Systematized Nomenclature of Medicine (SNOMED) code which describes the cell type of the malignancy/condition.). A 1 indicates ductal, 2 indicates lobular and 3 indicates other. See the tables in appendix for the coding for *histcat*.

- *sizles* indicates the actual size of the lesion (in cm). If lesion is greater than 9.9 cm 9.9 is entered and the actual size of lesion is recorded on the "note" variable. It is recorded from the pathology report, if missing, then the mammogram. If the pathologic and mammographic size are not available, then an estimate of the clinical

size by the surgeon, prior to operation, will be recorded. If lesion is stated as being greater than, e.g., 2.0 cm then 2.1 cm is entered (add 0.1 cm) and if the lesion is less than 2.0 cm then 1.9 is entered (subtract 0.1cm). If the tumour is recovered in several pieces over two operation, then the largest dimension from the largest piece is added to the smallest dimension of the smallest piece. Size of residual cavity is not used. For multifocal/multiple cancers, the dimension of the largest primary tumour is recorded. Recorded for Jan 1/93 admit date and onwards – Nov 23/93 not recorded for benign cases.

- *bccachyn* is coded from *bccach*: $bccachyn = Y$ if *bccach* is 1, 2, 3 or Y; $bccachyn = N$ if *bccach* is N. Value Y indicates the chemotherapy was given, N indicates chemotherapy was not given.

- *stapn* is coded from *stagepn*: $stapn = 0$ if *stagepn* is 0, -1. And the other levels of *stapn* are the same as *stagepn*.

- *Pair* is the number for matched patients.

- *idtrans* is the transformed agency *id* which is a unique identification number assigned to the patient upon initial contact with the BC Cancer Agency.

- *id*2 is the identification number for the patient in the treated group.

3.2 Data set cleanups

Using frequency analyses, one can see that there are 18 pairs of patients where survival times are missing, 1 patient in the treated group is male (with id2 is 728D), 1

patient had bilateral breast cancer still in the data set (with id2 is 241D), 5 patients in the treated group did not start the orthomolecular program. After these pairs were deleted only 106 pairs of observations were left in the data set. In order to consider $stagem = 1$ (with distant metastasis) patients separately from the $stagem = 0$ (without distant metastasis) patients, 18 pairs were deleted (the *id2* numbers are 158D, 250D, 390, 410, 420, 436, 492, 522, 660, 662, 712, 722, 729, 792D, 800D, 807D, 819D, 888D, 13 of which had *stagem* missing), only 88 pairs with $stagem = 0$ were left in the data set.

To check the accuracy of the response variable, the values of *dthcsurv* were compared with *dsurv*. There are two censored observations with *dsurv* values bigger than *dthcsurv*. The values of *dthcsurv* (2054 and 2065) were updated by *dsurv* (2148 and 2205) respectively for those two patients. One patient's *id2* number is 475 and another patient's *pair* number is 117 in the control group.

Chapter 4

Preliminary analysis

4.1 Data Checking

4.1.1 Check if data are well matched within pair

The control patient sample was generated by matching the 12 variables listed in Section 3.1 for each of the vitamin treated patients. We evaluated the quality of the matching for the continuous variables, *agedx* and *dxyear*, and the categorical variables, *staget*, *stagepn*, *bccasr*, *bccard*, *bccach*, *bccahr*, *dxgrade*, *dxer*, *dxlvn*, *histcat* and *dxyear* which can also be treated as a categorical variable. No test is needed of *stagem* since it is zero for all of the patients.

Paired-t test to check if the continuous variables are matched within pair

The paired-t test computes the difference between two values for each pair and tests whether the average of the differences differs from 0. Table 4.1 gives the results of the paired-t tests for *agedx* and *dxyear* separately, the P-values are all larger than 0.27.

Paired Samples Test

		Paired Differences		t	df	Sig. (2-tailed)
		Mean	Std. Error Mean			
Pair 1	Age at Diagnosis/control - Age at Diagnosis/case	.17	.16	1.098	87	.275
Pair 2	Diagnosis Year/control - Diagnosis Year/case	-3.41E-02	6.94E-02	-.491	87	.625

Table 4.1: Paired-t tests for *agedx* and *dxyear*

There is no evidence against the null hypothesis that the *agedx* and *dxyear* are both well matched within each pair.

Cross tabulation for treated versus control to check if the categorical variables are matched within pair

Tables 4.2 to 4.12 are the cross tabulations generated in SPSS version 9.0 with controls in the rows and treated in the columns for each of the 10 categorical matching variables and *histcat* which is usually believed to be associated with breast cancer survival. The variable *bccasr* was matched exactly for all the 88 pairs, *staget* was matched exactly except for one pair with a missing value, *dxyear* was matched within one year difference, and the other variables are partly matched. Table 4.13 is the summary of the matching.

Crosstab

Count

		STAGET1					
		-1	1	2	3	4	Total
T-stage	In-situ	3					3
	< 2.01 cm		48				48
	> 2 cm and < 5.01 cm			32			32
	> 5 cm				3		3
	Extended					1	1
Total		3	48	32	3	1	87

Table 4.2: Crosstab of *staget*

Crosstab

Count

		STAGEPN1				
		-1	0	1	2	Total
Pathological N-stage	No axil disect	1	2			3
	No Nodal Mets	5	33			38
	Axillary Nodal Mets	1		45		46
	Fixed Nodal Mets				1	1
Total		7	35	45	1	88

Table 4.3: Crosstab of *stagepn*

Crosstab

Count

		BCCASR1		
		N	Y	Total
bccasr - surgery or not	N	1		1
	Y		87	87
Total		1	87	88

Table 4.4: Crosstab of *bccasr*

Crosstab

Count

		BCCARD1		
		N	Y	Total
bccard - radiation therapy or not	N	20		20
	Y	4	64	68
Total		24	64	88

Table 4.5: Crosstab of *bccard*

Crosstab

Count

		BCCACH1					
		1	2	3	N	Y	Total
bccach - chemothe or not	1	41	1	3	2	1	48
	N	1	4		35		40
Total		42	5	3	37	1	88

Table 4.6: Crosstab of *bccach*

Crosstab

Count

		BCCAHR1			
		1	2	N	Total
bccahr - hormone therapy or not	1	26		3	29
	2			2	2
	N	2	6	49	57
Total		28	6	54	88

Table 4.7: Crosstab of *bccahr*

Crosstab

Count

		DXGRADE1			Total
		1	2	3	
Tumor Grade	Well Differentiated	7	1		8
	Moderately Differentiated	3	18	4	25
	Poorly Differentiated	2	3	30	35
Total		12	22	34	68

Table 4.8: Crosstab of *dxgrade*

Crosstab

Count

		DXER1		Total
		1	2	
Estrogen Receptor	Negative	18	4	22
	Positive	9	33	42
Total		27	37	64

Table 4.9: Crosstab of *dxer*

Crosstab

Count

		DXLVN1		Total
		0	1	
Invasive lvn	Negative	34	5	39
	Positive	7	32	39
Total		41	37	78

Table 4.10: Crosstab of *dxlvn*

Diagnosis Year/control * Diagnosis Year/case Crosstabulation

Count

		Diagnosis Year/case								
		89	90	91	92	93	94	95	96	Total
Diagnosis Year/control	89	8	1							9
	90	3	6	6						15
	91		2	8	3					13
	92				4	4				8
	93				5	7	5			17
	94					2	6	1		9
	95						3	7		10
	96							2	5	7
Total		11	9	14	12	13	14	10	5	88

Table 4.11: Crosstab of *dxyear*

Categorical variable for histology * HISTCAT1 Crosstabulation

Count

		HISTCAT1			
		1	2	3	Total
Categorical variable for histology	Ductal	62	4	2	68
	Lobular	14	1	2	17
	Other	3			3
Total		79	5	4	88

Table 4.12: Crosstab for *histcat*

Variable	Matched pairs	Valid pairs	Percent matched
stagem	88	88	100%
*staget**	87	87	100%
stagepn	80	88	90.9%
bccasr	88	88	100%
bccard	84	88	95.5%
bccach	76	88	86.4%
bccahr	75	88	85.2%
*dxgrade**	55	68	80.9%
*dxer**	51	64	79.7%
*dxlvn**	66	78	84.6%
histcat	63	88	71.6%
Average	73.9	83	89.0%

* Indicates there are missing values for those variables.

Table 4.13: Summary of the matching for the categorical variables

4.1.2 Test the base line homogeneity using the Pearson Chi-square test

In this section, we test the hypothesis of homogeneity of distributions of the matching variables for the treated versus the controls using the Pearson Chi-Square test. Tables 4.15 to 4.30 show the test results. One can see that except for *histcat* all the other matching variables have P-values larger than 0.20. There is no strong evidence against the assumption of homogeneity of the matching variables except for *histcat* (Tables 4.29 and 4.30). Note that in the Pearson Chi-square test for *histcat* two cells have small expected frequencies, therefore the P-values not accurate for that table; one can compare the expected counts and the observed values. From the case processing summary in Table 4.14 we see that there are missing values in *dxgrade*, *dxer* and *dxlvn*.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
case or control * Pathological N-stage	176	100.0%	0	.0%	176	100.0%
case or control * bccard - radiation therapy or not	176	100.0%	0	.0%	176	100.0%
case or control * bccach - chemotherapy or not	176	100.0%	0	.0%	176	100.0%
case or control * bccahr - hormone therapy or not	176	100.0%	0	.0%	176	100.0%
case or control * Tumor Grade	155	88.1%	21	11.9%	176	100.0%
case or control * Estrogen Receptor	145	82.4%	31	17.6%	176	100.0%
case or control * Categorical variable for histology	176	100.0%	0	.0%	176	100.0%
case or control * Invasive lvn	163	92.6%	13	7.4%	176	100.0%

Table 4.14: Case processing summary for the homogeneity tests

Crosstab

Count

		Pathological N-stage		
		No Nodal Mets or axil disect	Fixed Nodal Mets or Axillary Nodal Mets	Total
case or control	matched control	41	47	88
	Hoffer patient	42	46	88
Total		83	93	176

Table 4.15: Count of *stagepn* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.023[b]	1	.880		
Continuity Correction [a]	.000	1	1.000		
Likelihood Ratio	.023	1	.880		
Fisher's Exact Test				1.000	.500
Linear-by-Linear Association	.023	1	.880		
N of Valid Cases	176				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 41.50.

Table 4.16: Chi-square test of *stagepn*

Crosstab

Count

		bccard - radiation therapy or not		
		N	Y	Total
case or control	matched control	20	68	88
	Hoffer patient	24	64	88
Total		44	132	176

Table 4.17: Count of *bccard* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.485[b]	1	.486		
Continuity Correction[a]	.273	1	.602		
Likelihood Ratio	.485	1	.486		
Fisher's Exact Test				.602	.301
N of Valid Cases	176				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.00.

Table 4.18: Chi-square test of *bccard*

Crosstab

Count

		bccach - chemotherapy or not		
		N	Y	Total
case or control	matched control	40	48	88
	Hoffer patient	37	51	88
Total		77	99	176

Table 4.19: Count of *bccach* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.208[b]	1	.649		
Continuity Correction [a]	.092	1	.761		
Likelihood Ratio	.208	1	.648		
Fisher's Exact Test				.761	.381
N of Valid Cases	176				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 38.50.

Table 4.20: Chi-square test of *bccach*

Crosstab

Count

		bccahr - hormone therapy or not		
		N	Y	Total
case or control	matched control	57	31	88
	Hoffer patient	54	34	88
Total		111	65	176

Table 4.21: Count of *bccahr* for Chi–square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.220[b]	1	.639		
Continuity Correction [a]	.098	1	.755		
Likelihood Ratio	.220	1	.639		
Fisher's Exact Test				.755	.377
N of Valid Cases	176				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 32.50.

Table 4.22: Chi-square test of *bccahr*

Crosstab

Count

		Tumor Grade			Total
		Well Differentiated	Moderately Differentiated	Poorly Differentiated	
case or control	matched control	8	33	43	84
	Hoffer patient	12	24	35	71
Total		20	57	78	155

Table 4.23: Count of *dxgrade* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.965[a]	2	.374
Likelihood Ratio	1.963	2	.375
Linear-by-Linear Association	.668	1	.414
N of Valid Cases	155		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.16.

Table 4.24: Chi-square of *dxgrade*

Crosstab

Count

		Estrogen Receptor		
		Negative	Positive	Total
case or control	matched control	25	48	73
	Hoffer patient	32	40	72
Total		57	88	145

Table 4.25: Count of *dxer* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.580[b]	1	.209		
Continuity Correction [a]	1.182	1	.277		
Likelihood Ratio	1.583	1	.208		
Fisher's Exact Test				.236	.139
Linear-by-Linear Association	1.569	1	.210		
N of Valid Cases	145				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 28.30.

Table 4.26: Chi-square test of *dxer*

Crosstab

Count

		Invasive lvn		
		Negative	Positive	Total
case or control	matched control	41	43	84
	Hoffer patient	42	37	79
Total		83	80	163

Table 4.27: Count of *dxlvn* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.309[b]	1	.578		
Continuity Correction [a]	.159	1	.690		
Likelihood Ratio	.309	1	.578		
Fisher's Exact Test				.639	.345
Linear-by-Linear Association	.307	1	.579		
N of Valid Cases	163				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 38.77.

Table 4.28: Chi-square test of *dxlvn*

case or control * Categorical variable for histology Crosstabulation

			Categorical variable for histology			
			Ductal	Lobular	Other	Total
case or control	matched control	Count	68	17	3	88
		Expected Count	73.5	11.0	3.5	88.0
	Hoffer patient	Count	79	5	4	88
		Expected Count	73.5	11.0	3.5	88.0
Total		Count	147	22	7	176
		Expected Count	147.0	22.0	7.0	176.0

Table 4.29: Count of *histcat* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.511[a]	2	.023
Likelihood Ratio	7.883	2	.019
Linear-by-Linear Association	2.332	1	.127
N of Valid Cases	176		

a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 3.50.

Table 4.30: Chi-square test of *histcat*

4.2 Descriptive analysis of vitamin therapies

From the Tables 4.31 to 4.36 we see that in the treated group 72.7% of the patients were prescribed 1.5g niacin, 47.7% of the patients were prescribed 25,000iu beta-carotene, 40.9% of the patients were prescribed 200mcg selenium, 15.9% of the patients were prescribed 400mcg selenium, 30.7% of the patients were prescribed 600mcg selenium, 77.3% of the patients were prescribed 12g vitamin C, 67% of the patients were prescribed 50mg zinc and 42% of the patients were prescribed coenzyme Q10, all on a daily basis.

doses of B-3 (g/day)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.0	4	4.5	4.5	4.5
	.3	7	8.0	8.0	12.5
	.5	1	1.1	1.1	13.6
	1.0	9	10.2	10.2	23.9
	1.5	64	72.7	72.7	96.6
	3.0	3	3.4	3.4	100.0
	Total	88	100.0	100.0	

Table 4.31: Doses of B-3 (g/day)

doses of carot (iu/day)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	21	11.9	23.9	23.9
	20000	2	1.1	2.3	26.1
	25000	42	23.9	47.7	73.9
	30000	4	2.3	4.5	78.4
	50000	7	4.0	8.0	86.4
	75000	7	4.0	8.0	94.3
	100000	4	2.3	4.5	98.9
	250000	1	.6	1.1	100.0
	Total	88	50.0	100.0	
Missing	System	88	50.0		
Total		176	100.0		

Table 4.32: Doses of beta-carotene (iu/day)

doses of selenium (mcg/day)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	2.3	2.3	2.3
	200	36	40.9	41.4	43.7
	300	2	2.3	2.3	46.0
	400	14	15.9	16.1	62.1
	500	1	1.1	1.1	63.2
	600	27	30.7	31.0	94.3
	800	1	1.1	1.1	95.4
	1000	4	4.5	4.6	100.0
	Total	87	98.9	100.0	
Missing	System	1	1.1		
Total		88	100.0		

Table 4.33: Doses of selenium (mcg/day)

doses of vitamin C (g/day)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3.0	6	6.8	6.8	6.8
	4.0	1	1.1	1.1	8.0
	5.0	1	1.1	1.1	9.1
	6.0	7	8.0	8.0	17.0
	6.5	1	1.1	1.1	18.2
	8.0	1	1.1	1.1	19.3
	9.0	1	1.1	1.1	20.5
	10.0	1	1.1	1.1	21.6
	12.0	68	77.3	77.3	98.9
	24.0	1	1.1	1.1	100.0
	Total	88	100.0	100.0	

Table 4.34: Doses of vitamin C (g/day)

doses of zinc (mg/day)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	15	17.0	17.4	17.4
	50	67	76.1	77.9	95.3
	60	1	1.1	1.2	96.5
	100	3	3.4	3.5	100.0
	Total	86	97.7	100.0	
Missing	System	2	2.3		
Total		88	100.0		

Table 4.35: Doses of zinc (mg/day)

coenzyme Q10

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	51	58.0	58.0	58.0
	Yes	37	42.0	42.0	100.0
	Total	88	100.0	100.0	

Table 4.36: Frequency table of coenzyme Q10

4.3 Graphically compare the survival or censoring time of the treated and control groups

To draw inferences about the distributions of the right-censored survival data, we graph the empirical survival functions for the treated and control groups separately. A standard estimator of the survival function, proposed by Kaplan and Meier (1958), is called the Product-Limit estimator.

Suppose that events occur at D distinct times $a_1 < a_2 < \cdots < a_D$. Let h_j be the discrete hazard function at time a_j, which is defined as the conditional probability of

failure at a_j,

$$h_j = P(T = a_j | T \geq a_j),$$

where T denotes the discrete random variable representing the failure time. Then the survivor function at time t can be expressed in terms of h_j as

$$s(t) = \prod_{a_j < t} (1 - h_j). \tag{4.1}$$

A nonparametric estimator of the survivor function is

$$\widehat{s}(t) = \prod_{a_j < t} (1 - \widehat{h}_j), \tag{4.2}$$

where the $\widehat{h}_j$ are the maximum likelihood estimators of the h_j. From the log likelihood in terms of the h_j,

$$l = \sum_j \{d_j \log(h_j) + (r_j - d_j) \log(1 - h_j)\},$$

where r_j is the number of individuals at risk at a_j (including any individuals who are censored at a_j), and d_j is the number of individuals who fail at a_j, we have

$$\frac{\partial l}{\partial h_j} = \frac{d_j}{h_j} - \frac{r_j - d_j}{1 - h_j} = 0, \text{ and } \widehat{h}_j = d_j / r_j.$$

Then the corresponding estimator $\widehat{s}(t)$ of the survivor function is

$$\widehat{s}(t) = \prod_{a_j < t} (1 - \frac{d_j}{r_j}). \tag{4.3}$$

$\widehat{s}(t)$ is called the Kaplan-Meier estimator.

Figure 4.1 shows the plots of the Kaplan-Meier survival functions. The curve of estimated survival function for the control group, $\widehat{s_0}(t)$, lies above the curve of estimated survival function for the treated group, $\widehat{s_1}(t)$, in the plot, suggesting the superiority of control over treated in prolonging survival. Note however that this figure does not incorporate

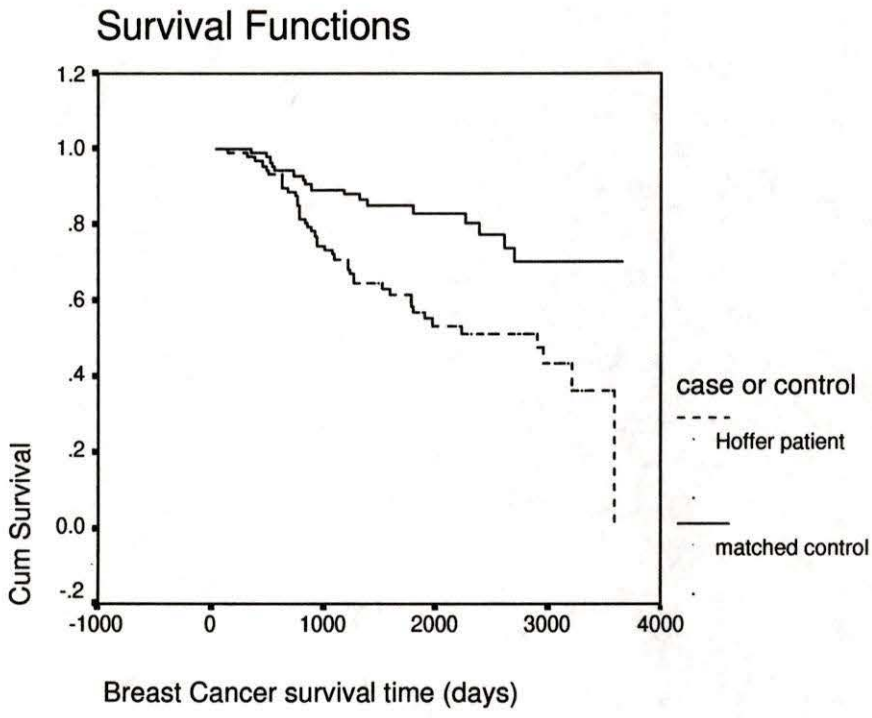

Figure 4.1: Plots of the K-M survival functions by treated and control groups

the time dependent nature of the treatment and is therefore biased. This is addressed in the model specification in subsequent chapters. Table 4.37 lists the status of the observations by the treated and control groups separately.

case or control * Breast Cancer Specific Death Crosstabulation

Count

		Breast Cancer Specific Death			
		Lost To Follow	No death	Death	Total
case or control	matched control	6	65	17	88
	Hoffer patient	1	45	42	88
Total		7	110	59	176

Table 4.37: The status of the observations by treated and control

Chapter 5

Cox proportional hazards model with time-dependent covariate

5.1 Cox proportional hazards model with time-dependent covariate

The Cox proportional hazards model (Cox, 1972) is widely used in analyzing the effect on survival time of explanatory variables (called covariates) by assuming their effect is to multiply the hazard by a constant factor. In most cases the covariates will be constant for each individual, but time-dependent covariates can arise, for example, if the treatment under study is not applied until some time after the time origin. Then a suitable covariate may be a time-dependent binary variable that jumps from 0 to 1 at the point of application of the treatment.

In the treated group of the breast cancer data, most of the patients did not start

vitamin treatment until some time after the diagnosis date. A time-dependent covariate is defined as

$$z_{i0}(t) = \begin{cases} 0 & \text{if patient } i \text{ has not yet started the vitamin treatment at time } t \\ 1 & \text{if patient } i \text{ has started the vitamin treatment at time } t \end{cases}, \quad (5.1)$$

and $z_i'(t) = (z_{i0}(t), z_{i1}, z_{i2}, ..., z_{ip-1})$ is the vector of covariates for the ith patient at time t. The proportional hazards model is

$$h_i(t) = e^{\beta' z_i(t)} h_0(t), \quad (5.2)$$

where $h_0(t)$ is an arbitrary base line hazard function, β is a p-vector of unknown regression coefficients and $e^{\beta z_i'(t)}$ incorporates the covariate effects for the ith individual. The parameters can be estimated by $\widehat{\beta}$ which maximizes the partial likelihood function (Cox, 1972 and 1975),

$$L(\beta) = \Pi_{i=1}^{n} \{ \frac{e^{\beta' z_i(T_i)}}{\Sigma_{j \in \mathcal{R}_i} e^{\beta' z_j(T_i)}} \}^{\delta_i}, \quad (5.3)$$

where $T_1, ..., T_n$ are the n possibly right censored survival times, $\mathcal{R}_i = \{j : T_j \geq T_i\}$ and δ_i is 0 if the ith patient is censored and 1 if she has been observed to fail. $\widehat{\beta}$ is asymptotically normally distributed and is a consistent estimator of β (Andersen and Gill, 1982).

