A Comparison Study of Nonmetastatic Breast Cancer Patients who were Prescribed Mega-doses of Vitamins and Minerals with a Matched Control group and a Control Population

by

Natalie Judith Forde
B.A., York University, 1995

A Thesis Submitted in Partial Fullfillment
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, Department of Mathematics & Statistics, University of Victoria

[redacted]
Dr. M. Tsao, Department of Mathematics & Statistics, University of Victoria

[redacted]
Dr. J.T. Buckley, Department of Biochemistry & Microbiology, University of Victoria

[redacted]
Dr. J. Chapman, Women's College Hospital, University of Toronto

Supervisor: Dr. M. Lesperance

Abstract

This study used the statistical methods of survival analysis to compare the life expectancy of breast cancer patients who received conventional cancer treatment along with mega-doses of vitamins and minerals, versus control patients who received conventional treatment alone. The vitamin and mineral regime was administered by Dr. Abram Hoffer, a practicing physician in Victoria, British Colombia. The Hoffer patients were compared to two different control groups. A set of matched controls was generated from the BC Cancer Agency database for the first analysis, and the entire population of nonmetastatic, nonbilateral breast cancer patients, diagnosed between the years 1989 to 1996, at the Vancouver Island BC Cancer Agency was used for the second analysis. The results of the analyses using parametric survival models indicate that women with nonmetastatic breast cancer, diagnosed between the years 1989 and 1996 who have no known bilateral cancer and received Hoffer's treatment, do not have longer survival times than those who did not receive Hoffer's treatment.

Examiners:

[redacted]

Dr. M. Lesperance, Department of Mathematics & Statistics, University of Victoria

[redacted]

Dr. M. Tsao, Department of Mathematics & Statistics, University of Victoria

[redacted]

Dr. J.T. Buckley, Department of Biochemistry & Microbiology, University of Victoria

[redacted]

Dr. J. Chapman, Women's College Hospital, University of Toronto

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Min Tsao and especially Dr. Mary Lesperance for all of their help and encouragement on this endeavour.

I would also like to thank Dr. Judy-Anne Chapman and Dr. J. Thomas Buckley for participating in my defence and providing suggestions and comments.

A special thanks to Caroline Speers and Donna Mates at the BC Cancer Agency in Vancouver, Dr. Abram Hoffer and Dr. Sean Bohun for all of their input.

Chapter 1

Introduction

Micronutrients such as vitamins and minerals are known to be vital for overall human health. Chemoprevention refers to the use of chemical agents such as vitamins, minerals and related nontoxic synthetic products and drugs to prevent the development of cancer. Vitamin A (retinol), the B vitamins, E (α-tocopherol), and C (ascorbic acid) as well as many minerals have been studied using experimental, clinical and epidemiological studies for cancer chemoprevention and treatment.

1.1 Recommended Intakes of Nutrients

A recommended nutrient intake is an estimate of the amount of dietary intake sufficient to meet the requirements of essentially all individuals in a group with similar specified characteristics {33}. It is assumed that requirements change progressively with age during growth, for males and females, therefore the levels of nutrient intakes take into account factors such as age and sex. The recommended intakes have been established for people who are considered healthy

individuals who eat a variety of common foods, and are expressed as intakes per day, however they represent average amounts and therefore can be ingested over a period of days or weeks {33}.

1.1.1 Vitamin A

Vitamin A found in food is comprised of two families, *preformed* vitamin A (*retinol*), and *provitamins* A (*carotenes*). β-carotene is considered one of the most important provitamins as it is efficiently converted into vitamin A in the human body. Other carotenes are assumed to have half the potency of β-carotene {33}.