For the pair-matched survival data there may be a distinct base line hazard function for each pair, $h_{i0}(t)$ for the ith pair, then the proportional hazards model is

$$h_{ij}(t) = e^{\beta' z_{ij}(t)} h_{i0}(t), \quad (j = 1, 2).$$

5.2 Proportional hazards tests and martingale residuals

5.2.1 A regression approach to test for nonproportionality

Grambsch and Therneau (1994) give a method using a regression approach to test the nonproportionality of the covariates in the proportional hazards model. Consider each subject to be an independent counting process $\{N_i(t), t \geq 0, i = 1, ..., n\}$ the proportional hazards model (5.1) can be formulated as

$$Y_i(t) \exp\{\beta' Z_i(t)\} h_0(t), \tag{5.4}$$

where $Y_i(t)$ is a $0-1$ process which indicates whether the ith subject is at risk at time t, β is a p vector of regression parameters, $Z_i(t)$ is a p vector of covariate processes at time t and $h_0(t)$ is an unspecified hazard function. The conditional weighted mean and variance of the covariate vector at time t given the process history up to time t are

$$\begin{aligned} M(\beta, t) &= S^{(1)}(\beta, t)/S^{(0)}(\beta, t), \\ V(\beta, t) &= \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \{\frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}\}^{\otimes 2}, \end{aligned}$$

where

$$S^{(r)}(\beta, t) = \Sigma_{i=1}^{n} Y_i(t) \exp\{\beta' Z_i(t)\} Z_i(t)^{\otimes r}$$

with $Z_i(t)^{\otimes 2}$ denotes the outer product $Z_i(t)Z_i(t)'$, $Z_i(t)^{\otimes 1}$ denotes the vector $Z_i(t)$, and $Z_i(t)^{\otimes 0}$ is the scalar 1.

An alternative to the proportional hazards model is a time-varying coefficients model with intensity given by

$$h_i(t) = Y_i(t) \exp\{\beta'(t) Z_i(t)\} h_0(t)$$

$$= Y_i(t)\exp[\{\beta + G(t)\theta\}' Z_i(t)]h_0(t), \tag{5.5}$$

where β in the proportional hazards model is replaced by $\beta + G(t)\theta$ to incorporate time-varying hazards, $G(t)$ is a $p \times p$ diagonal matrix of predictable processes $G_{jj}(t)$, and θ is a vector of parameters. If there is no evidence against the null hypothesis, $H_0 : \theta = 0$, then the hazard function (5.5) turns out to be the proportional hazards model. The Schoenfeld residuals are defined as

$$\begin{aligned} r_k(\beta) &= Z(t_k) - M(\beta, t_k) \\ &= [Z(t_k) - M\{(\beta(t_k), t_k\}] + [M\{\beta(t_k), t_k\} - M(\beta, t_k)]. \end{aligned}$$

Suppose β is known, expand $M\{(\beta(t_k), t_k\}$ in a Taylor's expansion about $\theta = 0$ in the second summand

$$\begin{aligned} M\{(\beta(t_k), t_k\} &= \frac{S^{(1)}(\beta + G(t_k)\theta, t_k)}{S^{(0)}(\beta + G(t_k)\theta, t_k)} \\ &\doteq \frac{S^{(1)}(\beta, t_k)}{S^{(0)}(\beta, t_k)} + \frac{S^{(2)}(\beta, t_k)S^{(0)}(\beta, t_k) - (S^{(1)}(\beta, t_k))^{\otimes 2}}{(S^{(0)}(\beta, t_k))^{\otimes 2}} G(t_k)\theta \\ &= M(\beta, t_k) + V(\beta, t_k)G(t_k)\theta, \end{aligned}$$

so,

$$E\{r_k(\beta)| \text{ the process history up to time } t_k\} \doteq V(\beta, t_k)G(t_k)\theta.$$

Let $r_k^*(\beta) = V^{-1}(\beta, t_k)r_k(\beta)$, then

$$E\{r_k^*(\beta)| \text{ the process history up to time } t_k\} \doteq G(t_k)\theta,$$

$$\text{and } Var\{r_k^*(\beta)| \text{ the process history up to time } t_k\} \doteq V^{-1}(\beta, t_k)$$

suggesting a standard linear model for $r_k^*(\beta)$. With $V_k \equiv V(\beta, t_k)$, the generalized least

squares equations gives

$$\widehat{\theta} = \{\Sigma G(t_k) V_k G(t_k)\}^{-1} \{\Sigma G(t_k) r_k(\beta)\}$$

which leads to an asymptotic χ^2 test statistic on p degrees of freedom:

$$\{\Sigma G(t_k) r_k(\beta)\}' \{\Sigma G(t_k) V_k G(t_k)\}^{-1} \{\Sigma G(t_k) r_k(\beta)\} \tag{5.6}$$

to test $H_0 : \theta = 0$.

When β is unknown and $\widehat{\beta}$ is the maximum partial likelihood estimate under H_0, let $\widehat{V}_k = V(\widehat{\beta}, t_k)$ and $\widehat{r}_k = r_k(\widehat{\beta})$. Then the asymptotic χ^2 test statistic is modified as

$$\{\Sigma G(t_k) \widehat{r}_k\}' D^{-1} \{\Sigma G(t_k) \widehat{r}_k\}, \tag{5.7}$$

with

$$D = \Sigma G(t_k) \widehat{V}_k G(t_k) - \{\Sigma G(t_k) \widehat{V}_k\} (\Sigma \widehat{V}_k)^{-1} \{\Sigma G(t_k) \widehat{V}_k\}',$$

where $\{\Sigma G(t_k) \widehat{V}_k\} (\Sigma \widehat{V}_k)^{-1} \{\Sigma G(t_k) \widehat{V}_k\}'$ is the consistent estimator of the covariance matrix of $(\widehat{r}_1, \widehat{r}_2, ..., \widehat{r}_d)'$ since, $\Sigma \widehat{r}_k = 0$, the residuals are correlated.

5.2.2 The martingale residuals

The martingale residual is defined as

$$\widehat{M}_i = \delta_i - \int_0^\infty Y_i(s) e^{\widehat{\beta}' z_i(s)} d\widehat{\Lambda}_0(s) \quad (i = 1, ..., n), \tag{5.8}$$

where $\widehat{\Lambda}_0(t)$ is the estimated cumulative base line hazard (Breslow, 1974) given by

$$\widehat{\Lambda}_0(t) = \Sigma_{T_i \le t} \frac{\delta_i}{\Sigma_{j \in \mathcal{R}_i} e^{\widehat{\beta}' z_j(T_i)}}.$$

The martingale residuals can be interpreted as the observed number of events minus the expected number of events under the assumed Cox model. The martingale residuals have

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Breast Cancer survival time (days)	88	150	3593	1687.92	911.18
#days to hof	88	0	2896	425.03	618.86
Valid N (listwise)	88				

Table 5.1: Compare *dthcsurv* with *daytohof*

the property $\Sigma_{i=1}^{n}\widehat{M_i} = 0$. Also, for large samples the $\widehat{M_i}$ are an uncorrelated sample from a population with a zero mean. Martingale residual plots are useful for assessing the functional form of the covariates in the model. Plots of $\widehat{M_i}$ versus a covariate not in the model can indicate the functional form for the covariate. Note however, that the martingale residuals can have a highly skewed distribution, which excludes their use for identifying influential observations.

5.3 Fit the Cox proportional hazards model for the breast cancer data set I

In the breast cancer data set, except for five patients who started the vitamin treatment immediately after diagnosis, all of the other patients in the treated group were under conventional treatment for a period of time before they underwent the additional vitamin treatment. Therefore the treatment covariate is time-dependent. See Table 5.1 for the comparison of the survival time and *daytohof*, the number of days from diagnosis date to the start date of the vitamin treatment.

Variable	No. of missing values	Used (Yes or No)
sizles	12	No
staget	1	Yes
posnod	13	No
dxposnod	0	Yes
dxer	31	No
dxgrade	21	Yes
dxlvn	13	Yes
dxyear	0	Yes
agedx	0	Yes
histcat	0	Yes

Table 5.2: Selection of the variables

The Cox proportional hazards model with time-dependent covariates is fitted for the pair-matched breast cancer survival data in the following steps:

5.3.1 Model 1. The full model

Table 5.2 shows the selection of the variables. Note that *sizles*, *posnod* and *dxer* have many missing values. We used *staget* and *dxposnod* which contain some of the information of *sizles* and *posnod* instead. Tables 5.3 to 5.6 give the results of tests the homogeneity of the distributions of the missing values between the treated and control groups for *dxgrade* and *dxlvn* separately. There are significantly more missing values in the treated group than that in the control group for variable *dxgrade*. The variables selected in Table 5.2 are called group I variables, and all the other variables are called the group II variables.

First the Cox regression model with time dependent treatment effect and the group I variables as covariates was fitted. There are 176 total cases read; 26 cases with missing values and one censored case before the earliest event in a stratum were dropped. At

case or control * dxgrade missing Crosstabulation

			dxgrade missing		
			No	Yes	Total
case or control	matched control	Count	84	4	88
		Expected Count	77.5	10.5	88.0
	Hoffer patient	Count	71	17	88
		Expected Count	77.5	10.5	88.0
Total		Count	155	21	176
		Expected Count	155.0	21.0	176.0

Table 5.3: Count of the missing values of *dxgrade* for treated and control separately

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.138[b]	1	.003		
Continuity Correction[a]	7.786	1	.005		
Likelihood Ratio	9.753	1	.002		
Fisher's Exact Test				.004	.002
Linear-by-Linear Association	9.086	1	.003		
N of Valid Cases	176				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.50.

Table 5.4: Test homogeneity of the missing values for *dxgrade* between treated and control

case or control * Invasive lvn Crosstabulation

			dxlvn missing		
			No	Yes	Total
case or control	matched control	Count	84	4	88
		Expected Count	81.5	6.5	88.0
	Hoffer patient	Count	79	9	88
		Expected Count	81.5	6.5	88.0
Total		Count	163	13	176
		Expected Count	163.0	13.0	176.0

Table 5.5: Count of the missing values of *dxlvn* for treated and control separately

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.076[b]	1	.150		
Continuity Correction[a]	1.329	1	.249		
Likelihood Ratio	2.127	1	.145		
Fisher's Exact Test				.248	.124
Linear-by-Linear Association	2.065	1	.151		
N of Valid Cases	176				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.50.

Table 5.6: Test homogeneity of the missing values for *dxlvn* between treated and control

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.5613	.3415	20.9076	1	.0000	.2092	4.7651
agedx	.0232	.0128	3.2629	1	.0709	.0541	1.0235
dxyear	.1592	.0903	3.1130	1	.0777	.0508	1.1726
dxlvn	-.7064	.3797	3.4607	1	.0628	-.0582	.4934
staget			5.0671	2	.0794	.0504	
staget(1)	-1.2105	.5880	4.2382	1	.0395	-.0720	.2980
staget(2)	-.5957	.5207	1.3086	1	.2526	.0000	.5512
dxgrade			5.5470	2	0624	.0598	
dxgrade(1)	-1.0123	.5523	3.3596	1	.0668	-.0561	.3634
dxgrade(2)	-.7851	.4149	3.5813	1	.0584	-.0605	.4561
dxposnod			.3306	2	.8476	.0000	
dxposnod(1)	-.6011	1.0504	.3275	1	.5671	.0000	.5482
dxposnod(2)	-.0540	.3837	.0198	1	.8880	.0000	.9474
histcat			.3483	2	.8402	.0000	
histcat(1)	.4696	1.0310	.2074	1	.6488	.0000	1.5993
histcat(2)	.6281	1.0917	.3310	1	.5651	.0000	1.8740

Table 5.7: Parameter estimates for model (1)

staget = 4 there is only one case. If this case was kept in the data, a parameter would be estimated to compare this particular case with other levels of the *staget*; the estimate would have a large standard error which is not reliable. After this case was dropped, we have 148 cases available for the analysis and among them 99 (66.9%) cases are censored. The proportional hazards assumption was tested for each covariate. The result shows that only *agedx* is not proportional. Tables 5.7 and 5.8 show the parameter estimate and the proportional hazards tests of the full model. The case processing summary and the indicator parameter coding are displayed in Table 5.9. Table 5.10 is the tabulation of the dropped cases.

Variable	rho	chisq	P-value
Treatment	.0778	.35063	.5538
agedx	.2373	2.80645	.0939
dxyear	.0213	.01976	.8882
dxlvn	-.0828	.33926	.5603
staget(1)	-.0987	.55163	.4577
staget(2)	-.1087	.58790	.4432
dxgrade(1)	.2396	2.14659	.1429
dxgrade(2)	-.0107	.00657	.9354
dxposnod(1)	-.0098	.00485	.9445
dxposnod(2)	.0981	.47943	.4887
histcat(1)	-.0158	.01157	.9143
histcat(2)	.0491	.11531	.7342
Global	NA	11.17003	.5144

Table 5.8: Tests of the proportional hazards assumption for model (1)

Value	Freq(Treated/Control)	(1)	(2)
staget **(T-stage)**			
<2.01cm	39/47	1	0
>2cm and<5.01cm	25/31	0	1
>5cm	3/3	0	0
dxposnod **(Nodal status)**			
No axillary dissection	5/0	1	0
No pos nodes	28/35	0	1
Pos nodes	34/46	0	0
dxgrade **(Tumor grade)**			
Well differentiated	12/8	1	0
Moderately differentiated	21/32	0	1
Poorly differentiated	34/41	0	0
dxlvn **(Invasive lvn)**			
Negative	36/38	1	0
Positive	31/43	0	0
histcat **(Categorical histology)**			
Ductal	61/62	1	0
Lobular	4/16	0	1
Other	2/3	0	0

Table 5.9: Case processing summary and the indicator parameter coding

	Lost to follow	**No death**	**Death**	**Total**
Treated	0	12	9	21
Control	2	4	1	7
Total	2	16	10	28

Table 5.10: The status of the dropped cases

To see why the *agedx* effect on the hazard function is not proportional, we categorize *agedx* into a new variable, *catage*, with the following five groups:

catage	*agedx*
1	$agedx \leq 40$
2	$40 < agedx \leq 45$
3	$45 < agedx \leq 50$
4	$50 < agedx \leq 55$
5	$55 < agedx$

.

Figures 5.1 and 5.2 are the plots of the Kaplan-Meier survival functions at each level of *catage* for the treated and control groups separately, since *treatment* is the only very significant factor in model (1). The curves cross each other in each of the two plots which gives evidence against the assumption of the proportional hazards effect of *agedx*.

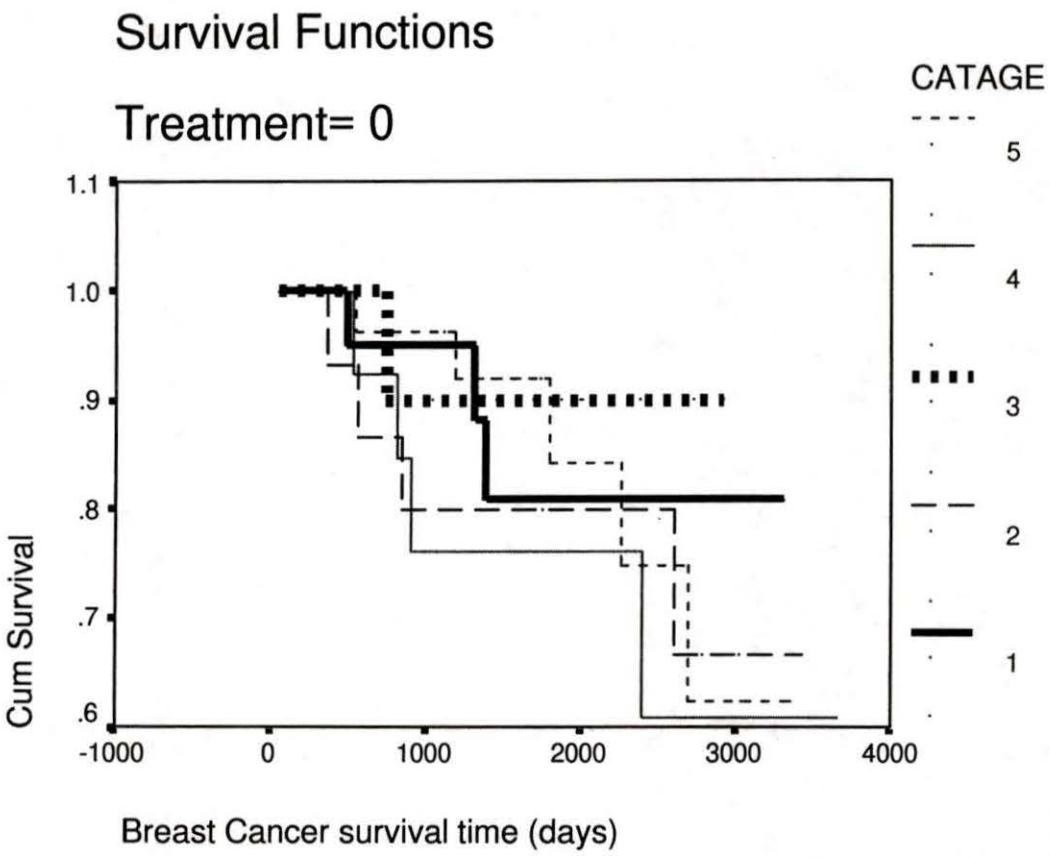

Figure 5.1: K-M survival functions for control group

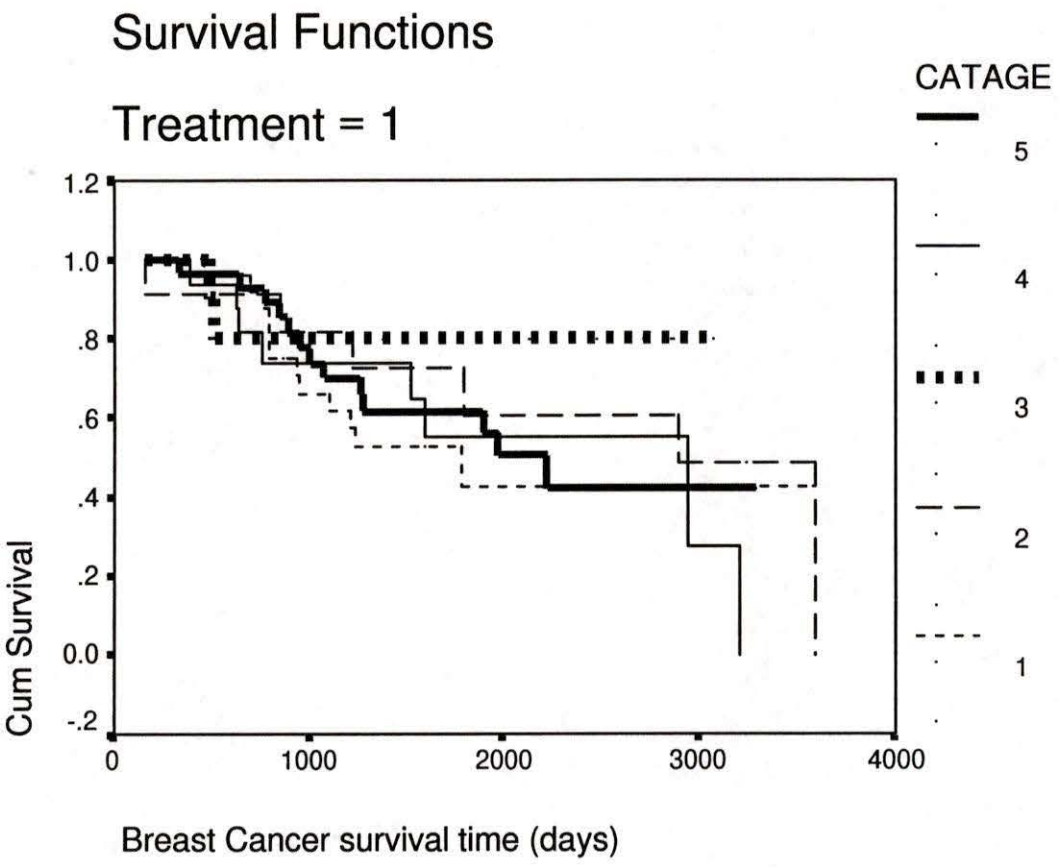

Figure 5.2: K-M survival funtions for treated group

5.3.2 Model 2. The full model using *catage* as strata

The Cox regression model with *catage* as strata, the treatment effect and the group I variables except for *agedx* as covariates was fitted. The proportional hazards assumption for each covariate was tested again. There is no evidence against the proportional hazards assumption since all the P-values are larger than 0.23. The estimated treatment effect is $\beta = 1.7524$ with P-value $= 0.0000$, which indicates in the treated group a patient's hazard of death is $\exp\{1.6804\} = 5.3676$ times that for a patient in the control group, all other

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.7524	.3706	22.3609	1	.0000	.2660	5.7683
dxyear	.1442	.0973	2.1979	1	.1382	.0262	1.1551
dxlvn	-.5652	.4211	1.8015	1	.1795	.0000	.5683
staget			6.4694	2	0394	.0926	
staget(1)	-1.3504	.6698	4.0651	1	.0438	-.0847	.2591
staget(2)	-.5198	.6001	.7502	1	.3864	.0000	.5946
dxgrade			8.1589	2	.0169	.1202	
dxgrade(1)	-1.6124	.7026	5.2673	1	.0217	-.1066	.1994
dxgrade(2)	-.9192	.4344	4.4779	1	.0343	-.0928	.3988
dxposnod			.8850	2	.6424	.0000	
dxposnod(1)	-1.0410	1.1405	.8331	1	.3614	.0000	.3531
dxposnod(2)	-.1745	.4323	.1630	1	.6864	.0000	.8399
histcat			.6013	2	.7404	.0000	
histcat(1)	.1500	1.0817	.0192	1	.8897	.0000	1.1618
histcat(2)	.5432	1.1504	.2230	1	.6368	.0000	1.7215

Table 5.11: Parameter estimates for Model (2)

variables fixed. We see that the estimated hazard contribution for *histcat* at the ductal level, $\exp\{\beta_d\} = 1.1618$, is lower than that for the lobular level, $\exp\{\beta_l\} = 1.7215$, which is not reasonable even though their effects are not significant. This may partly due to the inaccuracies in the variable *histcat*, see Appendix A for more information. Tables 5.11 and 5.12 are the parameter estimates and the proportional hazards tests for this model. Figure 5.3 is the plot of the martingale residuals versus index with the control patients first, the hazard of the patients in the treated group may be under estimated in this model since the smoothed line tends above zero for the treated patients. See Tables 5.9 and 5.13 for the case processing summary and the indicator parameter coding.

Variable	rho	chisq	P-value
Treatment	.0763	.34458	.557
dxyear	.1331	.63076	.427
dxlvn	-0.0142	.00941	.923
staget(1)	-.1829	1.39031	.238
staget(2)	-.1266	.59507	.440
dxgrade(1)	0.0734	.19224	.661
dxgrade(2)	0.0318	.05618	.813
dxposnod(1)	.0493	.09150	.762
dxposnod(2)	.1419	.92244	.337
histcat(1)	-.1248	.73676	.391
histcat(2)	-.0217	.02069	.886
Global	NA	5.58377	.900

Table 5.12: Tests of the proportional hazards assumption for Model (2)

catage	Events	Censored	Percent censored
1	14	26	56%
2	10	12	54.5%
3	2	15	88.2%
4	9	16	64%
5	14	30	68.2%
Total	49	99	66.9%

Table 5.13: Case processing summary by *catage*

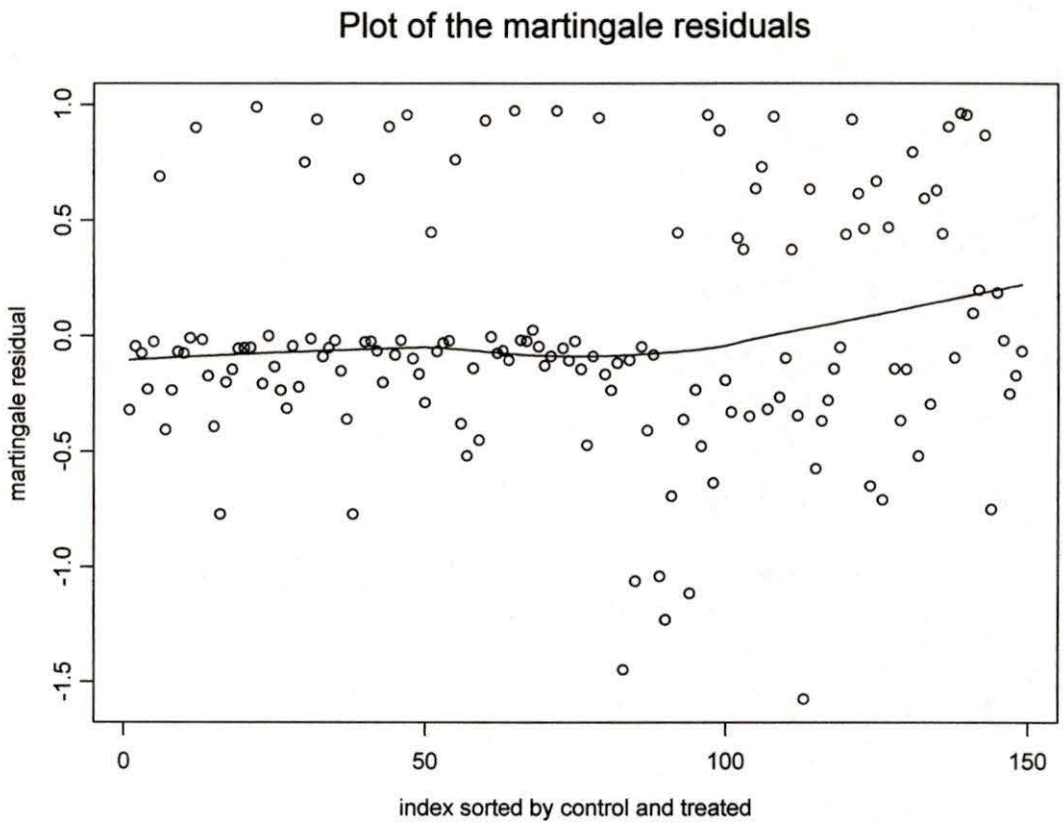

Figure 5.3: The plot of the martingale residuals for Model (2)

5.3.3 Model 3. The reduced model

From the output of model (2) we see that some covariates are not statistically significant. Using the backward elimination method, we arrive at the final model with covariates *treatment*, *dxlvn*, *staget* and *dxgrade*. From the parameter estimates in Table 5.14, we see first that the estimated treatment effect is still positive and a little smaller than that of model (2) which indicates the patients in the control group have a larger probability of living longer than the patients in the treated group after considering the

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.7325	.3445	25.2931	1	.0000	.2845	5.6548
dxlvn	-.6515	.3547	3.3739	1	.0662	-.0691	.5213
staget			7.3007	2	.0260	.1071	
staget(1)	-1.4686	.6592	4.9628	1	.0259	-.1015	.2302
staget(2)	-.6381	.5959	1.1467	1	.2842	.0000	.5283
dxgrade			8.7097	2	.0128	.1279	
dxgrade(1)	-1.4753	.6758	4.7652	1	.0290	-.0980	.2287
dxgrade(2)	-1.0223	.4169	6.0128	1	.0142	-.1181	.3598

Table 5.14: Parameter estimates for Model (3)

effects of significant factors *dxlvn*, *staget* and *dxgrade*; second that for *dxgrade*, "well or moderately differentiated" is significantly better than "poorly differentiated" with P-values of 0.0290 and 0.0142 respectively, "well differentiated" is the best case; third that for *staget*, tumor size "not larger than 2 cm" is significantly better than "tumor with direct extension" with P-value of 0.0259, and tumor size "between 2 cm and 5 cm" is also better than "tumor with direct extension" but not significant; fourth that for *dxlvn*, "negative" is better than "positive" with P-value of 0.0662.

Figures 5.4 to 5.7 are the plots which examine the influence of individual observations on the parameter estimates. The scaled change in the estimated coefficients due to dropping each observation from the fit is plotted versus index for each of the covariates in the model. The plots are reasonable for all of the observations. The plot of the martingale residuals in Figure 5.8 is quite like the plot for model (2). See Tables 5.9 and 5.13 for the case processing summary and the indicator parameter coding, Table 5.15 for the tests of the proportional hazards assumption.

Variable	rho	chisq	P-value
Treatment	.0460	.1221	.727
dxlvn	.0950	.3820	.537
staget(1)	-.1194	.6537	.419
staget(2)	-.0636	.1742	.676
dxgrade(1)	.1268	.6153	.433
dxgrade(2)	-.0258	.0362	.849
Global	NA	2.4206	.877

Table 5.15: Tests of the proportional hazards assumption for Model (3)

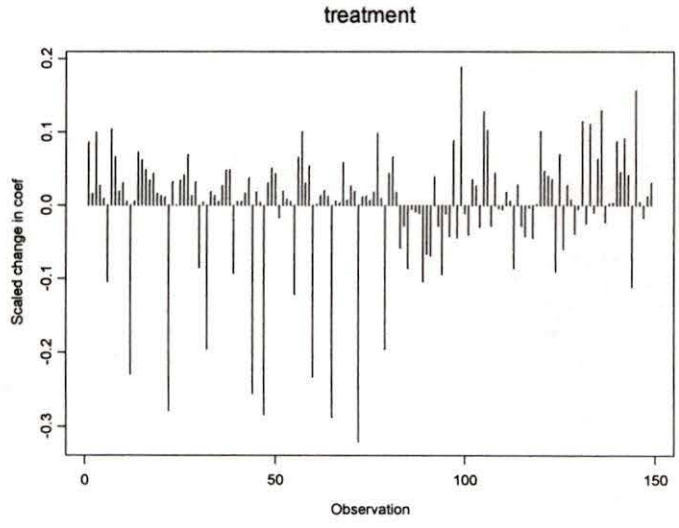

Figure 5.4: Plot of influence by observation number for *treatment* in model (3)

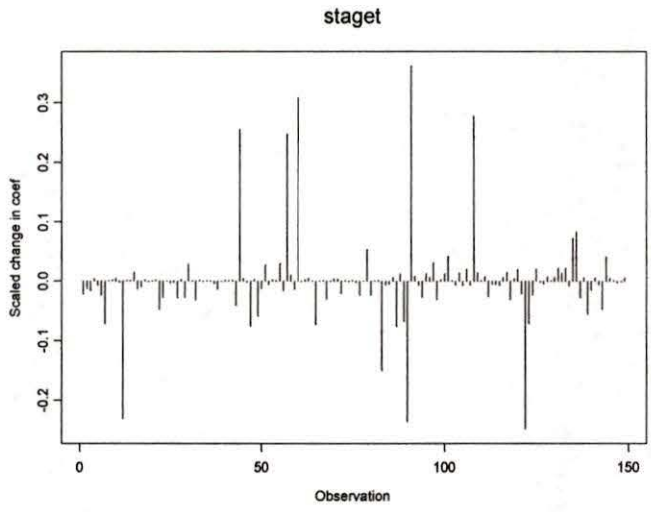

Figure 5.5: Plot of influence by observation number for *staget* in model (3)

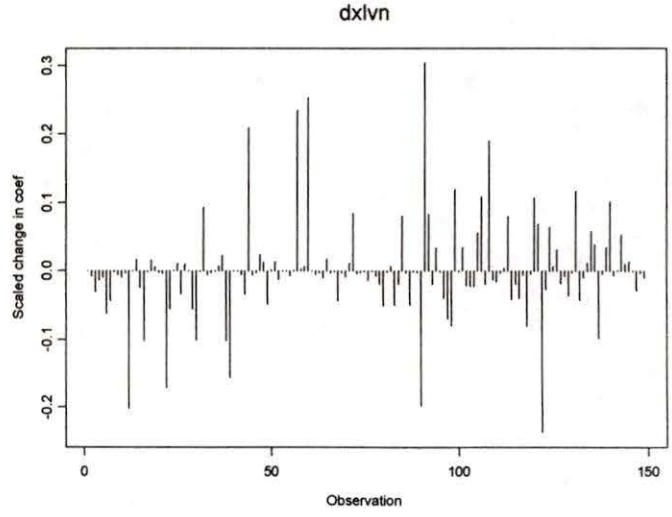

Figure 5.6: Plot of influence by observation number for *dxlvn* in model (3)

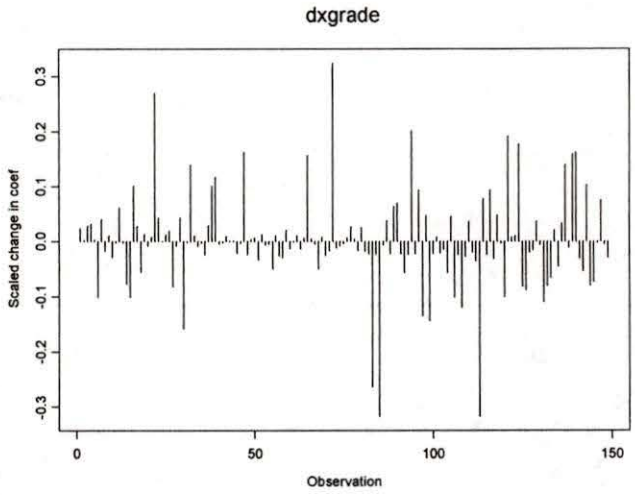

Figure 5.7: Plot of influence by observation number for *dxgrade* in model (3)

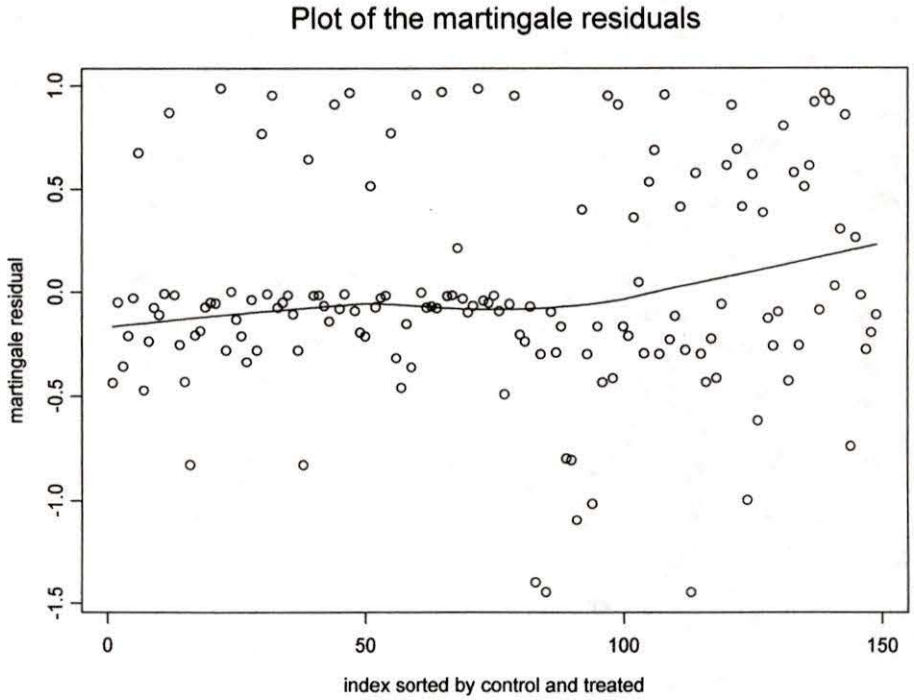

Figure 5.8: The plot of the martingale residuals for Model (3)

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.7949	.6288	8.1467	1	.0043	.3722	6.0188
agedx	-.4195	.4693	.7990	1	.3714.	.0000	.6574
dxyear	.3786	1.1478	.1088	1	.7415	.0000	1.4574
dxlvn	.7708	1.3839	.3102	1	.5776	.0000	2.1614
dxgrade			2.3943	2	.3021	.0000	
dxgrade(1)	-2.5215	2.4776	1.0357	1	.3088	.0000	.0803
dxgrade(2)	-.0409	2.4295	.0003	1	.9866	.0000	.9599

Table 5.16: Parameter estimates for Model (4)

5.3.4 Model 4. The full model using *pair* as strata

The data in Data set I is pair-matched. We incorporate the pair matching into the Cox model by stratifying on *pair* number. This should in a simplified Cox regression model. For example *staget* and *dxposnod* are matched exactly for all the pairs that are used in the model, therefore they need not appear in the stratified model since their effects are absorbed into the base-line hazard function for each pair.

Model 4 was fitted with *pair* as strata, the treatment effect and *agedx*, *dxyear*, *dxlvn* and *dxgrade* as covariates. The variable *histcat* is not used in this model. There are 176 total cases read, 26 cases with missing values; 75 censored cases before the earliest event in a stratum were dropped; 75 cases, which includes 32 complete pairs, are available for the analysis with 25 (33.3%) cases censored. Table 5.16 contains the parameter estimates for this model. Table 5.17 contains the tests of the proportional hazards assumption. We see that there is no evidence against the proportional hazards assumption and the estimated treatment effect is quite like that of the model (3). Figure 5.9 shows that the martingale residuals for most of the observations are very close to zero which indicates model (4) fits well for most of the observations.

Variable	rho	chisq	P-value
Treatment	.1292	.7969	.372
agedx	-.0065	.0020	.965
dxyear	.1428	.3217	.571
dxlvn	.0487	.0667	.796
dxgrade(1)	-.1046	.2134	.644
dxgrade(2)	-.0061	.0005	.983
Global	NA	1.1971	.977

Table 5.17: Tests of the proportional hazards assumption for Model (4)

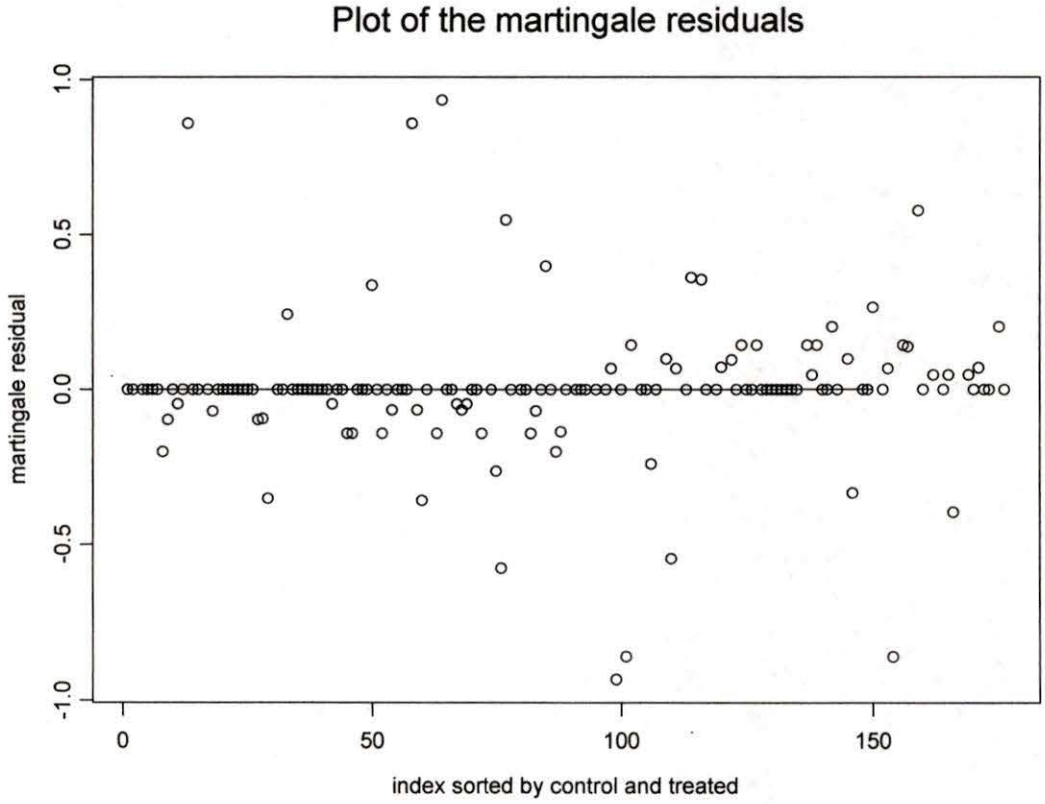

Figure 5.9: The plot of the martingale residuals for Model (4)

5.3.5 Model 5. Fit the reduced model stratifying by *pair* using all the possible observations

We see that in model (4) most of the covariates are not significant. We refit the model using the backward regression method, the reduced model has only treatment as covariate, all the other variables are removed because they are not significantly different from zero. There are 176 total cases read, 87 censored cases before the earliest event in a stratum were dropped, 89 cases available for the analysis and 30 (33.7%) cases censored. Among the 89 cases there are 42 complete pairs and 5 more unpaired treated patients. Tables 5.18 and 5.19 contain the parameter estimates and tests of the proportional hazards assumption for the model. There is no evidence against the proportional hazards assumption, and the estimated treatment effect is smaller than that of the above three models, but still very significant.

Figure 5.10 is a plot of the scaled change in the treatment parameter estimates versus observation number. No influential points are observed. Figure 5.11 is the plot of martingale residuals versus index with the control patients first. Only a few observations have residuals between 1 to 0.5 distance away from zero, all the other residuals are very close to zero. This indicates the excess number of events seen in the data but not predicted by this proportional hazard model are close to zero. Most of the treated patients have a small positive martingale residual, which indicates the treatment is slightly underestimated in this Cox model.

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.3863	.3953	12.2996	1	.0005	.4206	4.000

Table 5.18: Parameter estimates for Model (5)

Variable	rho	chisq	P-value
Treatment	.0037	.0008	.977

Table 5.19: Tests of the proportional hazards assumption for Model (5)

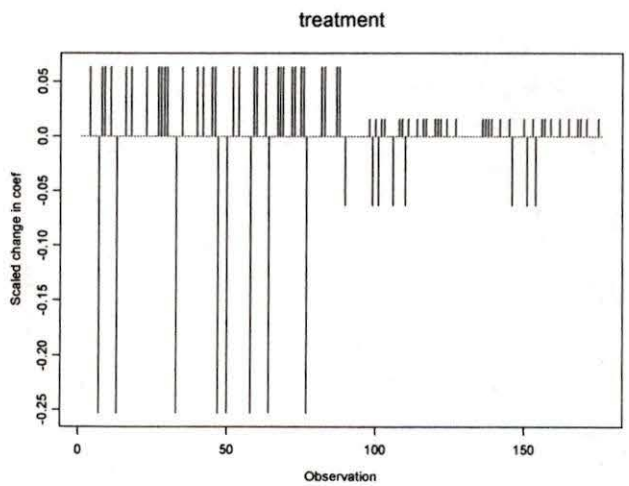

Figure 5.10: Plot of influence by observation number for *treatment* in model (5)

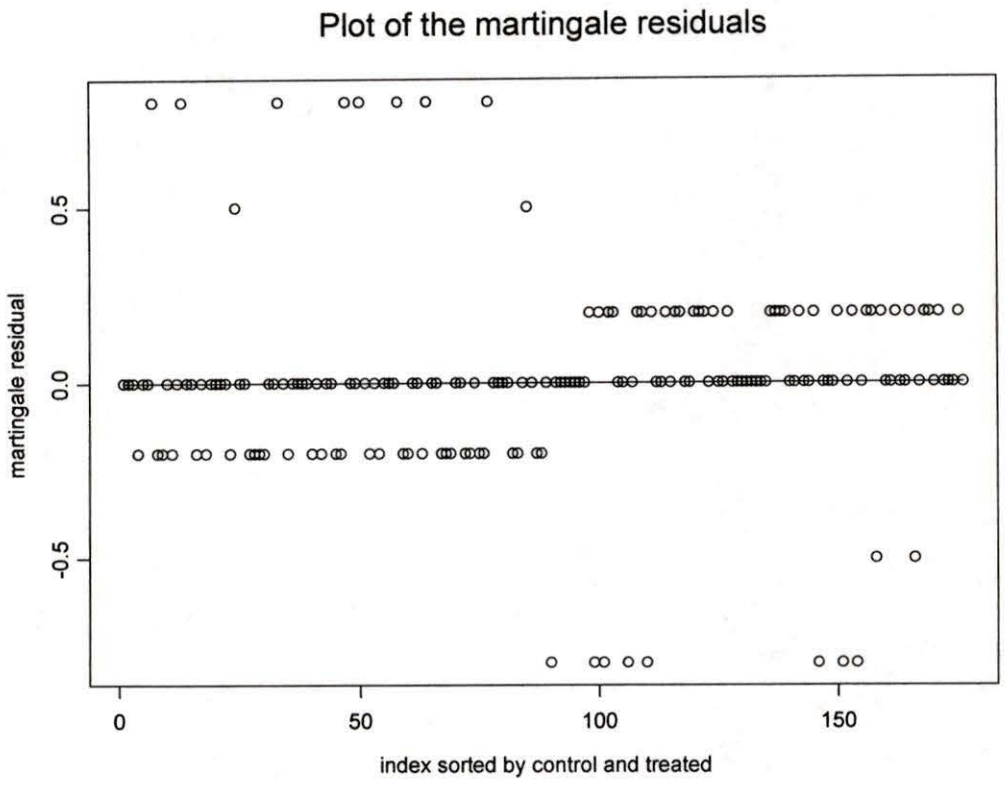

Figure 5.11: The plot of martingale residuals for Model (5)

Possible order of event	Partial likelihood	Rank of the survival times
$H < T < C$	$e^{\beta}/(1+e^{\beta})$	treated shorter
$H < (T) < C$	1	no contribution
$H < T < (C)$	$e^{\beta}/(1+e^{\beta})$	treated shorter
$C < H < T$	$1/2$	no contribution
$(C) < H < T$	1	no contribution
$C < H < (T)$	$1/2$	no contribution
$H < C < T$	$1/(1+e^{\beta})$	treated longer
$H < (C) < T$	1	no contribution
$H < C < (T)$	$1/(1+e^{\beta})$	treated longer

Table 5.20: The possible ordering of events and their partial likelihood

In model (5) there is only one covariate, the treatment effect, so the partial likelihood function is simplified. For any particular pair we let H be the number of days from diagnosis date to start of vitamin treatment, T and C be the possibly right censored survival time for the treated and the control patients respectively. Table 5.20 gives all the possible ordering of the events and the corresponding partial likelihood of this pair with T or C in parentheses giving the censoring time.