Vitamin A has primary roles in vision, bone and tooth development, growth, and resistance to infection {33}. With respect to provitamin activity, one-third of ingested carotene is absorbed, and the efficiency of conversion of absorbed β-carotene to retinol is 50% {33}. Intakes of vitamin A can be expressed in terms of retinol equivalents (RE), or in terms of the International System of Units (SI). For humans, 1 RE is equivalent to 3.33 IU of activity from retinol or 10 IU from β-carotene {33}. A recommended intake of 1000 RE/day (3300 IU) has been made for adult man, and 800 RE/day for women to accomodate individual variability and permit maintenance of a reserve (in the liver) {33}.

In young children, severe vitamin A deficiency is still a signficant cause of blindness, especially in developing countries {33}. An excess of retinol is considered toxic in large doses. The level of intake of vitamin A that is potentially hazardous varies with the form and source of the vitamin. The usual intakes of preformed vitamin A in food are not likely to be high

enough to cause toxicity, however, potent pharmaceutical sources are known to be potentially hazardous with respect to vitamin A toxicity {33}. A condition related to high intake of vitamin A known as hypervitaminosis A can cause a number of adverse effects such as severe headaches, nausea, insomnia, bone and joint pain, liver toxicity and birth defects {33}. While vitamin A is toxic in large doses, provitamins A are not toxic, although a condition known as hypercarotenemia (yellowing of the skin) may result when very high intakes of β-carotene are ingested {33}. This condition, however, is considered benign and disappears immediately after high-dose consumption of β-carotene is discontinued.

1.1.2 B Vitamins

Vitamins B_1 (*thiamine*), B_2 (*riboflavin*), B_3 (*niacin*), and B_6 (*pyridoxine*) are required to carry out many biochemical functions in the body. Thiamine is necessary for nerve cell function and myocardial function {32}. *Beriberi* is a common condition associated with thiamine deficiency, usually associated with excess alcohol consumption {33}. Riboflavin is involved in many aspects of energy and protein metabolism, while pyridoxine is necessary for the functional integrity of the central nervous system {32}. Niacin is involved in various metabolic processes (carbohydrate, fat and protein metabolism), as well as tissue and cell respiration. Deficiency of niacin results in pellagra, which is characterized by dementia, dermatitis and insomnia {32}.

Vitamin B_{12} (*cobalamin*) and folic acid (or *folate*, which is also a B vitamin) are necessary for the formation of red blood cells and DNA {32}. Vitamin B_{12} is found in meats, eggs and milk, and deficiency of this vitamin may arise from prolonged inadequate dietary intake, or

because of a specific disease that interferes with its absorption, such as pernicious anemia {33}. Folate is found in many fruits and vegetables, as well as dairy products, meat and liver. The recommended daily intake of folate is 210 micrograms (mcg) per day for men and 165 mcg/day for women and this requirement increases for women during preganancy {33}.

1.1.3 Vitamin C

Humans are only one of a few species that cannot synthesize vitamin C and are therefore dependent on dietary supplementation of this vitamin {33}. Vitamin C prevents scurvy, and is also required for tissue growth and repair, bone formation, and wound healing {32}. Commercially, vitamin C is sold as ascorbic acid, sodium ascorbate, calcium ascorbate, potassium ascorbate and ascorbate-acetate (fat-soluble vitamin C) and can be measured in blood samples (serum, plasma), or urine samples {16}.

Along with scurvy, other deficiency symptoms include loose teeth and unhealthy gums {32}. The recommended intake for the adult male, 19 years and older, is 60 mg/day, and for the adult female, 45 mg/day {33}. Further correction to allow for individual variation results in an overall recommended intake for adults of 100 mg/day {32}. With regard to high doses of vitamin C, known side effects include rebound scurvy, nausea, vomiting and kidney damage (oxalate stones) {32}. The 'rebound effect' or rebound scurvy can occur after sudden cessation of large doses of vitamin C, which causes the level of vitamin C circulating in the body to drop to well below normal.

Exposure to toxic pollutants like tobacco cigarettes increases vitamin C requirements as

smoking causes an increase in oxidative stress and decreased intake and absorption of vitamin C {3}. Individuals who smoke are extremely vulnerable to oxidant stress and susceptible to peroxidative damage to lipids, a cancer-causing chemical process due to the decrease in ascorbate levels {5}. In Canada, smoking 20 cigarettes per day or less is associated with a 25% lower (plasma) ascorbate level than that of nonsmokers, and with 20 or more cigaretted per day, a 40% reduction is observed {33}. A recent study indicated that smokers would need to consume at least 200 mg per day of vitamin C in order to achieve the same serum vitamin C concentrations as nonsmokers consuming the RDA of 60 mg per day {3}. Even individuals who do not smoke but are merely exposed to tobacco smoke in the environment may have an increased need for vitamin C. One recent study found that nonsmoking women with extensive exposure to environmental tobacco smoke had significantly lower plasma vitamin C levels than otherwise similar women who were rarely exposed to tobacco smoke {3}.

1.1.4 Vitamin E

Vitamin E consists of two families, the *tocopherols* and the *tocotrienols*. α-tocopherol is has the greatest nutritional importance and biological activity, however the intake of γ-tocopherol is approximately three times that of α-tocopherol as it is found in widely used food sources such as vegetable oils and margarines {33}. The tocotrienols are found in certain cereal grains, but are thought to have little nutritional significance. Synthetic forms of α-tocopherol are used as food supplements, the most common being dl-α-tocopherol {33}.

Vitamin E is considered a 'free radical scavenger'. There is some evidence that some

free radicals may be 'cancer-causing' molecules created by chemical processes in the body (i.e. oxidation), by UV exposure from the sun, smoking, some pollutants and many other factors. It has been suggested that vitamin E may inhibit the formation of tissue damage caused by this free radical activity {33}.

The requirement for vitamin E is determined primarily by the concentration of polyunsaturated fatty acids (PUFA) in the tissues. The recommended intake for adults is 10 IU per day, however it varies significantly with the PUFA content of the diet {33}. This relationship is due to the metabolic role of the vitamin as an antioxidant {33}. Deficiency symptoms are considered rare in the Canadian population, however risk increases with malabsorption such as in cystic fibrosis {32}. In terms of an optimal dose of vitamin E, no clear level of daily intake for humans has been firmly decided. One researcher contends that daily intake of an active tocopherol in excess of 100 to 300 units is regarded as a 'megadose' {12}. Until toxicity symptoms have been established, mega-doses of vitamin E should be used with caution.

1.1.5 Minerals

Selenium in small quantities is considered vital to human health. It is found in fish, meat and nuts, however the content in food varies with the amount of selenium in the soil {32}. This is mainly because of regional differences, as grains grown in the prairie provinces of Canada are higher in selenium than those grown in the eastern provinces {33}. Selenium is obtained from a typical Canadain diet in an average amount of 200 mcg/day, however individual intake may vary widely, and some diets may contain only half of this amount {33}. Where the selenium

content of the soil and of the food supply is exceptionally low, the dietary intake may be as low as 25 mcg/day {33}.

There is a mutual involvement between selenium and vitamin E in the metabolism of peroxides and free radicals {33}. Like vitamin E, selenium protects cells from oxidative damage and many of the effects produced by vitamin E deficiency can be reversed or prevented by selenium {33}.

Zinc is an important mineral that is required for fertility, growth, and maintenance of the skin {9}. Good sources of zinc include red meats, shellfish, grain cereals and dairy products {33}. Zinc has also been shown to be immunocompetent and thus supplementation with zinc as well as other nutrients may have a favourable influence on the immune system {14}.

Some of the symptoms of chronic zinc deficiency include growth retardation, anorexia, skin lesions and hair loss {32}. Zinc deficiency symptoms also include impaired wound healing, abnormal immune response and for genetic diseases like sickle cell disease {33}. In general, zinc should be ingested with caution. It is considered relatively toxic in large quantities and can cause nausea, vomiting, diarrhea and abdominal pain {33}. Therefore, although zinc is an essential nutrient for life, great care should be taken in ingesting the proper dose.