Then the partial likelihood function for the Cox proportional hazards model with only the treatment effect simplifies to:

$$\begin{aligned} L(\beta) &= \Pi_{i=1}^{n}\left(\frac{1}{1+e^{\beta}}\right)^{x_i}\left(\frac{e^{\beta}}{1+e^{\beta}}\right)^{1-x_i} \\ &= \left(\frac{1}{1+e^{\beta}}\right)^{n} e^{\beta(n-\Sigma_i^n x_i)}, \end{aligned} \tag{5.9}$$

where n is the number of pairs which contribute to the partial likelihood for β as listed in Table 5.20, and

$$x_i = \begin{cases} 1 & \text{if the treated has a longer survival/censoring time} \\ 0 & \text{if the treated has a shorter survival/censoring time} \end{cases}.$$

Treatment	B	S.E.	Sig	Exp(B)	Lower .95	Upper .95
Model (2)	1.7524	.3706	.0000	5.7683	2.7900	11.9258
Model (3)	1.7325	.3445	.0000	5.6548	2.8786	11.108
Model (4)	1.7949	.6288	.0043	6.0188	1.7548	20.6436
Model (5)	1.3863	.3953	.0005	4.0000	1.8433	8.6801

Table 5.21: The estimates of the treatment effect from model (2) to model (5)

The parameter is estimated by

$$\widehat{\beta} = \log(\frac{n - \Sigma_i^n x_i}{\Sigma_i^n x_i})$$

which maximizes the logarithm of the partial likelihood function (5.9). The sum, $\Sigma_i^n x_i$ is the number of comparable pairs where the control patient dies before the treated patient, and $n - \Sigma_i^n x_i$ is the number of comparable pairs where the treated patient dies before the control patient.

5.3.6 Summary of the model fitting for Data set I

From the proportional hazards tests and the plots of the residuals we see that except for model (1) all the other five models are well fitted. To analyze the treatment effect the estimates of the treatment effect from model (2) to model (5) are summarized into the Table 5.21.

From the above five models, we see that treatment effect is highly significant with P-value $< .005$, and the hazard of death of the patient in the vitamin treated group is from four to six times that of the patient in the control group.

Chapter 6

Aligned rank test for pair-matched censored survival data

6.1 Test statistics

The aligned rank test for the null hypothesis of no treatment differences of Hodges and Lehmann (1962) was extended to pair-matched censored survival data by Schluchter (1985).

6.1.1 Notation and assumptions

Let X_{ir}^0 denote the survival days of the patient receiving treatment r in pair i,

$$r = \begin{cases} 1 & \text{if the patient in the treated group} \\ 0 & \text{if the patient in the control group} \end{cases},$$

and $i = 1, ..., n$. We actually observe $X_{ir} = \min(X_{ir}^0, C_{ir})$ and δ_{ir}, where C_{ir} is a censoring variate, and

$$\delta_{ir} = \begin{cases} 1 \text{ if } X_{ir} = X_{ir}^0 \\ 0 \text{ if } X_{ir} = C_{ir} \end{cases}$$

is an indicator variate. The survival days X_{ir}^0 are assumed to follow the model

$$X_{ir}^0 = \tau_i + \beta_r + e_{ir} \ ,$$

where e_{ir} are independent identically distributed random variables from a continuous distribution, τ_i are nuisance block effects, and β_r are treatment parameters. We assume that the censoring times in pair i, C_{i1}, C_{i0}, have an exchangeable distribution, and are independent of the outcomes and censoring times in other pairs.

6.1.2 Aligned rank test statistic for censored matched pairs

The true aligned observations in pair i are defined as $X_{i1}^0 - X_{i0}^0$ and $X_{i0}^0 - X_{i1}^0$. The observed aligned observations in pair i, $X_{i1} - X_{i0}$ and $X_{i0} - X_{i1}$, equal the true aligned observations if neither X_{i1}^0 nor X_{i0}^0 is censored. If one member of the pair, say X_{i1}, is censored then $(X_{i1} - X_{i0})$ is right censored and $(X_{i0} - X_{i1})$ is left censored. Under the model we assume, the fact that both pair members are censored could be due entirely to the presence of an arbitrarily large block effect. In that sense doubly censored pairs add no extra information about the treatment parameters and will be ignored in the analysis. In the generalized aligned rank test the rank assigned to $X_{i1}^0 - X_{i0}^0$, $R_i(1,0)$, is the number of other $2n-1$ aligned observations definitely smaller than that observation minus the number definitely larger. The statistic $W(1,0)$ is defined to be the sum of the ranks assigned to

treatment 1:

$$W(1,0) = R_1(1,0) + ... + R_n(1,0).$$

The scoring function $\psi(a,b)$ is defined as

$$\psi(a,b) = \begin{cases} 1 & \text{if } a > b \\ -1 & \text{if } a < b \\ 0 & \text{otherwise} \end{cases}, \tag{6.1}$$

then $R_i(1,0)$ is equal to

$$\psi(X_{i1}^0 - X_{i0}^0, X_{i0}^0 - X_{i1}^0) + \Sigma_{j\neq i}\psi(X_{i1}^0 - X_{i0}^0, X_{j0}^0 - X_{j1}^0) + \Sigma_{j\neq i}\psi(X_{i1}^0 - X_{i0}^0, X_{j1}^0 - X_{j0}^0). \tag{6.2}$$

Summing over i in (6.2), the third term on the right-hand side sums to zero, so that we have

$$W(1,0) = \Sigma_i \psi_{ii} + \Sigma_{i\neq j}\psi_{ij}, \tag{6.3}$$

where $\psi_{ij} = \psi(X_{i1}^0 - X_{i0}^0, X_{j0}^0 - X_{j1}^0)$. Note that $W(1,0)$ is the sum of two terms: the term $\Sigma\psi_{ii}$ is the sign-test statistic, whereas the second term recovers interblock information by comparing aligned observations across blocks. Large positive values of $W(1,0)$ are evidence that $\beta_1 > \beta_0$.

To test H_0: $\beta_1 = \beta_0$, it is easy to see that $E_{H_0}\{W(1,0)\} = 0$ since (X_{i1}, δ_{i1}) and (X_{i0}, δ_{i0}) are exchangeable for all i under H_0. Schluchter (1985) proved that $W(1,0)$ is asymptotically normally distributed, and the estimator $\{R_1^2(1,0) + ... + R_n^2(1,0)\}$ converges in probability to $\text{var}\{W(1,0)\}$ under the null hypothesis. It then follows that

$$Z = W(1,0)/\{\Sigma R_i^2(1,0)\}^{\frac{1}{2}} \tag{6.4}$$

has asymptotically a standard normal distribution when the null hypothesis is true.

6.2 Test the treatment effect in the breast cancer study using the aligned rank test

We see that the estimated treatment effect in the Cox proportional hazards model stratified by pair with only treatment as covariate is the ratio of the number of comparable pairs where the treated patient dies before the control patient and the number of comparable pairs where the control patient dies before the treated patient, which is same as the information used in the rank test for the paired data. To recover the "interpair" information, the aligned rank test is used in the following analysis.

For the breast cancer data set, 47 pairs of the 88 pair-matched observations which are not doubly censored were used to calculate the test statistic $W(1,0)/\{\Sigma R_i^2(1,0)\}^{\frac{1}{2}}$. Here,

$$\begin{aligned} W(1,0) &= -1015, \\ \widehat{var}\{W(1,0)\} &= \Sigma R_i^2(1,0) = 138,415, \end{aligned}$$

and $Z = W(1,0)/\{\Sigma R_1^2(1,0)\}^{\frac{1}{2}} = -2.728189$, the corresponding P-value is 0.0065. There is a very strong evidence against the null hypothesis that $\beta_1 = \beta_0$. And $W(1,0)$ has a very big negative value, which implies that the survival time of the patient in the vitamin treated group is significantly shorter than that of the control group.

In the above model, we ignored the fact that most of the patients in the treated group did not start the vitamin treatment at the time of diagnosis. Table 6.1 gives ranks for fixed and time dependent treatment effect separately for all the possible orderings of $X_{i1}^0 - X_{i0}^0$, $X_{j0}^0 - X_{j1}^0$ and ${}_aH_i$. To simplify the table, we suppose that both $X_{i1}^0 - X_{i0}^0$

Ordering	Rank for fixed treatment	Rank for time-dependent treatment
${}_aH_i < X^0_{i1} - X^0_{i0} < X^0_{j0} - X^0_{j1}$	-1	-1
${}_aH_i < X^0_{j0} - X^0_{j1} < X^0_{i1} - X^0_{i0}$	1	1
$X^0_{j0} - X^0_{j1} < {}_aH_i < X^0_{i1} - X^0_{i0}$	1	0

Table 6.1: Comparing the rank for the fixed and time dependent treatment effect

and $X^0_{j0} - X^0_{j1}$ are not censored, and the adjusted time to the start of vitamin treatment, ${}_aH_i$, is comparable with $X^0_{j0} - X^0_{j1}$ and always less than $X^0_{i1} - X^0_{i0}$. The rank for the time-dependent treatment effect, in the case when a patient in the control group died before the patient in the treated group started the vitamin treatment, is 0 instead of 1 as for the fixed treatment effect, and same as the fixed treatment effect in all the other cases. Therefore, the actual value of $W(1,0)$ should be smaller than -1015, which indicates after adjusted treatment as a time-dependent covariate the treatment effect would be more significant.

Chapter 7

Cox model and the results for Data set II

7.1 Description and preliminary analysis of Data set II

There are 2461 women breast cancer patients in Data set II who did not take large doses of vitamins as far as can be known with year of diagnosis between 1989 and 1996, and no distant metastasis. After 106 bilateral breast cancer women were deleted, the remaining 2355 cases were used as the control group for Data set II. The treated group are selected from Data set I using the following criteria: year of diagnosis within 1989 to 1996, no missing values in *dthcsurv*, female, no bilateral breast cancer, no distant metastasis and administered the additional vitamin treatment. We have 100 cases that satisfied the above criteria, which formed the treated group for Data set II.

7.1.1 Homogeneity tests for the matching variables including *dxposnod* and *histcat*

Table 7.1 displays the summary of the variables considered, including the number of valid cases and the missing values for each of them. Tables 7.2 to 7.25 provide the tests of the homogeneity of the covariate distributions for the treated versus the control group for the discrete variables which were used as matching variables in data set I as well as *dxposnod* and *histcat*. There is no evidence against the homogeneity hypothesis for variables *bccasr*, *dxer*, *dxlvn*, *dxyear*, *histcat*, *staget* with P-values > 0.1, moderate evidence against homogeneity for variables *bccahr* and *bccard* with P-values within $(0.1, 0.05)$, and strong evidence against homogeneity for variables *bccach*, *dxgrade*, *dxposnod* and *stagepn*. Tables 7.26 and 7.27 are the t-tests for equality of the means of *agedx* between treated and control groups, which indicates that the mean of age at diagnosis for the patients in treated group is significantly larger than that of the control group. Table 7.28 gives the summary of the difference between the treated and control groups. Note that in the Pearson Chi-square test some of the cells have small expected frequency, for example *staget*, *histcat*, *stagepn* and *bccasr*, and the P-values are not accurate for those tables.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
treated/control * bccach (yes or no)	2455	100.0%	0	.0%	2455	100.0%
treated/control * bccahr (yes or no)	2455	100.0%	0	.0%	2455	100.0%
treated/control * bccard	2455	100.0%	0	.0%	2455	100.0%
treated/control * bccasr	2455	100.0%	0	.0%	2455	100.0%
treated/control * Estrogen Receptor	2014	82.0%	441	18.0%	2455	100.0%
treated/control * Tumor Grade	1935	78.8%	520	21.2%	2455	100.0%
treated/control * Invasive Ivn	2126	86.6%	329	13.4%	2455	100.0%
treated/control * Nodal Status	2455	100.0%	0	.0%	2455	100.0%
treated/control * Diagnosis Year	2455	100.0%	0	.0%	2455	100.0%
treated/control * Categorical Histology	2455	100.0%	0	.0%	2455	100.0%
treated/control * Pathological N-stage	2454	100.0%	1	.0%	2455	100.0%
treated/control * T-stage	2435	99.2%	20	.8%	2455	100.0%

Table 7.1: Case processing summary for homogeneity tests

Crosstab

			bccach (yes or no)		
			N	Y	Total
treated/control	control	Count	1920	435	2355
		Expected Count	1883.0	472.0	2355.0
	treated	Count	43	57	100
		Expected Count	80.0	20.0	100.0
Total		Count	1963	492	2455
		Expected Count	1963.0	492.0	2455.0

Table 7.2: Count of *bccach* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	88.864[b]	1	.000		
Continuity Correction [a]	86.476	1	.000		
Likelihood Ratio	69.509	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	2455				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 20.04.

Table 7.3: Chi-square test of *bccach*

Crosstab

			bccahr (yes or no)		
			N	Y	Total
treated/control	control	Count	1268	1087	2355
		Expected Count	1276.8	1078.2	2355.0
	treated	Count	63	37	100
		Expected Count	54.2	45.8	100.0
Total		Count	1331	1124	2455
		Expected Count	1331.0	1124.0	2455.0

Table 7.4: Count of *bccahr* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.241[b]	1	.072		
Continuity Correction [a]	2.882	1	.090		
Likelihood Ratio	3.289	1	.070		
Fisher's Exact Test				.081	.044
N of Valid Cases	2455				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 45.78.

Table 7.5: Chi-square test of *bccahr*

Crosstab

			bccard		
			N	Y	Total
treated/control	control	Count	844	1511	2355
		Expected Count	835.5	1519.5	2355.0
	treated	Count	27	73	100
		Expected Count	35.5	64.5	100.0
Total		Count	871	1584	2455
		Expected Count	871.0	1584.0	2455.0

Table 7.6: Count of *bccard* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.274[b]	1	.070		
Continuity Correction [a]	2.899	1	.089		
Likelihood Ratio	3.412	1	.065		
Fisher's Exact Test				.087	.042
N of Valid Cases	2455				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 35.48.

Table 7.7: Chi-square test of *bccard*

Crosstab

			bccasr			Total
			D	N	Y	
treated/control	control	Count	2	40	2313	2355
		Expected Count	1.9	39.3	2313.8	2355.0
	treated	Count	0	1	99	100
		Expected Count	.1	1.7	98.2	100.0
Total		Count	2	41	2412	2455
		Expected Count	2.0	41.0	2412.0	2455.0

Table 7.8: Count of *bccasr* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.371[a]	2	.831
Likelihood Ratio	.498	2	.780
N of Valid Cases	2455		

a. 3 cells (50.0%) have expected count less than 5. The minimum expected count is .08.

Table 7.9: Chi-square test of *bccasr*

Crosstab

			Estrogen Receptor		
			Negative	Positive	Total
treated/control	control	Count	730	1202	1932
		Expected Count	733.9	1198.1	1932.0
	treated	Count	35	47	82
		Expected Count	31.1	50.9	82.0
Total		Count	765	1249	2014
		Expected Count	765.0	1249.0	2014.0

Table 7.10: Count of *dxer* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.801[b]	1	.371		
Continuity Correction [a]	.607	1	.436		
Likelihood Ratio	.790	1	.374		
Fisher's Exact Test				.416	.217
Linear-by-Linear Association	.801	1	.371		
N of Valid Cases	2014				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 31.15.

Table 7.11: Chi-square test of *dxer*

Crosstab

			Tumor Grade			
			Well Differentiated	Moderately Differentiated	Poorly Differentiated	Total
treated/control	control	Count	360	950	543	1853
		Expected Count	359.1	937.5	556.4	1853.0
	treated	Count	15	29	38	82
		Expected Count	15.9	41.5	24.6	82.0
Total		Count	375	979	581	1935
		Expected Count	375.0	979.0	581.0	1935.0

Table 7.12: Count of *dxgrade* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11.569[a]	2	.003
Likelihood Ratio	10.973	2	.004
Linear-by-Linear Association	5.370	1	.020
N of Valid Cases	1935		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.89.

Table 7.13: Chi-square test of *dxgrade*

Crosstab

			Invasive lvn		
			Negative	Positive	Total
treated/control	control	Count	1352	683	2035
		Expected Count	1344.9	690.1	2035.0
	treated	Count	53	38	91
		Expected Count	60.1	30.9	91.0
Total		Count	1405	721	2126
		Expected Count	1405.0	721.0	2126.0

Table 7.14: Count of *dxlvn* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.610[b]	1	.106		
Continuity Correction[a]	2.258	1	.133		
Likelihood Ratio	2.532	1	.112		
Fisher's Exact Test				.114	.068
Linear-by-Linear Association	2.609	1	.106		
N of Valid Cases	2126				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 30.86.

Table 7.15: Chi-square test of *dxlvn*

Crosstab

			Nodal Status			
			No Axillary Dissection	No pos nodes	Pos nodes	Total
treated/control	control	Count	265	1350	740	2355
		Expected Count	260.0	1333.4	761.7	2355.0
	treated	Count	6	40	54	100
		Expected Count	11.0	56.6	32.3	100.0
Total		Count	271	1390	794	2455
		Expected Count	271.0	1390.0	794.0	2455.0

Table 7.16: Count of *dxposnod* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.602[a]	2	.000
Likelihood Ratio	21.174	2	.000
Linear-by-Linear Association	19.120	1	.000
N of Valid Cases	2455		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.04.

Table 7.17: Chi-square test of *dxposnod*

treated/control * Diagnosis Year Crosstabulation

Count

	Diagnosis Year								
	89	90	91	92	93	94	95	96	Total
treated/cor contro	268	256	304	283	309	300	285	350	2355
treate	12	10	14	13	14	14	12	11	100
Total	280	266	318	296	323	314	297	361	2455

Table 7.18: Count of *dxyear* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.414[a]	7	.985
Likelihood Ratio	1.496	7	.982
Linear-by-Linear Association	.336	1	.562
N of Valid Cases	2455		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.84.

Table 7.19: Chi-square test of *dxyear*

Crosstab

			Categorical Histology			Total
			Ductal	Lobular	Other	
treated/control	control	Count	2030	267	58	2355
		Expected Count	2033.6	261.9	59.5	2355.0
	treated	Count	90	6	4	100
		Expected Count	86.4	11.1	2.5	100.0
Total		Count	2120	273	62	2455
		Expected Count	2120.0	273.0	62.0	2455.0

Table 7.20: Count of *histcat* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3.516[a]	2	.172
Likelihood Ratio	3.861	2	.145
Linear-by-Linear Association	.264	1	.607
N of Valid Cases	2455		

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 2.53.

Table 7.21: Chi-square test of *histcat*

Crosstab

			Pathological N-stage					Total
			No axil disect	No Nodal Mets	Axillary Nodal Mets	Fixed Nodal Mets	Mammary Nodal Mets	
treated/control	control	Count	270	1353	701	29	1	2354
		Expected Count	266.7	1336.2	721.4	28.8	1.0	2354.0
	treated	Count	8	40	51	1	0	100
		Expected Count	11.3	56.8	30.6	1.2	.0	100.0
Total		Count	278	1393	752	30	1	2454
		Expected Count	278.0	1393.0	752.0	30.0	1.0	2454.0

Table 7.22: Count of *stagepn* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	20.362[a]	4	.000
Likelihood Ratio	18.822	4	.001
Linear-by-Linear Association	13.122	1	.000
N of Valid Cases	2454		

a. 3 cells (30.0%) have expected count less than 5. The minimum expected count is .04.

Table 7.23: Chi-square test of *stagepn*

Crosstab

			T-stage						
			In-situ	No Evidence of Tumor	< 2.01 cm	> 2 cm and < 5.01 cm	> 5 cm	Extended	Total
treated/contr	control	Count	144	2	1435	614	73	68	2336
		Expected Cou	143.9	1.9	1426.5	622.6	73.9	67.2	2336.0
	treated	Count	6	0	52	35	4	2	99
		Expected Cou	6.1	.1	60.5	26.4	3.1	2.8	99.0
Total		Count	150	2	1487	649	77	70	2435
		Expected Cou	150.0	2.0	1487.0	649.0	77.0	70.0	2435.0

Table 7.24: Count of *staget* for Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.764[a]	5	.445
Likelihood Ratio	4.651	5	.460
Linear-by-Linear Association	.847	1	.357
N of Valid Cases	2435		

a. 4 cells (33.3%) have expected count less than 5. The minimum expected count is .08.

Table 7.25: Chi-square test of *staget*

Group Statistics

	treated/control	N	Mean	Std. Deviation	Std. Error Mean
Age at Diagnosis	treated	100	49.68	11.75	1.18
	control	2355	62.58	13.10	.27

Table 7.26: Means and standard deviations of *agedx* by treated and control

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age at Diagnos	Equal variance assumed	3.446	.064	-9.680	2453	.000	-12.90	1.33	-15.51	-10.28
	Equal variance not assumed			-10.693	109.705	.000	-12.90	1.21	-15.29	-10.51

Table 7.27: Test the equality of means of *agedx* between treated and control

Variables	Difference between the treated and control groups	Significant level
bccahr	treated have more "N"	moderately
bccard	treated have more "Y"	moderately
bccach	treated have more "Y"	strongly
dxgrade	treated have more "Poorly differentiated"	strongly
dxposnod	treated have more "Pos nodes"	strongly
stagepn	treated have more "Axillary nodel mets"	strongly
agedx	treated are older than controls	strongly

Table 7.28: Summary of the difference between the treated and control groups

	Total	No. of events	No. of censored	Percent censored
Control	2355	319	2036	86.45%
Treated	100	45	55	55.00%
Overall	2455	364	2091	85.17%

Table 7.29: Case processing summary for K-M survival functions

7.1.2 Plot of Kaplan-Meier survival function

Figure 7.1 shows the plots of Kaplan-Meier survival function by treated and control, which is quite like the plots for the pair-matched data set. The control patients tend to have longer lifetimes. This is consistent with the analyses of the staging variables in Section 7.1.1 that the treated group has more seriously diseased patients than the control group. To evaluate the treatment effect we need to adjust for age and disease severity. Table 7.29 is the case processing summary for the plots.

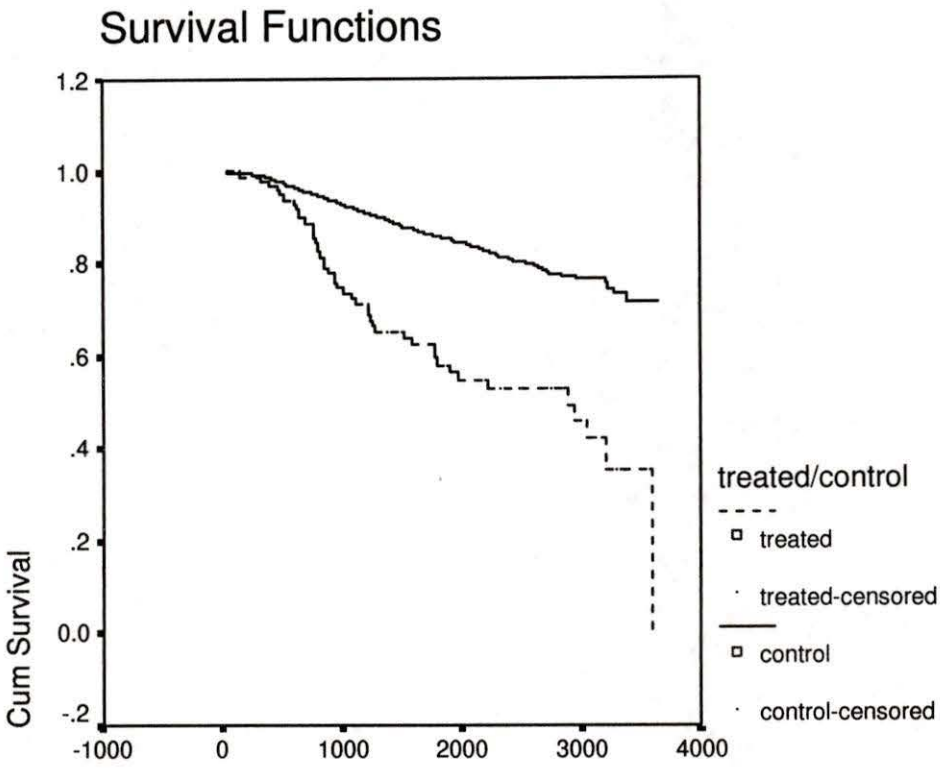

Figure 7.1: Plots of Kaplan-Meier survival functions by treated and control

7.2 Cox model and the results for Data set II

Cox models for Data set II were fitted in the following steps:

(1) Selection of the covariates: The time-dependent treatment effect and all the variables that are significantly effective in the models for Data set I plus the variables which are not homogeneously distributed between the treated and control groups are used as covariates for the full model. Those variables include *treatment*, *dxlvn*, *staget*, *dxgrade*, *stagepn*, *dxposnod*, *bccard*, *bccach* and *agedx*. Since the *agedx* effect on the hazard function is not proportional (Section 5.3.1), *catage* was used instead.

(2) Model fitting and the tests of the proportional hazards assumption for each of the covariates: The Cox model with time-dependent treatment effect and all the selected variables as covariates was fitted and the proportional hazards assumption was tested. We found that there is evidence against the proportional hazards assumption for the variables *dxgrade*, *stagepn* and *bccach*, the smallest P-values for each of the three variables are 0.00326, 0.05831 and 0.09406 respectively.

To examine the nonproportional hazards of *bccach* we plot the Kaplan-Meier survival function at each level of *bccach* in Figure 7.2. We see that there is no clear difference between each level and the curves cross each other. In order to include *bccach* in the Cox regression model we combined levels 1, 2, 3 and Y as a new level Y, indicating that the chemotherapy was given. The new *bccach*, denoted by *bccachyn*, has two levels, Y and N which indicates that chemotherapy was not given. The plot of the Kaplan-Meier survival function for *bccachyn* in Figure 7.3 shows the clear difference between level Y and N, and *bccachyn* was used as covariate in the Cox model instead of *bccach*.

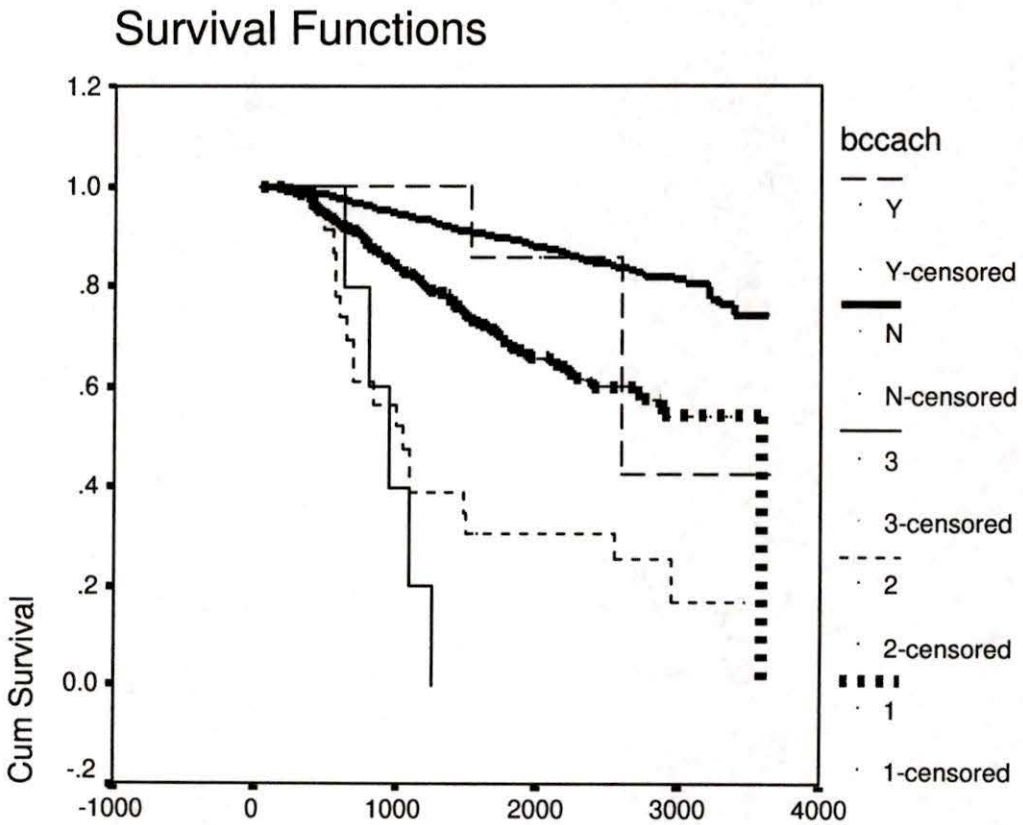

Figure 7.2: Plot of Kaplan-Meier survival functions for each level of *bccach*

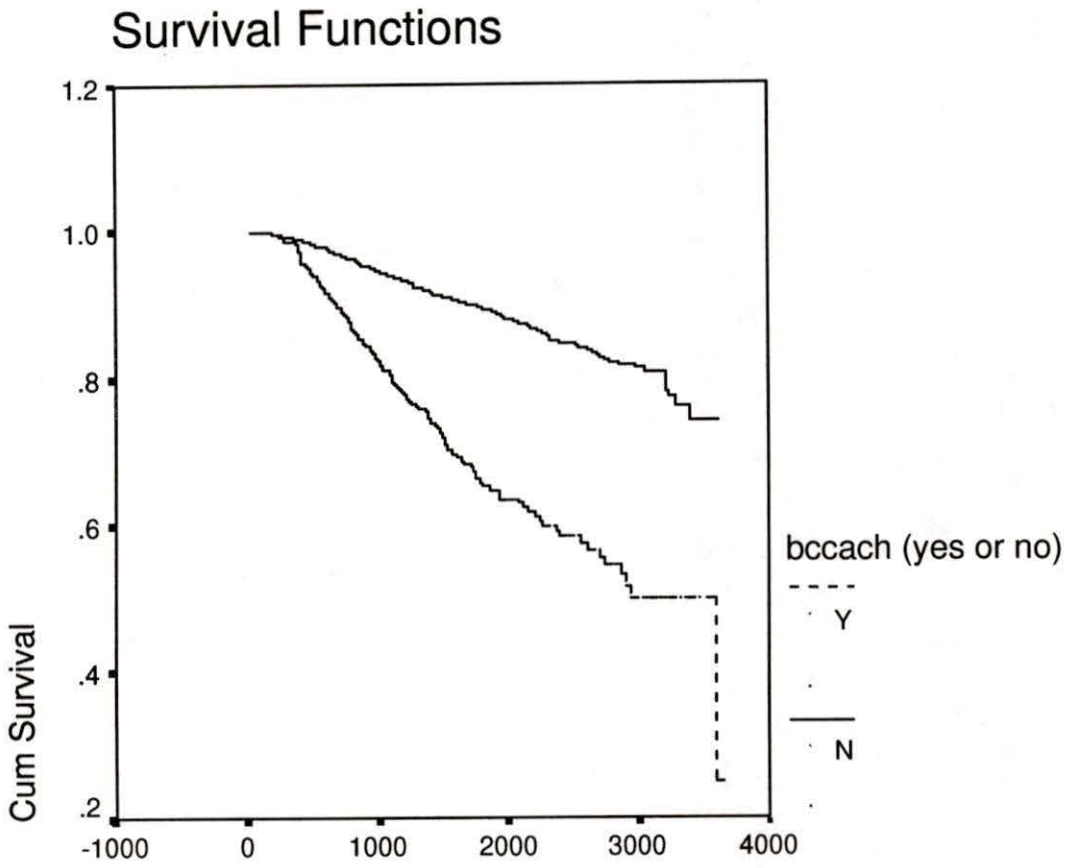

Figure 7.3: Plot of Kaplan-Meier survival functions for *bccachyn* with two levels

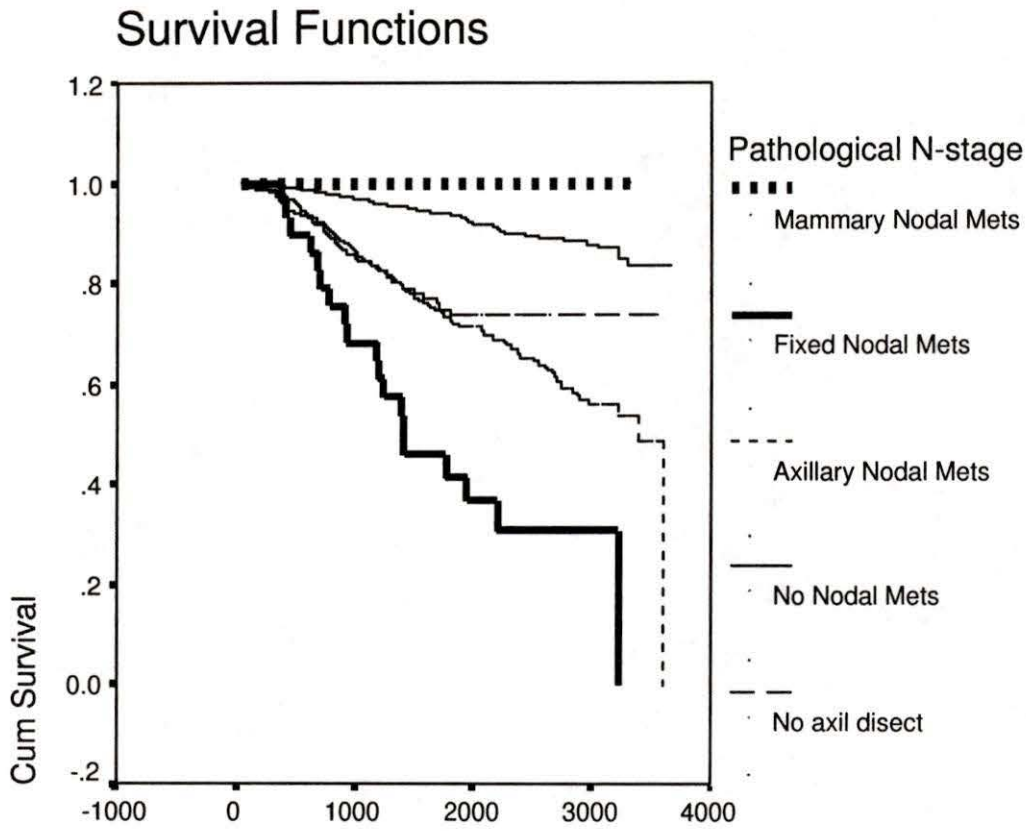

Figure 7.4: Plot of Kaplan-Meier survival functions for each level of *stagepn*

Similarly the variable *stagepn* is coded as *stapn* with $stagepn = -1$(no nodal mets) and $stagepn = 0$ (no axillary dissection) combined into one level for *stapn*. Figures 7.4 and 7.5 are the plots of the Kaplan-Meier survival functions for each level of *stagepn* and *stapn*; we can see the difference between the two levels of *stapn*. The variable *stapn* was used in the model instead of *stagepn*.

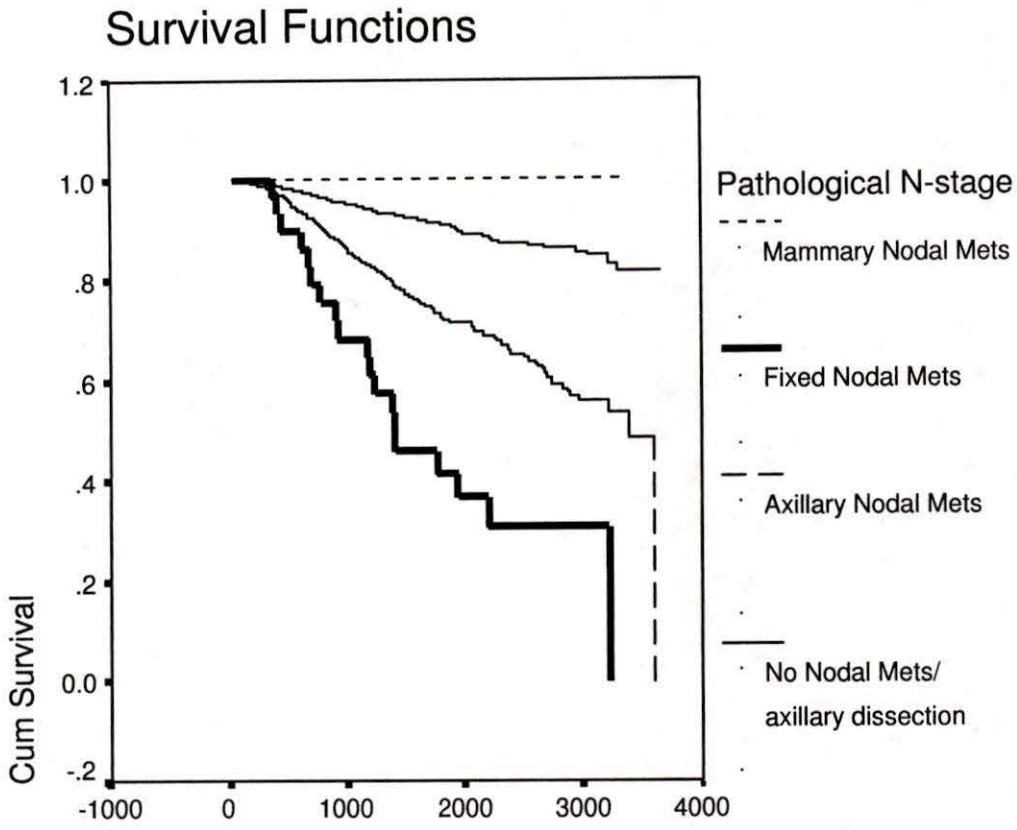

Figure 7.5: Plot of Kaplan-Meier survival functions for *stapn* with two levels

(3) We refitted the Cox model using *dxgrade* as strata, all the selected variables substituting *bccachyn* for *bccach* and *stapn* for *stagepn* as covariates. The proportional hazards assumption was tested again, no evidence against the assumption, then using the backward regression method, the reduced model with covariates *treatment*, *bccard*, *dxlvn*, *dxposnod*, *staget* and *stapn* is obtained.

(4) In order to identify the influential points the plots of scaled change in parameter estimate by observation number for the variables in the reduced model were checked.

Two cases (one control with *idtrans* = 64414070, one treated with *idtrans* = 62482070) have the largest influence on the estimation of the effects of the variables *staget* and *dxposnod*, and another case in the treated group (*idtrans* = 66654063), has the largest influence on the estimation of the effects of *stapn*.

(5) The reduced model in step (3) was fitted again without the three influential cases: There are 2455 total cases read, 646 cases with missing values, 73 censored cases before the earliest event in a stratum, the only case at *staget* = 0 and three influential cases were dropped. Table 7.30 is the case processing summary by *dxgrade*. Table 7.31 shows the parameter estimates for this model. After considering the effects of the covariates, *bccard*, *dxlvn*, *dxposnod*, *staget* and *stapn*, the treatment effect is still very significant, and the estimates of the effects of variables *dxlvn* and *staget* are similar to those for the models for Data set I. The estimated effect of "no pos nodes" of *dxposnod* is significantly better than "pos nodes" with P-value is close to zero, "axillary nodal mets" of *stapn* is significantly better than "fixed nodal mets" with P-value is 0.0117. The estimated effect of radiation therapy is significantly not good as no radiation therapy, one conclusion is that the radiation is not helpful in the treatment for the breast cancer, the other explanation is that the radiation therapy may relate with some staging variables that are not included in the model since only the staging variables that are not homogeneously distributed between the treated and control groups were considered in the Cox model. See Table 7.32 for the tests of the proportional hazards assumption and Table 7.33 for the indicator parameter coding.

dxgrade	Events (treated/control)	Censored (treated/control)	Percent censored
1	6/9	8/253	94.6%
2	7/103	19/793	88.1%
3	24/115	12/383	74.0%
Total	37/227	39/1429	84.8%

Table 7.30: Case processing summary by *dxgrade*

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.2364	.1859	44.2304	1	.0000	.1173	3.4431
bccard	-.5255	.1729	9.2352	1	.0024	-.0486	.5913
dxlvn	-.8170	.1409	33.6142	1	.0000	-.1015	.4418
dxposnod			28.7006	2	.0000	.0897	
dxposnod(1)	-.9222	.4935	3.4920	1	.0617	-.0220	.3976
dxposnod(2)	-1.9522	.4667	17.4957	1	.0000	-.0711	.1420
staget			24.8364	4	.0000	.0741	
staget(1)	-10.9584	172.5544	.0040	1	.9494	.0000	1.74E-5
staget(2)	-.9468	.2576	13.5.73	1	.0002	-.0612	.3880
staget(3)	-.3487	.2533	1.8952	1	.1686	.0000	.7066
staget(4)	-.7373	.3592	4.2140	1	.0401	-.0269	.4784
stapn			13.4271	2	.0044	.0554	
stapn(1)	.5545	.5293	1.0974	1	.2948	.0000	1.7410
stapn(2)	-.7149	.2835	6.3592	1	.0117	-.0377	.4892

Table 7.31: Parameter estimates for reduced model

Figures 7.6 to 7.11 are the plots which examine the influence of individual observations on the parameter estimates. The plots indicate that there are no seriously influential observations. Figure 7.12 is the plot of the martingale residuals. Most of the residuals are very close to zero.

Variable	rho	chisq	P-value
Treatment	.02503	.16878	.681
bccard	-.08797	2.13432	.144
dxlvn	.09934	2.61244	.106
dxposnod(1)	.02200	.16440	.685
dxposnod(2)	.00827	.02439	.876
staget(1)	.14950	.00744	.931
staget(2)	.02298	.12458	.724
staget(3)	.03848	.35234	.553
satget(4)	-.00392	.00400	.950
stapn (1)	-.03592	.42186	.516
stapn(2)	-.02344	.14749	.701
Global	NA	8.66041	.732

Table 7.32: Tests of the proportional hazards assumption for reduced model

Value	Freq (Treated/Control)	(1)	(2)	(3)	(4)
staget **(T-stage)**					
In-situ	2/18	1	0	0	0
<2.01cm	41/1072	0	1	0	0
>2cm and <5.01cm	27/468	0	0	1	0
>5cm	4/54	0	0	0	1
Extended	2/44	0	0	0	0
dxposnod **(Nodal status)**					
No axillary dissection	6/120	1	0	0	
No pos nodes	31/956	0	1	0	
Pos nodes	39/580	0	0	0	
dxlvn **(Invasive lvn)**					
Negative	45/1095	1	0		
Positive	31/561	0	0		
bccard **(Radiation therapy)**					
No	21/515	1	0		
Yes	55/1141	0	0		
stapn **(Coded from** *stagepn*)					
No axil disect/nodal mets	38/1083	1	0	0	
Axillary nodal mets	37/552	0	1	0	
Fixed nodal mets	1/21	0	0	0	