Coenzyme Q_{10}, or, *ubiquinone* plays an essential role in mitochondria, the energy-producing part of cells. Since it is found in almost every cell of the body, it is currently being studied as an adjunct treatment in individuals with heart disease and cancer.

1.2 Antioxidants as Adjuncts to Standard Cancer Treatment

Conventional cancer treatment involves a combination of surgery (usually lumpectomy or a modified radical), chemotherapy, hormonal therapy and radiation therapy. Chemotherapy refers to *cytotoxic*, or cancer killing drugs. Chemotherapeutic agents are used to induce cytotoxicity of malignant cells {11}. However, normal cells and malignant cells are very similar and therefore cytotoxic chemotherapy administered to cancerous (neoplastic) cells can also be toxic to normal cells. Additional adverse effects experienced during chemotherapy may include cardiac toxicity, low white blood cell count (higher risk of infection), hair loss, nausea, vomiting, and occasionally even second cancers {11, 17}. Heavy doses of anti-cancer drugs can damage the immune system, which can lead to infectious complications - one of the major causes of death in cancer patients {11}. Hormonal therapy may be appropriate for patients, especially elderly, who have hormonally sensitive cancers (estrogen and/or progesterone receptor positive tumours), since there is usually minimal toxicity {17}. Radiotherapy is usually mandated with a lumpectomy, but may be used in the adjuvant setting in combination with chemotherapy and/or hormonal therapy for patients with more advanced breast cancer. New treatments that decrease the cytotoxic effect of standard therapy on normal cells but not on cancer cells would improve the efficacy of standard treatment. This is one reason why nontoxic compounds such as antioxidants are being studied as possible adjuncts.

Studies on the effects of antioxidant vitamins individually or in combination (synergistically) with standard therapy are as yet only in progress. One study found that the lethal effects of

vitamin C on Ehrlich ascites carcinoma cells in vitro are synergistically enhanced by 3-amino-1,2,4-triazole (ATA) {2}. Vincenzo Noto et al. (1989) investigated the effects of sodium ascorbate and vitamin K3 administered separately or in combination in vitro on human breast cancer cells MCF-7. Vitamin K3 when administered on its own was essentially harmless to the breast cancer cells, however the synergistic killing effects when combined with vitamin C was better than the effects of each vitamin administered separately {10}. An enhanced antitumour effect was shown to exist when combining certain retinoids (tretinoin, isotretinoin, acitretin) with other agents such as chemotherapy drugs or cytokines (IFNα). All three retinoids inhibited mammary carcinoma cell lines, however, the cytokine did not have an inhibitory effect on the MCF-7 cells alone. A significant reduction in proliferation of the breast cancer cells was seen when the retinoids were combined with IFNα - more profound than the retinoids administered alone {6}.

1.3 Cholesterol

Cholesterol is an important substance needed for the proper functioning of all of the cells in the human body. High-density lipoprotein (HDL) is considered the 'good' cholesterol, and low-density lipoprotein (LDL) the 'bad' cholesterol. It has been suggested that dietary antioxidants such as vitamins C, A and β-carotene may be protective against oxidized LDLs {8}. Reports on the relation between vitamin deficiency and cholesterol levels in the blood show that increased cancer incidence is related to a decrease in HDL cholesterol and an increase in plasma levels of

LDL cholesterol {4}. Vitamin C is known to act as a regulator for cholesterol levels. Studies in experimental animals and humans have shown a direct relationship between vitamin C status and HDL & LDL cholesterol. An increase in vitamin C intake raises the HDL level, and higher than RDA intakes of vitamin C have been associated with lowered cardiovascular disease risk including increases in HDL, and decreases in LDL oxidation, blood pressure and cardiovascular mortality {4}.