Table 7.33: Case processing summary and indicator parameter coding

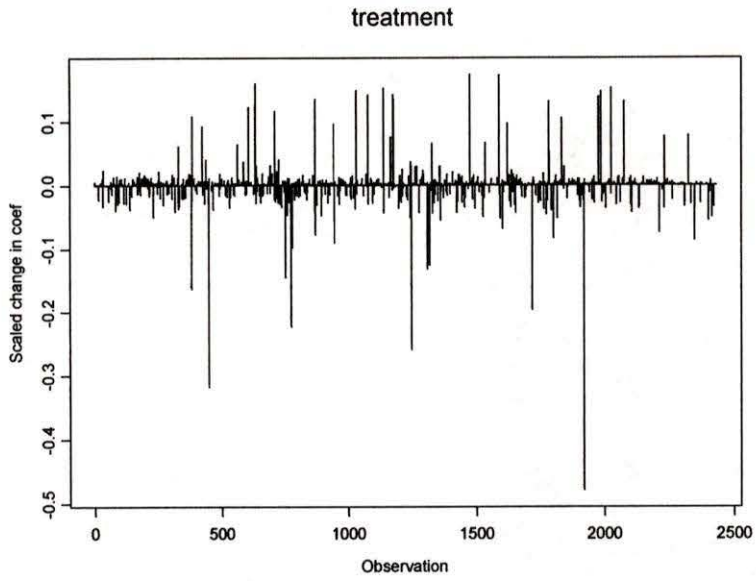

Figure 7.6: Plot of influence by observation number for *treatment*

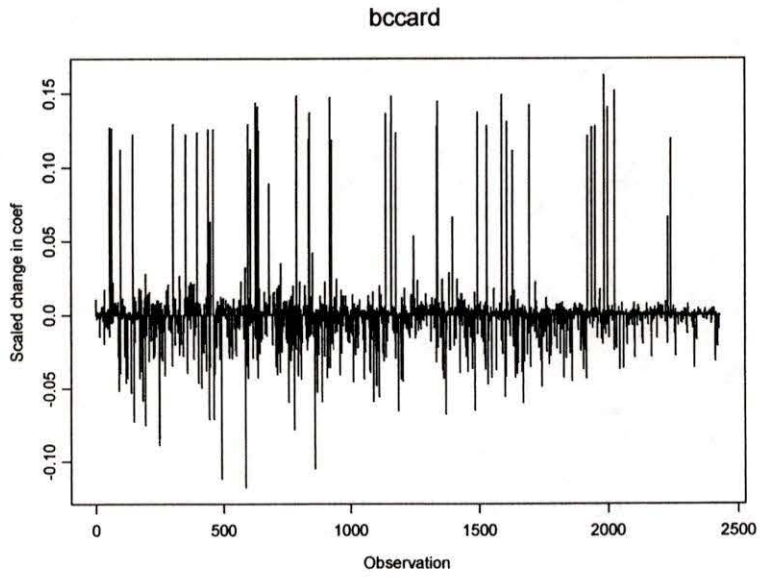

Figure 7.7: Plot of influence by observation number for *bccard*

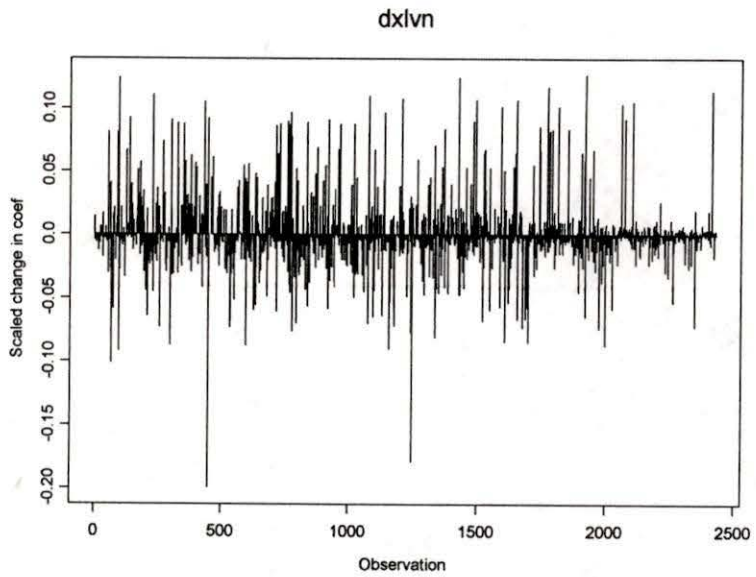

Figure 7.8: Plot of influence by observation number for *dxlvn*

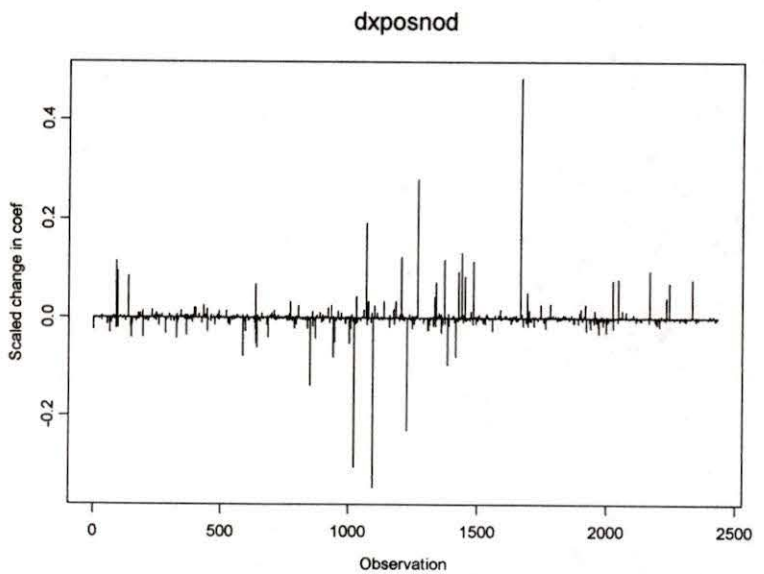

Figure 7.9: Plot of influence by observation number for *dxposnod*

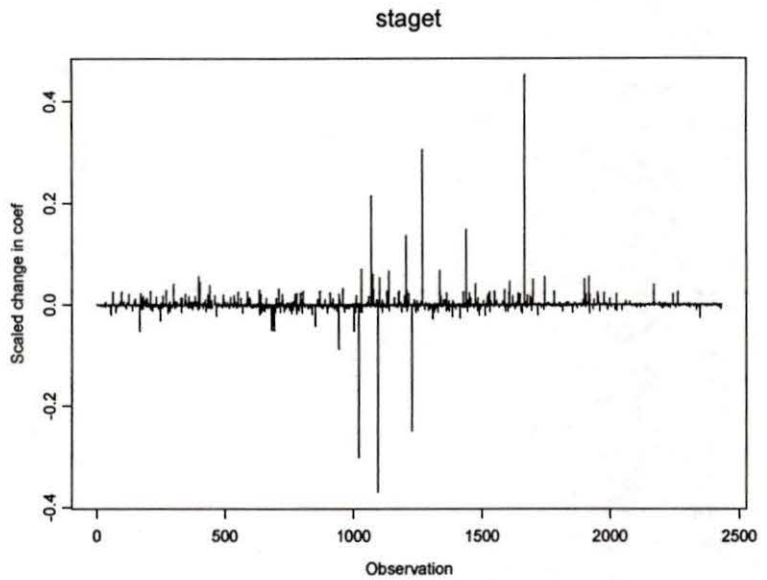

Figure 7.10: Plot of influence by observation number for *staget*

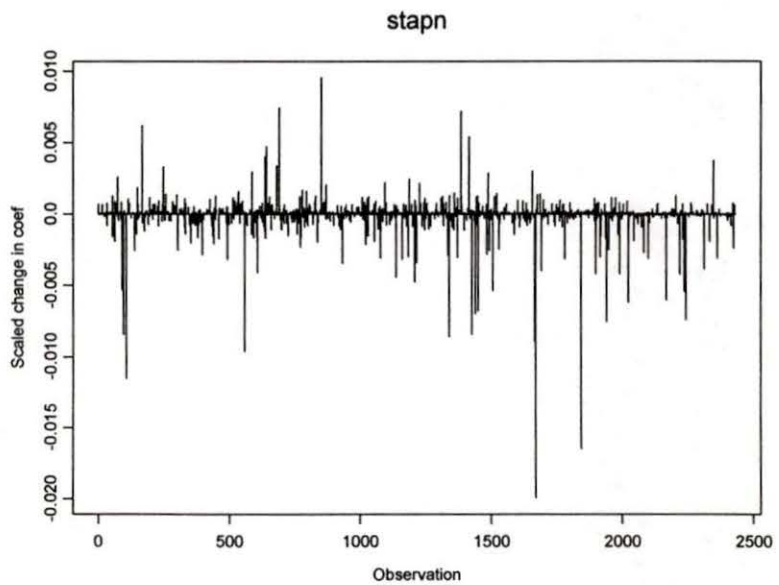

Figure 7.11: Plot of influence by observation number for *stapn*

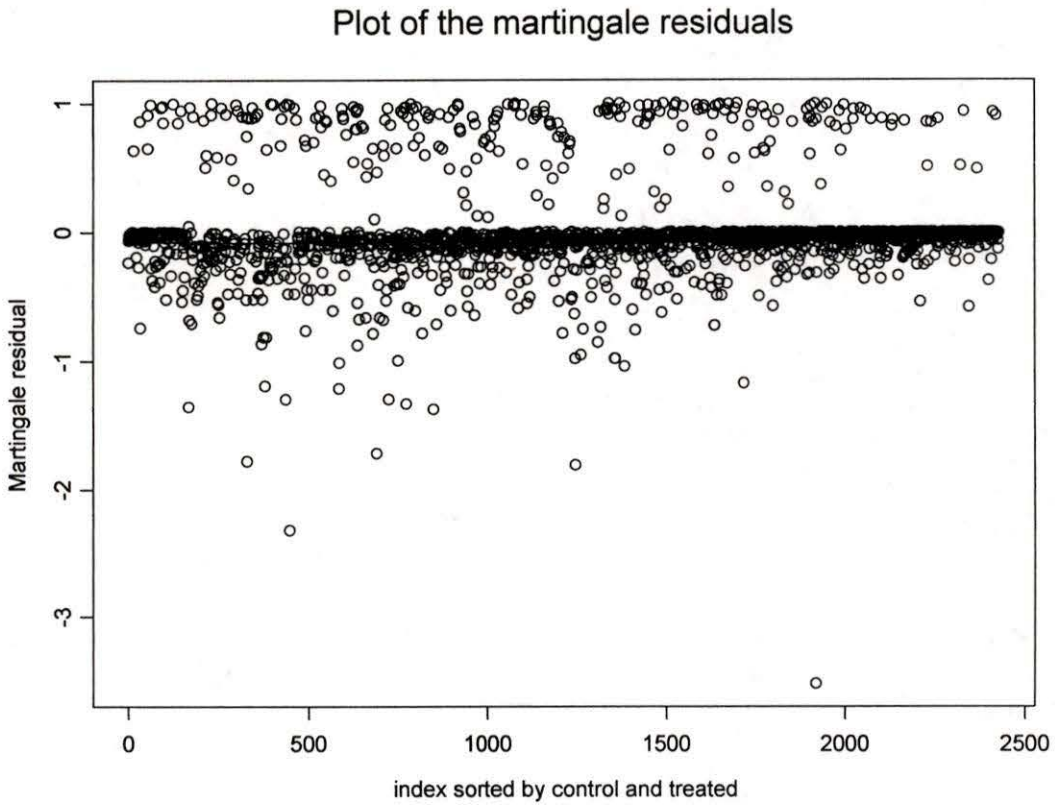

Figure 7.12: Plot of the martingale residuals

Chapter 8

Compare the time to distant relapse between treated and control groups

8.1 Cox model for distant relapse for Data set I

In this chapter, we focus on the event of distant relapse since it is a important outcome variable and compare the time to distant relapse of patients in the treated group with those in the control group. The patients in the treated group who did not see Dr. Hoffer before distant relapse were considered to be controls, who saw Dr. Hoffer before relapse were made in to two entries, the control entry is the time before she saw Dr. Hoffer and the treated entry is the time from she saw Dr. Hoffer. The Kaplan-Meier survival functions for the distant relapse free survival were plotted in Figure 8.1 for the treated and control groups

separately. The curve for the treated is under that of the control for the survival time over 1000 days, but before that they very close to each other. This graph takes into account the time dependent nature of the treatment.

Table 8.1 is the case processing summary for the plots of Kaplan-Meier survival functions. The 164 controls include the 82 patients who had no vitamin treatment, 16 patients who started the vitamin treatment after relapse and 66 control entries of the treated patients who started the vitamin treatment before relapse. The 71 treated include 5 patients who started the vitamin treatment at the diagnosis date and 66 treated entries of the treated patients who started the vitamin treatment before relapse. The 7 observations, including 1 treated and 6 controls, with missing value in status variable *devent* are not included in the analysis.

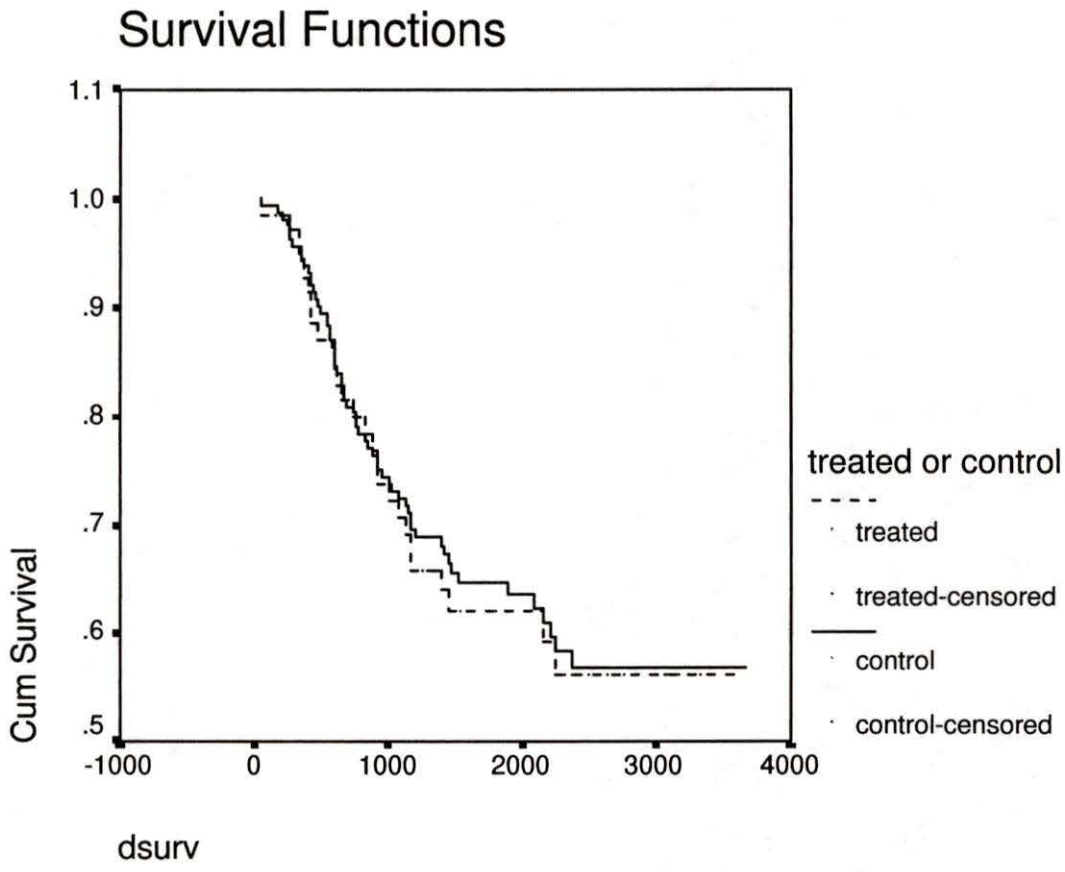

Figure 8.1: Plots of K-M survival functions for *dsurv* by treated and control groups

	Total	No. Event	No. Censored	Percent censored
Treated	71	27	44	61.97%
Control	164	60	104	63.41%
Overall	235	87	148	62.98%

Table 8.1: Case processing summary for K-M plots

The Cox model for distant relapse for Data set I was fitted in the following steps:

(1) Using *dsurv* as the response variable, the Cox model with time-dependent treatment effect and all the group I variables from Section 5.3.1 as covariates was fitted. Using the test described in Section 5.2.1, we found that there is evidence against the proportional hazards assumption for the variable *dxgrade*, the smallest P-value is 0.0159.

(2) We fitted the Cox model stratified by *dxgrade*, and tested the proportional hazards assumption. There is no evidence against the assumption, and we see that some of the variables are not significantly related with the time to the distant relapse.

(3) The model in step (2) was fitted again using backward elimination and variables *dxyear*, *agedx*, *histact* and *dxposnod* were removed with only the significant predictors, *staget*, *dxlvn* and *treatment* left in the reduced model.

(4) The plots of influence by observation number for the variables in the reduced model were checked. One censored case in the control group ($pair = 155$) with the largest influence on the estimation of the effects of *staget* and *dxlvn* was dropped from the data set.

(5) The reduced model in step (3) was fitted again without the influential case. There are 176 total cases read, 32 cases with missing values, 1 censored case before the earliest event in a stratum, the only case at $staget = 4$ and the influential case were

dxgrade	Events	Censored	Percent censored
1	5	14	73.7%
2	10	38	79.2%
3	34	40	54.1%
Total	49	92	65.2%

Table 8.2: Case processing summary by *dxgrade*

dropped. Table 8.2 is the case processing summary by *dxgrade*. Tables 8.3 and 8.4 are the parameter estimates and the tests of the proportional hazards assumption for this model. The estimated treatment effect on the hazard function is

$$\exp\{\widehat{\beta}_t\} = \exp\{.4105\} = 1.5076$$

with P-value 0.1943 which is not significant. This estimation is consistent with the above analysis according to the plots of the Kaplan-Meier survival functions. The estimated effects of *dxlvn* and *staget* are similar to that in the models for the survival time. See Table 5.5 for the indicator parameter coding.

Figures 8.2 to 8.4 are the plots of influence. All of the observations are reasonable. Figure 8.5 is the plot of the martingale residuals versus index with the control patients first. The hazard of remission of the patients in the treated group may be under estimated in this model since the smoothed line tends above zero for the treated patients in the plot.

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	.4105	.3163	1.6849	1	.1943	.0000	1.5076
dxlvn	-.5343	.3373	2.5111	1	.1130	-.0376	.5861
staget			10.5150	2	.0080	.1343	
staget(1)	-1.8934	.5940	10.1595	1	.0014	-.1502	.1506
staget(2)	-1.2473	.5446	5.2455	1	.0220	-.0948	.2873

Table 8.3: Parameter estimates for reduced model

Variable	rho	chisq	P-value
Treatment	-.000154	1.41e-06	.999
dxlvn	.007907	2.74e-03	.958
staget(1)	-.075119	3.26e-01	.568
staget(2)	-.013684	9.85e-03	.921
Global	NA	7.83e-01	.941

Table 8.4: Tests of the proportional hazards assumption for reduced model

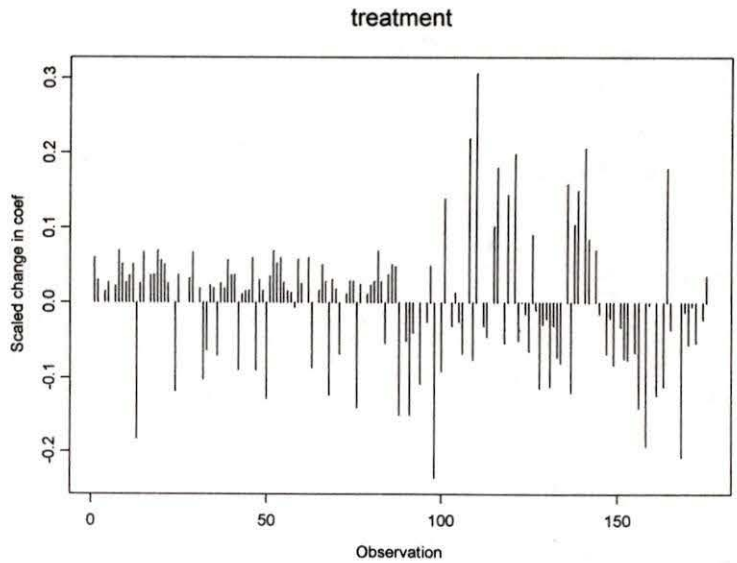

Figure 8.2: Plot of influence by observation number for *treatment*

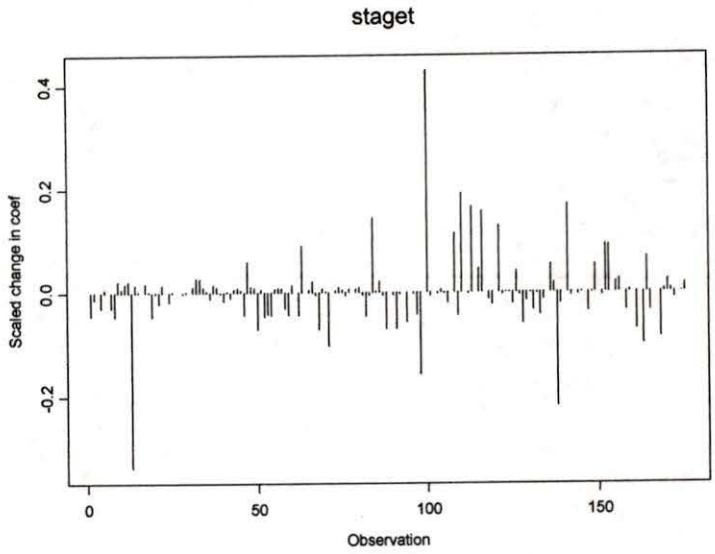

Figure 8.3: Plot of influence by observation number for *staget*

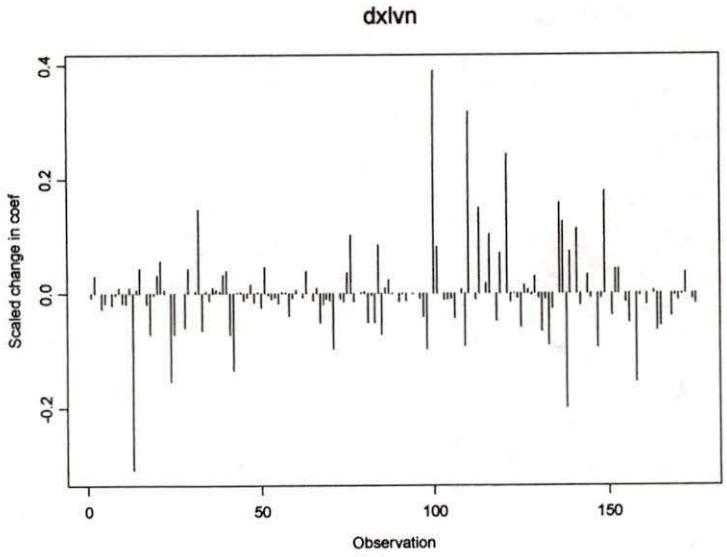

Figure 8.4: Plot of influence by observation number for *dxlvn*

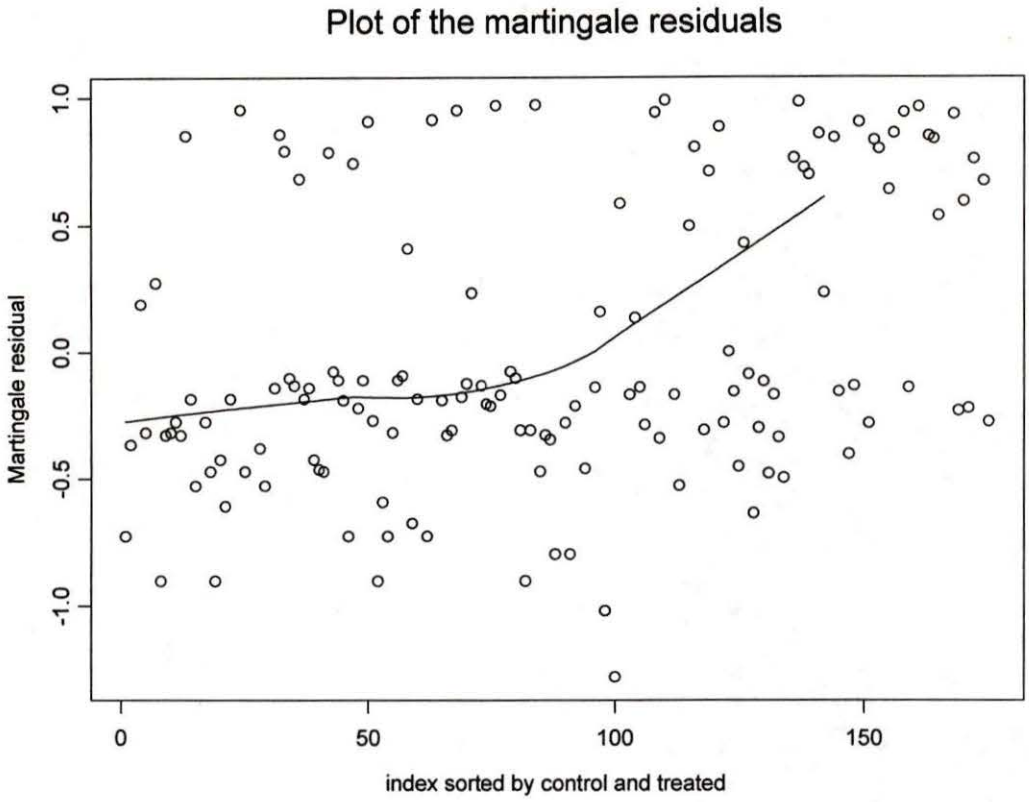

Figure 8.5: Plot of the martingal residuals

8.2 Cox model for distant relapse for Data set II

As for Data set I, the Kaplan-Meier survival functions for distant relapse were plotted in Figure 8.6 for the treated and control groups separately. The curve for the treated is under that of the control, which indicates that breast cancer patients' times to remission for the treated women are shorter than those of the control women. Table 8.5 is the case processing summary for the plots of Kaplan-Meier survival functions.

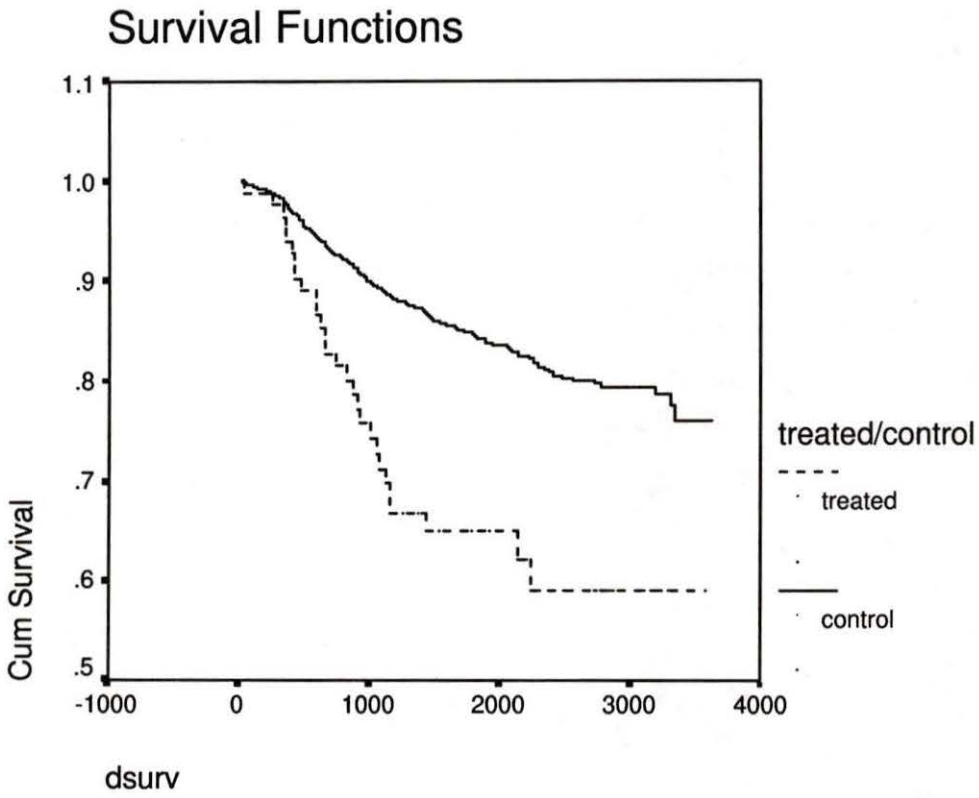

Figure 8.6: Plots of K-M survival functions for *dsurv* by treated and control groups

	Total	**No. Events**	**No. Censored**	**Percent censored**
Treated	83	28	55	66.27%
Control	2427	342	2085	85.91%
Overall	2510	370	2140	85.26%

Table 8.5: Case processing summary for K-M plots

The Cox model for distant relapse for Data set II was fitted in the following steps: First the Cox model with time-dependent treatment effect and all the selected variables in Section 7.2 as covariates was fitted and the proportional hazards assumption was tested. We found that there is evidence against proportional hazards assumption for the variable *bccach* and *staget*.

Second we refitted the Cox model using *staget* as strata, all the selected variables except *bccach* which was replaced with *bccachyn* as covariates. The proportional hazards assumption was tested again, no evidence against the assumption, then using the backward regression method we obtained the reduced model with covariates *treatment*, *bccard*, *dxlvn*, *dxgrade*, *stagepn* and *bccachyn*.

Third the plots of influence by observation number for the variables in the reduced model were checked. One censored case in the treated group (*idtrans* = 66535098) with the largest influence in the estimation of the effect of *bccachyn* was dropped from the data set.

Finally the reduced model in the second step was fitted again without the influential case. There are 2455 total cases read, 646 cases with missing values, 23 censored cases before the earliest event in a stratum and one influential case were dropped. Table 8.6 is the case processing summary. Table 8.7 shows the parameter estimates for this model. After considering the effects of the covariates, *bccard*, *dxlvn*, *dxgrade*, *stagepn* and *bccachyn*, the treatment effect on the distant relapse free survival time is still very significantly different from zero. The estimated treatment effect on the hazard function is $\exp\{\widehat{\beta}_t\} = \exp\{1.0002\} = 2.7187$, which indicates that the breast cancer patients' relapse

staget	Events	Censord	Percent censored
-1	1	0	0%
1	101	1067	91.4%
2	117	392	77%
3	15	45	75%
4	12	35	74.5%
Total	246	1539	86.2%

Table 8.6: Case processing summary by *staget*

times for the treated women are shorter than those of the control women after accounting for the matching variables. The estimated effects of *bccard*, *dxlvn* and *dxgrade* are similar to that in the models for the survival time. For *stagepn*, the estimated effect of "no nodal mets" is better than "fixed nodal mets" with P-value is 0.0001, "axillary nodal mets" or "no axillary nodal mets" is better than "fixed nodal mets" with P-values are 0.0074 and 0.0979 respectively, and "no nodal mets" is the best case; for *bccachyn*, "no chemotherapy" is better than "have chemotherapy" with P-value is 0.0104, which may because chemotherapy is related with some staging variables that are not included in the Cox model.

See Table 8.8 for the tests of the proportional hazards assumption and Tables 7.33 and 8.9 for the indicator parameter coding.

Figures 8.7 to 8.12 are the plots to examine the influence of individual observations on the parameter estimates. The plots are reasonable for all of the observations. Figure 8.13 is the plot of the martingale residuals for this model. Most of the residuals are very close to zero.

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Treatment	1.0002	.2339	18.2861	1	.0000	.0753	2.7187
bccachyn	-.3803	.1484	6.5666	1	.0104	-.0398	.6837
bccard	-.8712	.1962	19.7208	1	.0000	-.0785	.4185
dxlvn	-.7622	.1438	28.1104	1	.0000	-.0953	.4666
dxgrade			17.3427	2	.0001	.0681	
dxgrade(1)	-.9614	.2614	13.53.4	1	.0002	-.0633	.3824
dxgrade(2)	-.4092	.1384	8.7367	1	.0031	-.0484	.6642
stagepn			19.5830	3	.0005	.0687	
stagepn(1)	-.6530	.3945	2.7401	1	.0979	-.0160	.5205
stagepn(2)	-1.2951	.3316	15.2579	1	.0001	-.0679	.2739
stagepn(3)	-.8497	.3174	7.1682	1	.0074	-.0424	.4275

Table 8.7: Parameter estimates for reduced model

Variable	rho	chisq	P-value
Treatment	-.03879	.4059	.524
bccachyn	-.00628	.0118	.913
bccard	-.07434	1.3992	.237
dxlvn	.01175	.0316	.859
dxgrade(1)	.03989	.3948	.530
dxgrade(2)	.00726	.0135	.908
stagepn(1)	-.07564	1.4404	.230
stagepn(2)	-.02869	.2026	.653
stagepn(3)	-.03148	.2430	.622
Global	NA	4.6611	.863

Table 8.8: Tests of the Proportional hazards assumption for reduced model

Value	(1)	(2)	(3)
bccachyn (coded from bccach)			
No	1	0	
Yes	0	0	
stagepn			
No axil disect	1	0	0
No nodal mets	0	1	0
Axillary nodal mets	0	0	1
Fixed nodal mets	0	0	0

Table 8.9: Indicator parameter coding

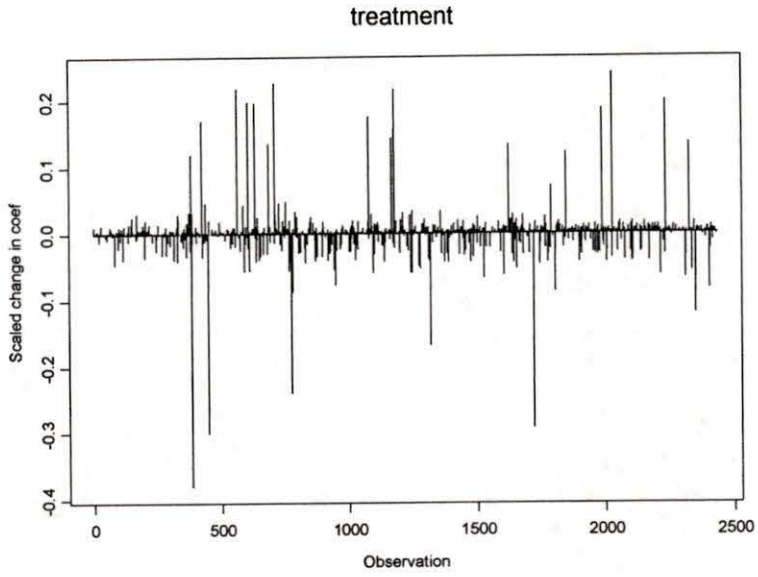

Figure 8.7: Plot of influence by observation number for *treatment*

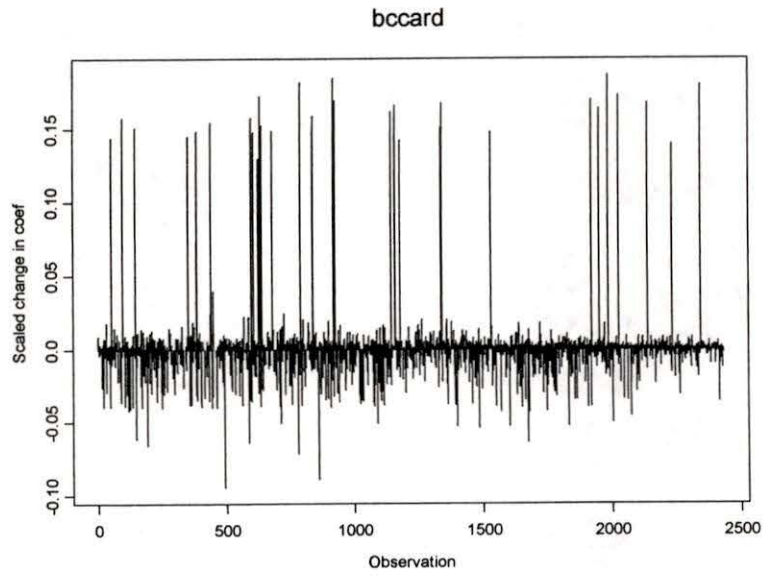

Figure 8.8: Plot of influence by observation number for *bccard*

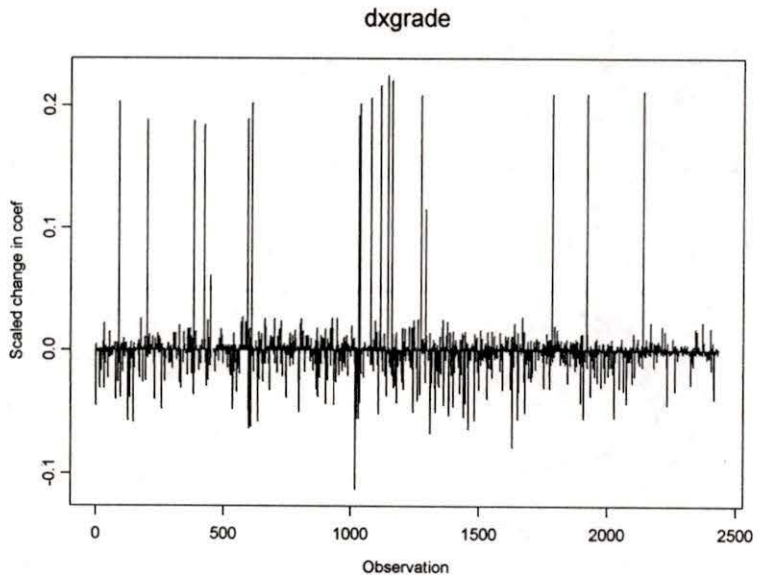

Figure 8.9: Plot of influence by observation number for *dxgrade*

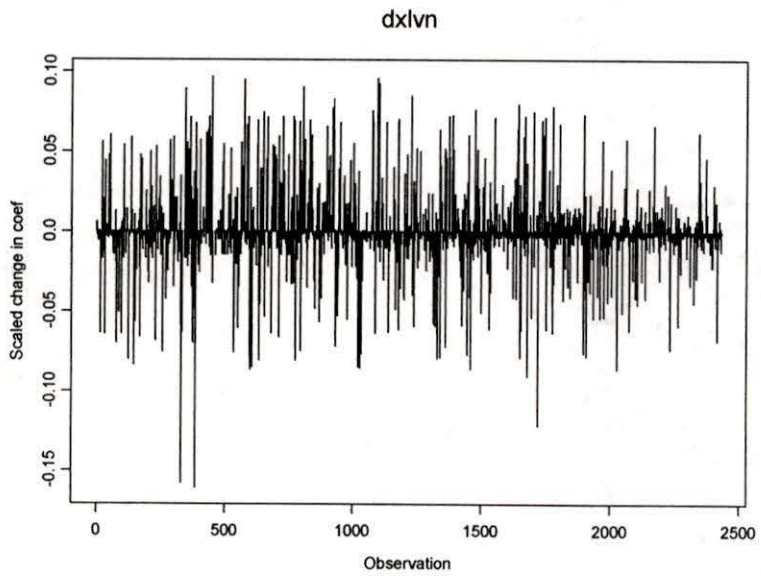

Figure 8.10: Plot of influence by observation number for *dxlvn*

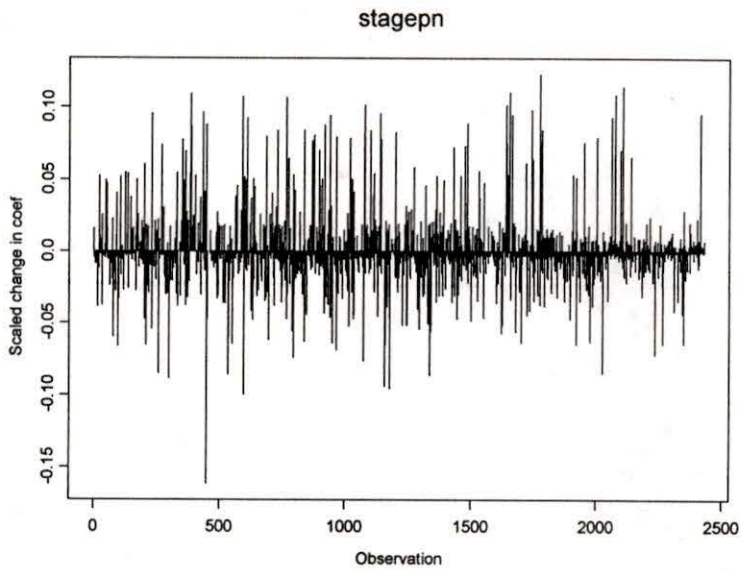

Figure 8.11: Plot of influence by observation number for *stagepn*

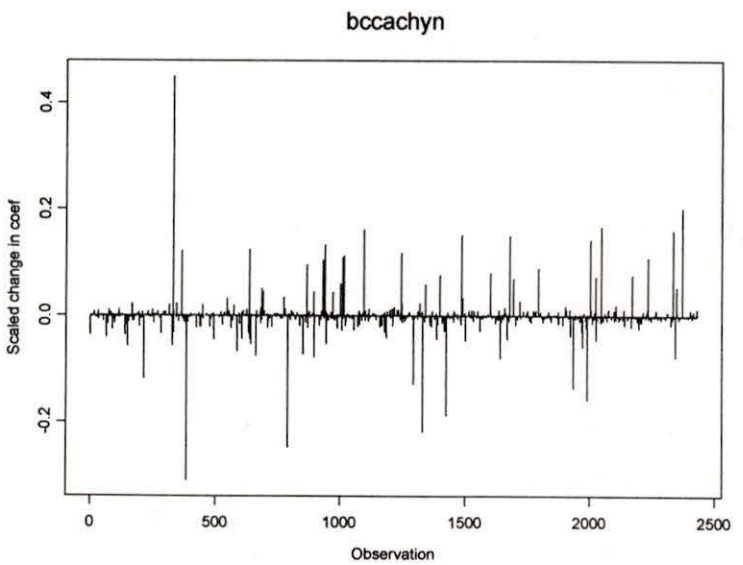

Figure 8.12: Plot of influence by observation number for *bccachyn*

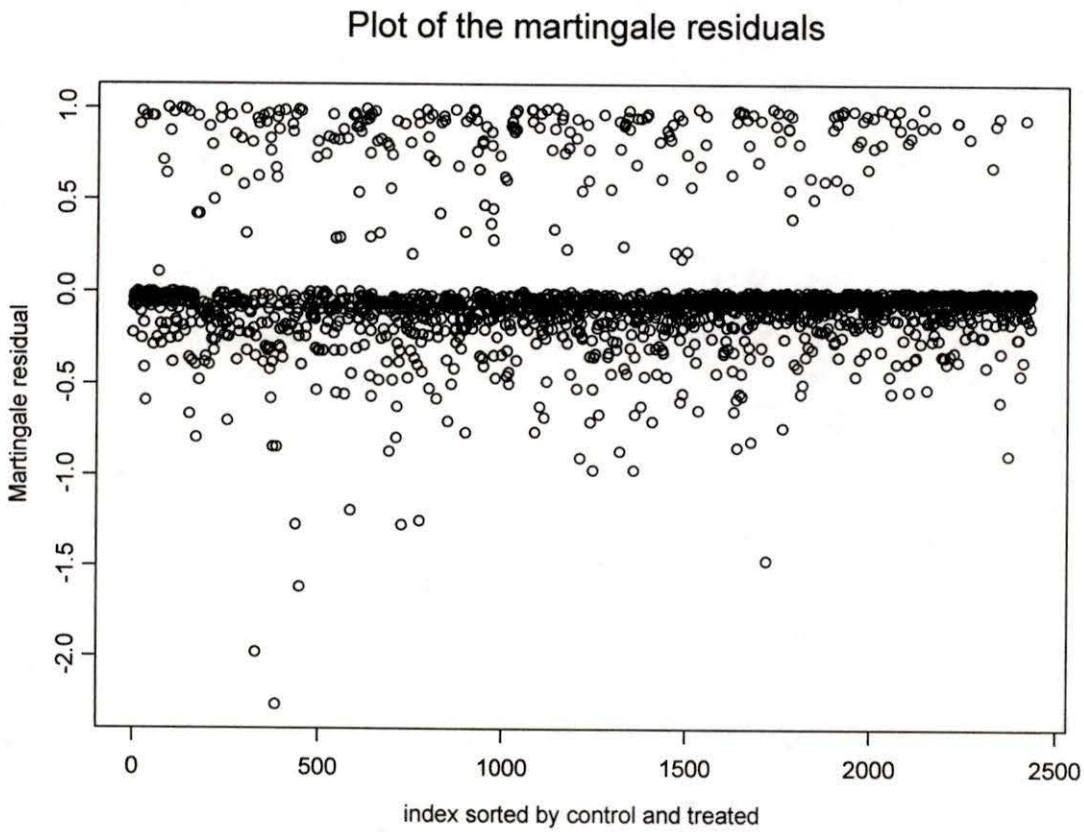

Figure 8.13: Plot of the martingal residuals of the Cox model for distant relapse

Chapter 9

Conclusions and comments

From the above analyses of the data sets of women breast cancer patients diagnosed between 1989 and 1996 without distant metastasis and bilateral breast cancer, we conclude that breast cancer patients' life times for the treated women are shorter than those of the control women after accounting for the matching variables in our study. The study on the distant relapse free survival gives a similar result in Data set II, but no significant difference for Data set I.

We have provided a comparison of the treated patients versus two control groups where we have controlled for diagnostic variables thought to be related to breast cancer survival. There may be confounding factors that we have not accommodated in our analyses, for example the nutritional status and *dxer*, but which explain the differences between the treated and control groups. The patients in the treated group could be taking higher doses of vitamins and minerals than recommended or having other damaging treatment but we do not have information. Therefore we cannot conclude that vitamins and minerals

shorten the lives of breast cancer patients. Dr. Hoffer's patients maybe a subgroup of those receiving conventional treatment who feel that it is not working; they seck Hoffer's regime looking for alternatives and the alternatives do not seem to work. Note that this is an observational study. The treated women chose to take the vitamin treatment at some point in time; they were not randomized to the treatment. Also, the treated women chose **when** they started the vitamin tratment. There may have been some aspects of their disease which prompted them to seek alternative therapy about which we do not have information. Further study in the form of clinical trials could be done to shed light on the effect of vitamin treatment on survival time.

The reliability of the analyses of the survival time with heavy censoring depends on the assumption of the same censoring mechanism for the treated and control groups. For example, among the 176 observations in Data set I, 66.5% of the survival times are censored. It difficult to test the assumption of the censoring. To date, we know that 7 patients were lost to follow-up, 4 patients died because of causes other than breast cancer, 106 patients survived to the date when the data was collected and the remaining 59 patients died in breast cancer. If the study period is extended until most of the patients die we can obtain more reliable results, pairs in which both patients were censored provide no contribution to the estimation of the treatment effect.

The parameter estimates of the treatment effect in the Cox model for both Data set I and Data set II are quite close, which indicates the pair-matched experiment is a reliable study.

In the above analyses the treatment effect is used as the time dependent binary

variable that jumps from 0 to 1 at the point of application of the treatment. A Cox model stratified by pair with fixed treatment effect as covariate was also fitted. From the analyses of the model stratified by *pair* we know that the model with fixed treatment effect will slightly under estimate the treatment effect since there are only 2 pairs of the observations with control patient died before the patient in the treated group start the vitamin treatment. The parameter estimate of this model gives $\widehat{\beta}_t = 1.1632$ which is a little smaller than $\widehat{\beta}_t = 1.3863$ in model (5). So the parametric model with fixed treatment effect may be used to analyze this data set.

All of the Cox models were fitted in both SPSS and S-plus, and the results are the same.

Bibliography

[1] Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. The Annals of Statistics, Vol. 10, No. 4, 1100-1120.

[2] Bendich A. (1991). Beta-carotene and the immune response. Proceedings of the Nutrition Society, 50:253-274.

[3] Bollage, W. and Holdener E. E. (1992). Retinoids in cancer prevention and therapy. Annals of Oncology, 3:513-526.

[4] Breslow, N. (1972). Contribution to the discussion of the paper by D.R. Cox. Journal of the Royal Statistical Society B. 34 187-220.

[5] Breslow, N. (1974). Covariance analysis of censored survival data. Biometrics, 30 89-99.

[6] Cameron, E. and Campbell, A. (1974). The orthomolecular treatment of Cancer II. Clinical trial of high-dose ascorbic supplements in advanced cancer. Chemical-Biological Interactions, 9: 285-315.

[7] Cameron, E. and Campbell, A. (1991). Innovation versus quality control: an unpublish-

able clinical trial of supplemental ascorbate in incurable cancer. Medical Hypotheses, 36: 185-189.

[8] Cameron, E. and Pauling, L. (1979). Cancer and vitamin C. W.W. Norton and Co. New York.

[9] Cameron, E. and Pauling, L. (1993). Cancer and vitamin C. Camino Books: Philadelphia.

[10] Campbell, A., Jack, T. and Cameron, E. (1991). Reticulum cell sarcoma: two complete 'spontaneous' regressions, in response to high-dose ascorbic acid therapy. A report on subsequent progress. Oncology, 48: 495-7.

[11] Canadian Breast Cancer Research Initiative. (1996). Vitamin A, C and E supplements: an information package CBCRI: Toronto.

[12] Cobleigh, M. A. Breast cancer and fenretinide, an analogue of vitamin A. Leukemia, 8 (Supplements 3): S59-S63.