1.4 Hormone Replacement Therapy and Breast Cancer Risk

The risk factors for breast cancer include personal history of breast disease, radiation exposure and age. Early menarche, late menopause and late or no child bearing may also increase the chances of developing breast cancer because of the increased lifetime exposure to estrogen {1}. Estrogen is one of the female sex hormones produced by the ovaries (along with progesterone). Although our body naturally produces estrogen, we can also ingest it from external sources such as birth control and through Postmenopausal Hormone Replacement Therapy (HRT). Some cancer cells are estrogen-sensitive therefore high levels of estrogen may induce or promote tumour formation in the breast {8}. Although estrogen may promote the growth of cancer cells, it also helps bones maintain their bone mass density by assisting in the absorption of calcium {18}. After menopause, natural estrogen levels decline and this estrogen deficiency will lead to a loss of calcium absorbed in the bones and thus overall lower bone mass density in the body. Therefore, low levels of estrogen may decrease the risk of breast cancer formation but increase the risk of

the bone-thinning disease, osteoporosis. Women take Hormone Replacement Therapy (HRT) to combat menopausal symptoms that include osteoporosis, as well as heart disease, another serious long-term effect of estrogen deficiency. HRT is the administration of the female hormones estrogen and progesterone; the administration of estrogen alone is called estrogen replacement therapy, or ERT. Therefore a trade-off exists as HRT appears not only to maintain bone mass, but also decrease the risk of two severe diseases, however HRT is related to a higher risk of breast cancer which is just as serious. Just as breast tissue is highly responsive to estrogen, bone is estrogen-sensitive tissue. Hormone Replacement Therapy is thought to help prevent the devastating effects of heart disease and osteoporosis, but hormone treatment during menopause still remains controversial because of its possible association with breast cancer. However, as heart disease is the leading killer in postmenopausal women and claims six times as many lives as breast cancer, the known benefits of HRT are likely to outweigh the possible increased risk of breast cancer for most women in the face of uncertainty {19}.

1.5 Conclusions

The recent interest in nutritional therapies has prompted researchers to take a closer look at the current RDA levels of micronutrients. The use of unconventional nutritional therapies is due primarily to its host-nontoxic compounds which are primarily health-promoting, however, without proper supervision and knowledge of dose requirements, this therapy can be dangerous. Therefore the use of vitamin supplementation needs to be administered with great caution, as it

is important to ensure that supplements of antioxidants are completely safe and free from side effect. Active ongoing research to answer this question is required to reveal if more diseases can be treated with clinical nutrition by providing adequate amounts of vitamins and other nutrients in conjunction with well-established conventional therapies.

Chapter 2

Description of Data set I

2.1 Background Information

The medical management of breast cancer, from the time it is first suspected (suspicious lump or abnormal mammogram) up until a final treatment strategy is in place, follows a number of steps including *diagnosis, primary surgery, staging*, and if necessary, *adjuvant therapy* {17}.

Definitive diagnosis is standardly made by a pathologist after examination of breast tissue, although tissue may frequently be removed by fine needle aspiration in a physician's office, or by core biopsy under radiologic guidance with a stereo tactic procedure, rather than by open biopsy. If cancer is present, the information provided by the pathologist's summmary report is used to plan an effective treatment strategy for the patient.

For those tumours that are identified as malignant, the extent or stage of the cancer is considered before deciding on the treatment plan. An 'early' stage cancer is one that is small and confined to the breast, 'regional' stage implies involvement of the axillary lymph nodes, and 'advanced' stage cancer implies that is has spread to other parts of the body (metastatic).

There are several staging systems for classifying the extent or stage of breast cancer. Clinical staging is often done using the Stage I, II, II, IV system (corresponds to Cancer Control Agency of British Colombia (CCABC) Pretreatment Clinical Staging in BC), and the TNM Clinical Classification, while the pathologist uses the TNM Pathological Staging System; criteria from pathological staging are usually employed in research investigations. Other prognostic factors, or diagnostic variables are also used to help predict the future behaviour of an invasive cancer. Some key prognostic variables are nodal status (positive lymph node involvement), tumour size, lymphatic or blood vessel invasion (positive invasion of blood, nerves, veins of the breast tumour at diagnosis), extension of the tumour to the skin (margins) and the grade of the cancer {17}. Other important factors include age at diagnosis, date of diagnosis, treatment regime, and estrogen and/or progesterone receptor status.

2.1.1 Pathologist's Report

The important features in the pathology report are tumour size, nodal status, lymphatic invasion, tumour type, estrogen and/or progesterone receptor status, grade, and margins {17}. In terms of the size of the cancer, the larger the cancer, the more likely it will spread to regional or distant sites. The TNM pathological staging categories for tumour size are:

-1	0	1	2	3	4
In situ	no tumour	small (<2cm)	med. 2 to 5 cm	large (>5cm)	Extended

In situ is the least severe stage. It refers to cancer that is still within the milk ducts or glands (lobules) of the breast. Cancers that are not invasive are more likely to be cured. Categories

0 through 4 represent the size of the tumour from smallest to largest with a 4 representing a patient who is very sick. An *extended* classification refers to a patient who is most sick. The patient has a *locally advanced* form of cancer which is most likely invading or growing into the chest wall, muscles or bones, no matter what the size of the cancer. At this stage there is a fairly high risk of recurrence and spread {17}. In research investigations, in situ cancer patients and stage'0' patients may not standardly be included in studies with invasive cancer. For patients with stage '4' disease, the tumours may never fully be removed in primary therapy, in which case these patients would not be eligible for a study of disease-free period. It should be noted that a cancer that spreads to other parts of the body is still a *breast* cancer and treatment should reflect this {17}.

The number of lymph nodes infected by cancer is one of the most important factors in determining the future behaviour of a breast cancer. The main groups of lymph nodes are *axillary* (above armpit), *supraclavicular* (above collarbone), *internal mammary* (along sternum or breast bone), and the *apex of axilla* (between axillary and supraclavicular) lymph nodes. The risk of cancer spreading increases according to the number of axillary lymph nodes found to have cancer. If the patient is treated with surgery alone, the chance of the cancer spreading and reappearing within 5 to 10 years is: 30% to 50% if 1 to 3 lymph nodes are infected, 50 to 75% if 4 to 9 are infected, and 75% or more for 10 or greater nodes involved {17}. Breast cancer cells can enter the *lymphatic* vessels that drain the breast and may be carried to the lymph nodes where they can settle and grow {17}. If cancer cells are found in the lymph channels or blood

vessels in the breast, the prognosis is similar to having 1 to 3 cancerous lymph nodes {17}.

The type of tumour refers to the behaviour of the tumour. A tumour can be in situ, benign (noncancerous), or malignant. Malignant cancers represent the most serious types of cancer as they have most likely spread past the primary site with larger tumours likely to have systemic spread to other parts of the body. If the cancer is large enough, a biochemical test is done to measure the estrogen receptors (ER) and/or progesterone receptors (PgR) levels, otherwise the diagnostic slides may be used for immunohistochemical assessment. ER/PgR status is a predictor of prognosis and response to hormone therapy {17}. Current methods of ER and PgR assessment include biochemical, immunohistochemical, as well as high technology methods (image, flow, and laser scanning) and molecular biology {21}. Depending on the results of the biochemical test, the ER status can be divided into two categories, estrogen receptor positive, and estrogen receptor negative. For a biochemical test, when there are less than 10 to 15 fmol/mg protein ER (ER -ive), this implies there are very few estrogen receptors on the cell and the tumour will most likely not respond well to cancer hormonal therapy. A person with ER negative status may have worse prognosis than a patient who is ER positive {27}. One method of scoring for the immonohistochemical is intensity of stain 0, +1, +2, or +3, with a 0 representing virtually no ER content, (ER -ive), and a +3 for high ER content (ER +ive) {17}. This is the method employed by the BC Cancer Agency. The higher the estrogen receptor level, the more likely the cancer will respond to anti-estrogen hormone drugs like tamoxifen.