[13] Cook, R. and Lawless, J. F. (2000). Analysis of event history data. Boistatistics workshop. Annual meeting of Statistical Society of Canada. Ottawa, Canada, June 4, 2000.

[14] Cox, D. R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society, B. 34 187-220.

[15] Cox, D. R. (1975). Partial likelihood. Biometrika 62 269-276.

[16] Frey, J. R., Peck, R. and Bollag, W. (1991). Antiproliferative activity of retinoids,

interferon p and their combination in five human transformed cell lines. Cancer Letters, 57: 223-227.

[17] Gopalakrishna, R. (1994). Vitamin E succinate inhibits protein kinase C., correlation with its unique inhibitory effects on cell growth and transformation. Second Denver Conference on Nutrition and Cancer, September 7-11, 1994. Denver, Co.

[18] Grambsch, P. M., and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81, 3, pp. 515-526.

[19] Groth, A. and Litmann, I. (1948). Ascorbic acid content in human cancer tissue. Cancer Research, 8: 349-351.

[20] Hartmann et al. (1996). An absent correlation between antioxidant blood concentrations and the remission response of preoperatively treated breast carcinomas. Strahlenther Onkol, 172(8): 434-8.

[21] Hodges, J. D. and Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. Ann. Math. Statist. 33, 482-97.

[22] Ip, C. and Ip, M.M. 1981 cited in Clark R. L., Cumley, R. W. and Hickey, R. C. The Year Book of Cancer 1983. Chicago: Year Book Medicinal Publishers, 385-387.

[23] Jackson, J.A., Riordan, H. D., Hunninghake, R.E. and rioradan, N. (1995). High dose intravenous vitamin C and long term survival of a patient with cancer of the head of the pancreas. Journal of Orthomolecular Medicine, 10(2), 87-88.

[24] Kalbfleisch, J. D. and Prentice, R. L. (1980). Analysis of paired failure times. The statistical analysis of failure time data, Chapter 8.1.

[25] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.

[26] Kline, K. (1994). Vitamin E succinate: mechanisms of action as tumor cell growth inhibitor. Second Denver Conference on Nutrition and Cancer, September 7-11, 1994. Denver, Co.

[27] Knekt, P. (1991). Role of vitamin E in the prophylaxis of cancer. Annals of Medicine, 23: 3-12.

[28] Knekt, P., Aromaa A., Maatela, J., Aaran, R.K., Nikkara, T., Hakama, M., Makulinen, T., Peto, R. and Teppo, L. (1991). Vitamin E and cancer prevention. Journal of Clinical Nutrition, 53: 283S-286S.

[29] Lippman, S.M., Kavanagh, J.J., Paredes-Espinoza, M., Delgadillo-Madrueno, F., Paredes-Casillas, P. and Hong, W.K. (1992). 13-cis-retinoic acid plus interferon p-2a: highly active systemic therapy for squamous cell carcinoma of the cervix. Journal of National Cancer Institute, 84(4): 241-245.

[30] Longnecker, M.P., Martin-Morreno, J.M., Kneke, P., Nomura A.M., Schober, S.E., Stahelin, H.B., Wald, N.J., Grey, K.F. and Willett, W.C. (1992). Serum alpha-tocopherol concertrations in relation to subsequent colorectal cancer: pooled data from five cohorts. Journal of National Cancer Institute, 84: 430-435.

[31] Mates, Donna, Data Analyst, Breast Outcomes Unit. (1999). Private communication at BC Cancer Agency, Vancouver.

[32] Moriguchi, S. and Kishino Y. (1990). In vitro activation of tumoricidal properties of human monocytes by beta-carotene encapsulated in liposomes. Nutrition Research, 10:837-846.

[33] Palan, P.R., Mikhail, M.S. and Rommey, S.L. (1991). Plasma levels of antioxidant beta-carotene and alpha tocopherol in uterine cervix dysplasia and cancer. Nutrition and Cancer, 15: 13-20.

[34] Richdan, H.D., Jackson, J.A. and Schulz, M. ((1990). Case study: High dose intravenous vitamin C in the treatment of a patient with adenocaricinoma of the kidney. Journal of Orthomolecular Medicine, 5: 5-7.

[35] Rustin, G.J. (1992). Therapy of solid tumors with retinoids, monotherapy and combination therapy (abstract). European Journal of Cancer, 29A(Supplements 6): S42 (Abstract No. 202).

[36] Schluchter, M. D. (1985). An aligned rank test for censored data from randomized block designs. Biometrika, 72, 3, pp. 609-618.

[37] Shimpo, K., Nagatsu, T., Kamadak, Sato, T., Niini, H., Shamoto, M. et al. (1991). Asorbic acid and adriamycin toxicity. American Journal of Clinical Nutrition, 54(Supplement): 1298S-1301S.

[38] Stone, I. (1972). The healing Factor: vitamin C against disease. Grosset and Dunlap, New York.

[39] SPSS version 9.0.

[40] S-plus version 2000.

[41] Turley et al. (1997). Vitamin E succinate induces Fas-mediated apoptosis in estrogen receptor-negative human breast cancer cells. Cancer Research, 57(5): 881-90

[42] Veronesi et al. (1999). Randomized trial of fenretinide to prevent second breast malignancy in women with early breast cancer. Journal of the National Cancer Institute, 91(21): 1847-56

Appendix A

Tables of the Codes for *histcat*

Tables 1 to 3 display the codes of *hist*1, *hist*2 and *hist*3 for *histcat* = 1. Tables 4 to 6 are the codes of *hist*1, *hist*2 and *hist*3 for *histcat* = 2.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	80102	1	.0	.0	.0
	80103	13	.6	.6	.6
	80503	3	.1	.1	.7
	81403	23	1.0	1.0	1.8
	81413	6	.3	.3	2.0
	82012	5	.2	.2	2.2
	82013	4	.2	.2	2.4
	82113	88	3.9	3.9	6.3
	82603	1	.0	.0	6.4
	84012	1	.0	.0	6.4
	84013	1	.0	.0	6.4
	84803	57	2.5	2.5	9.0
	84813	1	.0	.0	9.0
	85002	66	2.9	2.9	11.9
	85003	1845	81.4	81.4	93.3
	85012	56	2.5	2.5	95.8
	85032	1	.0	.0	95.8
	85033	2	.1	.1	95.9
	85103	19	.8	.8	96.7
	85223	29	1.3	1.3	98.0
	85303	20	.9	.9	98.9
	85403	3	.1	.1	99.0
	85412	2	.1	.1	99.1
	85413	18	.8	.8	99.9
	85433	1	.0	.0	100.0
	85723	1	.0	.0	100.0
	Total	2267	100.0	100.0	

Table 1: Code of *hist*1 for $histcat = 1$

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1853	81.7	81.7	81.7
	80213	1	.0	.0	81.8
	80502	8	.4	.4	82.1
	80503	1	.0	.0	82.2
	80953	1	.0	.0	82.2
	81402	1	.0	.0	82.3
	81403	2	.1	.1	82.4
	81413	3	.1	.1	82.5
	82012	104	4.6	4.6	87.1
	82013	9	.4	.4	87.5
	82113	4	.2	.2	87.6
	82302	50	2.2	2.2	89.9
	82303	1	.0	.0	89.9
	82603	1	.0	.0	89.9
	84013	1	.0	.0	90.0
	84803	4	.2	.2	90.2
	84813	1	.0	.0	90.2
	85002	23	1.0	1.0	91.2
	85003	5	.2	.2	91.4
	85012	168	7.4	7.4	98.9
	85013	3	.1	.1	99.0
	85032	2	.1	.1	99.1
	85042	1	.0	.0	99.1
	85202	18	.8	.8	99.9
	85213	1	.0	.0	100.0
	85433	1	.0	.0	100.0
	Total	2267	100.0	100.0	

Table 2: Code of *hist*2 for $histcat = 1$

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		2202	97.1	97.1	97.1
	80502	5	.2	.2	97.4
	82012	31	1.4	1.4	98.7
	82302	20	.9	.9	99.6
	84012	1	.0	.0	99.6
	85012	8	.4	.4	100.0
	Total	2267	100.0	100.0	

Table 3: Code of $hist3$ for $histcat = 1$

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	85202	13	4.4	4.4	4.4
	85203	187	63.4	63.4	67.8
	85212	2	.7	.7	68.5
	85213	93	31.5	31.5	100.0
	Total	295	100.0	100.0	

Table 4: Code of $hist1$ for $histcat = 2$

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		251	85.1	85.1	85.1
	81403	1	.3	.3	85.4
	82013	2	.7	.7	86.1
	82113	3	1.0	1.0	87.1
	84903	1	.3	.3	87.5
	85002	4	1.4	1.4	88.8
	85003	15	5.1	5.1	93.9
	85012	1	.3	.3	94.2
	85013	4	1.4	1.4	95.6
	85202	2	.7	.7	96.3
	85203	9	3.1	3.1	99.3
	85212	2	.7	.7	100.0
	Total	295	100.0	100.0	

Table 5: Code of *hist*2 for *histcat* = 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		289	98.0	98.0	98.0
	82112	1	.3	.3	98.3
	85002	2	.7	.7	99.0
	85003	1	.3	.3	99.3
	85212	2	.7	.7	100.0
	Total	295	100.0	100.0	

Table 6: Code of *hist*3 for *histcat* = 2

The following syntax codes, given by Mates, which only use *hit*1 was also used to calculate the variable *histcat*. Table 7 is the cross-tabulation of the *histcat* given in the Data set II, we call it *histactold*, and the new *histcat* calcualted by the following codes, we call it *histcatnew*.

Syntax codes:

recode hist1 ('85003' =1) ('80102' =1) ('80103' =1) ('80113' =1) ('80123'=1) ('80203' =1) ('80213' =1) ('80303' =1) ('80313' =1) ('80323'=1) ('80343' =1) ('80353' =1) ('80502' =1) ('80503' =1) ('81402'=1) ('81403' =1) ('81408' =1) ('81409' =1) ('81413' =1) ('82012'=1) ('82013' =1) ('82112' =1) ('82113' =1) ('82313' =1) ('82401'=1) ('82602' =1) ('82603' =1) ('84012' =1) ('84013' =1) ('84703'=1) ('84803' =1) ('84813' =1) ('85002' =1) ('85008' =1) ('85009'=1) ('85012' =1) ('85013' =1) ('85018' =1) ('85032' =1) ('85033'=1) ('85042' =1) ('85043' =1) ('85103' =1) ('85113' =1) ('85123' =1) ('85213'=1) ('85222' =1) ('85223' =1) ('85303' =1) ('85403' =1) ('85412'=1) ('85413' =1) ('85433' =1) ('85503' =1) ('85603' =1) ('85703'=1) ('85723' =1) ('85733' =1) ('85202' =2) ('85203' =2) ('85209'=2) ('85223' =1) ('84903' =2) ('85212' =1) (else = 3) into histcat.

		histcatold			Total
		1	2	3	
histcatnew	1	396	13	12	421
	2	0	39	0	39
	3	0	0	2	2
Total		396	52	14	462

Table 7: Crosstab of *histcatold* and *histcatnew*

Appendix B

SPSS and S-plus codes

B.1 SPSS codes

- Code 9 to systemissing

  RECODE

  dxer dxgrade dxlvn stagepn staget (9=SYSMIS) .

  EXECUTE .

- Paired-t test to check if the continuous variables, *agedx* and *dxyear* are matched within pair

  T-TEST

  PAIRS= aged0 dxyear0 WITH aged1 dxyear1 (PAIRED)

  /CRITERIA=CIN(.95)

  /MISSING=ANALYSIS.

- Crosstabulation of the matching variables for control versus case

```
CROSSTABS
/TABLES=agedx0 BY agedx1
/TABLES=dxyear0 BY dxyear1
/TABLES=staget0 BY staget1
/TABLES=stagepn0 BY stagepn1
/TABLES=bccasr0 BY bccasr1
/TABLES=bccard0 BY bccard1
/TABLES=bccach0 BY bccach1
/TABLES=bccahr0 BY bccahr1
/TABLES=dxgrade0 BY dxgrade1
/TABLES=dxer0 BY dxer1
/TABLES=dxlvn0 BY dxlvn1
/FORMAT= AVALUE TABLES
/CELLS= COUNT
```

- Chi-square tests for the homogeneity of the unmatched variables

```
CROSSTABS
/TABLES=control BY stagepn
/TABLES=control BY bccard
/TABLES=control BY bccach
/TABLES=control BY bccahr
/TABLES=control BY dxgrade
/TABLES=control BY dxer
```

```
/TABLES=control BY dxlvn
/FORMAT= AVALUE TABLES
/STATISTIC=CHISQ
/CELLS= COUNT
CROSSTABS
/TABLES=control2 BY bccach bccahr bccard bccasr dxer dxgrade dxlvn
dxposnod dxyear histcat stagepn staget stagem
/FORMAT= AVALUE TABLES
/STATISTIC=CHISQ
/CELLS= COUNT EXPECTED .
```

- Construct the frequency table for variable dthcevt

```
FREQUENCIES
 VARIABLES=dthcevt
 /ORDER ANALYSIS .
```

- Plot the Kaplan-Meier survival function with *treatment* as factor, *bccach* as strata

```
KM
dthcsurv BY control /STRATA=bccach /STATUS=dthcevt(1)
/PRINT NONE
/PLOT SURVIVAL HAZARD
```

- Create the summary table for the 26 missing observations

```
FILTER OFF.
USE ALL.
SELECT IF((dxlvn = 9) | (dxposnod = 9) | (dxgrade = 9) | (staget = 9)).
EXECUTE .
CROSSTABS
 /TABLES=control BY dthcevt dxlvn dxposnod staget dxgrade
 /FORMAT= AVALUE TABLES
 /CELLS= COUNT .
```

- Summary of the observations used in the regression

```
FILTER OFF.
USE ALL.
SELECT IF((dxlvn < 9) & (dxposnod < 9) & (staget < 4) & (dxgrade < 9)).
EXECUTE .
 /TABLES=dthcevt dxlvn dxposnod dxgrade staget BY control
 /FORMAT= AVALUE TABLES
 /CELLS= COUNT .
```

- Fit the Cox regression model stratified by *dxgrade* with time-dependent treatment effect and *bccachyn*, *bccard*, *dxlvn*, *dxposnod*, *catage*, *staget* as covariates using the backward regression method.

```
TIME PROGRAM.
COMPUTE T_COV_ = (T_>daytoh1)*1 .
```

COXREG

dthcsurv /STATUS=dthcevt(1) /STRATA=dxgrade

/CONTRAST (bccachyn)=Indicator /CONTRAST (bccard)=Indicator

/CONTRAST (dxlvn)=Indicator /CONTRAST (dxposnod)=Indicator

/CONTRAST (staget)=Indicator

/METHOD=BSTEP(COND) bccachyn bccard dxlvn dxposnod catage staget

/PRINT=CI(95) SUMMARY

/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .

- Fit the Cox model by fixing the treatment effect in the model

TIME PROGRAM.

COMPUTE T_COV_ = (T_ > daytoh1)*1 .

COXREG

ddsurv /STATUS=devent(1) /STRATA=dxgrade

/CONTRAST (dxlvn)=Indicator /CONTRAST (dxposnod)

=Indicator /CONTRAST

(staget)=Indicator /CONTRAST (histcat)=Indicator

/METHOD=ENTER T_COV_ /METHOD=BSTEP(COND)

agedx dxyear dxlvn dxposnod

staget histcat

/PRINT=CI(95) SUMMARY

/CRITERIA=PIN(.10) POUT(.15) ITERATE(20) .

- Select the treated group for Data set II from Data set I

FILTER OFF.

USE ALL.

SELECT IF((dxyear > 88)&(dxyear<97)&(control = 1)

&(dthcsurv >= 0)&(sex="F")).

EXECUTE .

- T-test for equality of means for agedx between the treated and control groups

T-TEST

GROUPS=control2(1 0)

/MISSING=ANALYSIS

/VARIABLES=agedx

/CRITERIA=CIN(.95) .

- Form the frequency table for the variable *dthcevt*

FREQUENCIES

VARIABLES=dthcevt

/ORDER ANALYSIS .

B.2 S-plus codes

- Fit the Cox regression model stratified by *dxgrade* with time-dependent treatment effect and the variables *bccard*, *dxlvn*, *dxposnod*, *staget* and *stapn* as covariates.

> cox1_coxph(formula = Surv(START, STOP, EVENT, type =

"counting") ~TREATMENT + BCCARD + strata(

DXGRADE) + DXLVN + DXPOSNOD + STAGET + STAPN,

data = data2, na.action = na.exclude, eps =

0.0001, iter.max = 10, method = "efron", robust

= F)

- Test the proportional hazards assumption for model cox1.

```
> test.cox1_cox.zph(cox1)
```

- Plot of the influence by the number of observations for each of the variables in model cox1.

```
> df1_resid(cox1, collapse=data2["IDTRANS"], type="dfbetas")
> plot(1:2430,df1[,1], type="h",ylab="Scaled change in coef",
  xlab="Observation", main="treatment")
> plot(1:2430,df1[,2], type="h",ylab="Scaled change in coef",
  xlab="Observation", main="bccard")
> plot(1:2430,df1[,3], type="h",ylab="Scaled change in coef",
  xlab="Observation", main="dxlvn")
> plot(1:2430,df1[,4], type="h",ylab="Scaled change in coef",
  xlab="Observation", main="dxposnod")
> plot(1:2430,df1[,5], type="h",ylab="Scaled change in coef",
  xlab="Observation", main="staget")
> plot(1:2430,df1[,6], type="h",ylab="Scaled change in coef",
```

```
xlab="Observation", main="stapn")
```

- Plot of the martingale residuals for model cox1.

```
> plot(residuals(cox1, collapse=data2[,"IDTRANS"]), xlab="index sorted
 by control and treated", ylab="Martingale residual",
 main="Plot of the martingale residuals")
> lines(lowess(na.omit(residuals(cox1, collapse=data2[,"IDTRANS"]))))
```

- Calculate the statistics for the aligned rank test

(1) Calculate the status variable for the aligned observations. A 1 indicates faliure, 0 indicates right censored and -1 indicates left consored with Ct is for the treated group and Cc for the control group.

```
> Ct_function(a, b)
{
 n <- length(a)
 ct <- c(rep(1, n))
 for(i in 1:n) {
 if((a[i] == 1) & (b[i] == 0))
 ct[i] <- -1
 if((a[i] == 0) & (b[i] == 1))
 ct[i] <- 0
 ct[i]
 }
```

```
ct
}

> Cc_function(a)
{
n <- length(a)
cc <- c(rep(1, n))
for(i in 1:n) {
if(a[i] == -1)
cc[i] <- 0
if(a[i] == 0)
cc[i] <- -1
cc[i]
}
cc
}
```

- (2) Form the function, R(x, s, y, a), for calculating the vector of rank for the treated group.

```
> R_function(x, s, y, a)
{
n <- length(x)
r <- c(rep(0, n))
```

```
for(i in 1:n) {
for(j in 1:n) {
if((s[i] == a[j]) & (a[j] == 1) & (x[i] > y[j]))
r[i] <- r[i] + 1
if((s[i] == a[j]) & (a[j] == 1) & (x[i] < y[j]))
r[i] <- r[i] - 1
if((s[i] < 0) & (0 < a[j]) & (x[i] < y[j]))
r[i] <- r[i] - 1
if((s[i] > 0) & (0 > a[j]) & (x[i] > y[j]))
r[i] <- r[i] + 1
if((s[i] < a[j]) & (a[j] == 0) & (x[i] < y[j]))
r[i] <- r[i] - 1
if((0 == s[i]) & (s[i] > a[j]) & (x[i] > y[j]))
r[i] <- r[i] + 1
if((s[i] > a[j]) & (a[j] == 0) & (x[i] < y[j]))
r[i] <- r[i] - 1
if((0 == s[i]) & (s[i] < a[j]) & (x[i] > y[j]))
r[i] <- r[i] + 1
r[i]
}
for(j in 1:n) {
if((s[i] == s[j]) & (s[j] == 1) & (x[i] > x[j]))
```

```
r[i] <- r[i] + 1
if((s[i] == s[j]) & (s[j] == 1) & (x[i] < x[j]))
r[i] <- r[i] - 1
if((s[i] < 0) & (0 < s[j]) & (x[i] < x[j]))
r[i] <- r[i] - 1
if((s[i] > 0) & (0 > s[j]) & (x[i] > x[j]))
r[i] <- r[i] + 1
if((s[i] < s[j]) & (s[j] == 0) & (x[i] < x[j]))
r[i] <- r[i] - 1
if((0 == s[i]) & (s[i] > s[j]) & (x[i] > x[j]))
r[i] <- r[i] + 1
if((s[i] > s[j]) & (s[j] == 0) & (x[i] < x[j]))
r[i] <- r[i] - 1
if((0 == s[i]) & (s[i] < s[j]) & (x[i] > x[j]))
r[i] <- r[i] + 1
r[i]
}
r[i]
}
r
}
```

- (3) Use (1) and (2) to calcualte the test statistic z.

Xt - The vector of survival time for the treated group;

St - Status vector for Xt;

Xc - The vector of survival time for the control group;

Sc - Status vector for Xc;

Dt - The vector of aligned observation for the treated group;

Et - Status vector for Dt;

Dc - The vector of aligned observation for the control group;

Ec - Status vector for Dc;

r - The vector of rank for the treated group.

```
> Xt_rankvit2[,"DTHCSURV"]
> St_rankvit2[,"DTHCEVT"]
> Xc_rankvit2[,"DTHCSUR0"]
> Sc_rankvit2[,"DTHCEVT0"]
> Dt_Xt-Xc
> Dc_-Dt
> Et_Ct(St, Sc)
> Ec_Cc(Et)
> r_R(Dt, Et,Dc,Ec)
> w_sum(r)
> std_sqrt(sum(r*r))
> z_w/std
```

- Identify the treated patients for Data set II by calculated the variable *control2*

  a - the id number of the whole data set;

  b - the id number of the Hoffer patients;

  control2 - with 1 indicates treated, 0 indicates control patient.

```
> L_function(a,b) {
n_length(a)
m_length(b)
g_c(rep(F,n))
for (i in 1:m)
g_g|(a==c(rep(b[i],n)))
control_as.numeric(g)
control
}
```

- Identify the bilateral breast patient by calculating the variable *bilateral*. A 0 indicates not bilateral and 1 indicates bilateral breast cancer.

```
> Bi
function(a)
{
 n <- length(a)
 b <- c(rep(0, n))
 for(i in 1:n) {
 b[i] <- b[i] + as.numeric(a[i] == a[i - 1])
```

```
b[i] <- b[i] + as.numeric(a[i] == a[i + 1])
b[i]
}
b
}
```

VITA

Surname: Zhao Given Name: Yang

Place of Birth: Shenyang, Liaoning province, PR China

Educational Institutions Attended:

University of Victoria, Canada	1998 to 2000
University of Louisville, USA	1997 to 1998
Nankai University, PR China	1988 to 1992

Degrees Awarded:

B.Sc.	Nankai University July	1992

Honours and Awards:

Nankai University Scholarship	1989 to 1991
Scholarship at University of Victoria	1998 to 2000

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis:

Mega-dose Vitamins and Minerals for the Treatment of Breast Cancer: A Comparison Study of Treated Nonmetastatic Patients Versus Two Control Groups

Author [redacted]

Yang Zhao

August 21, 2000