The grading of the disease represents the aggressiveness of the cancer (severity of disease). The International Classification Codes (ICD-0) grading system was standardly used in the province of British Colombia during the study period. Cells viewed under the microscope by the pathologist, are classified as I, II, or III which are coded based on how well the tumour ressembles the normal tissue from which it arose.

Grading classifications, after microscopic examination by a pathologist,are as follows:

Grade	Degree of differentiation
I	'Well' differentiated (tumour most ressembles normal tissue)
II	'Moderately" (moderately well, intermediate)
III	'Poor' differentiation (tumour mostly neoplastic)

A high grade of cancer is usually associated with a greater incidence of axillary lymph node invasion, which may lead to a higher chance of recurrence and worse survival {17}. It is also important to note how close the cancer is to the walls of the specimen (margins). If the cancer is near the edges of the lump removed by the surgeon rather than centred within a block of normal tissue, the surgeon may have left some of the cancer behind in the breast.

2.1.2 Treatment Plan

The cancer is then treated with a combination of surgery, radiation, chemotherapy and hormone therapy that will reduce the chance of further spread of the cancer and increase the chances of being cured. When the cancer is small and confined to one area, initial treatment is usually surgery (lumpectomy), and then adjuvant radiation to prevent recurrence. Chemotherapy and hormone therapy are systemic, or whole-body treatments, and are used for invasive cancers to reach regional and distant sites. Chemotherapy is the adjuvant therapy used for hormonally

insensitive tumours, or for advanced/high risk cancer; it may be used by itself or in combination therapy (adjuvant chemotherapy with or without hormonal therapy, with or without radiotherapy). An individualized approach is used for each patient as different treatment options are more effective for different types of breast cancer and for different types of women {17}. The goal is to give the best combination of treatment strategies for that individual's unique situation.

2.1.3 International Classification Codes: Histological Classification of Tumours (ICD-0)

Breast cancer can be separated into in situ cancer and invasive cancer. There are many types of non-invasive (in situ) and invasive breast cancers, all with potentially different levels of severity. Studies usually evaluate separately patients with in situ and invasive breast cancers. The main types of *in situ* breast cancers are ductal carcinoma in situ (DCIS) and lobular carcinomas in situ (LCIS). The different types of *invasive* cancer include ductal carcinoma, lobular carcinoma, and inflammatory carcinoma. Lobular carcinomas are believed to have less chance of spreading than ductal carcinomas because the milk glands (lobules) are located deeper into the breast and are therefore considered more 'contained'. Therefore it is believed that ductal cancer may be worse than lobular cancer. Inflammatory cancers were included here if patients had definitive surgery dates, (i.e., the primary surgical treatment was thought to be successful).

Cancers from the ducts and glands of the breast are called *adenocarcinomas.* Cancers that begin in other areas of the breast include sarcomas, lymphomas, and cystosarcoma phyllodes {17}. Sarcomas, lymphomas, and cystosarcoma phyllodes are very rare types of invasive breast

cancers, and their behaviour and treatment are considered completely different from the other breast cancers {17}. The 'International Classification of Diseases for Oncology', or (ICD-0), provides all codes that represent breast cancer as primary site (three digit *topography* codes, **C**00.0 - **C**80.9, see Table 2.1 and Table 2.2), as well as codes that provide detailed information of the neoplastic cell type and tumour type (five digit *morphological* codes, **M**-0000/0). Non-metastatic, first primary, unilateral invasive breast cancer patients will be considered in this study

SKIN C44	C44.5	skin of breast, chest wall, axilla (underarm area)
TISSUES OF THORAX C49	C49.3	soft tissues of axilla, chest, chestwall, thorax, infraclavicular region
	C49.9	soft tissues, lymphatic, subcutaneous tissue
BREAST C50 (excludes skin C44.5)	C50.0	nipple, areola
	C50.1	central portion of breast
	C50.2	upper-inner quadrant
	C50.3	lower-inner quadrant
	C50.4	upper-outer quadrant
	C50.5	lower-outer quadrant
	C50.6	axillary tail of breast
	C50.8	overlapping lesion of breast (inner, lower, mid, outer, upper)
	C50.9	breast/mammary gland, NOS

Table 2.1: Topography (Site) Codes for the Breast, Skin and Thorax

OTHER/ILL-DEFINED SITES C76	C76.1	thorax, axilla, chest, chest wall, thoracic wall
LYMPH NODES C77	C77.0	supraclavicular: head, face and neck
	C77.1	intrathoracic/thoracic
	C77.3	axilla or arm: axillary brachial, infraclavicular subclavicular, pectoral
	C77.9	lymph node

Table 2.2: Topography Codes for Other Sites Related to the Breast

Histological/Morphological Codes that relate to the **breast:** the first four digits represent the specific neoplastic cell type of a given tumour, and the fifth digit represents the type of tumour (M-0000/0). The behaviour codes for the last digit of the morphological classification are:

0	benign neoplasm
1	uncertain behaviour (uncertain whether malignant or benign tumour)
2	in situ neoplasms
3	malignant neoplasms (primary cancer - in the original site, first detected)
6	malignant neoplasms (secondary - metastatic cancer)
9	uncertain whether primary or secondary site (not normally used)

Table 2.3: Behaviour codes of the tumour

Morphology index, numerical (M-0000/0):

Note: The code 'NOS', not otherwise specified, is when pathologist reports a cell type with no adjective or an adjective not in the list.

805-808 squamous cell neoplasms

8050/2 papillary carcinoma in situ

814-838 adenomas and adenocarcinomas

8140/2 adenocarcinoma in situ, NOS

8140/3 adenocarcinoma, NOS

8140/6 adenocarcinoma (mets), NOS

8201/3 cribriform carcinoma

8211/3 tubular adenocarcinoma/tubular carcinoma (subtype of invasive **ductal**)

8230/3 solid carcinoma, NOS (subtype of **ductal**)

844-849 mucinous, colloid neoplams - subtypes of invasive **ductal**

8480/3 mucinous (adeno) carcinoma, colloid (adeno) carcinoma

850-854 ductal, lobular and medullary neoplasms

8500/2 intraductal (adeno) carcinoma, noninfiltrating, NOS

8500/3 ductal (adeno) carcinoma, infiltrating ductal (adeno) carcinoma, NOS

8501/2 comedocarcinoma, noninfiltrating

8501/3 comedocarcinoma, NOS

8510/3 medullary (adeno) carcinoma, NOS

8520/2 lobular carcinoma in situ, lobular carcinoma, noninfiltrating

8520/3 lobular (adeno) carcinoma, infiltrating lobular carcinoma, NOS

8521/3 infiltrating ductular carcinoma

8530/3 inflammatory (adeno) carcinoma

Paget's disease (854's)

8540/3 Paget's disease, mammary/Paget's disease of the breast

8541/3 Paget's disease + infiltrating ductal carcinoma of breast

8542/3 Paget's disease, extramammary (except Paget's disease of bone)

8543/3 Paget's disease + intraductal carcinoma of breast

2.2 Description of the Database and Data set 1 variables

The Breast Cancer Outcome Evaluation Unit (BCOU) at the BC Cancer Agency (BCCA) in Vancouver, is made up of a variety of breast cancer specialists (oncologists, pathologists, statisticians) who conduct research as well as access and maintain databases for analyses. Donna Mates, an Outcome Data Analyst at the BCCA, accessed the breast cancer database to generate suitable matched controls for each Hoffer patient from the BC Cancer Agency Information System (CAIS). Caroline Speers is a Data Analyst at the BCCA and was available to answer questions pertaining to the data and medical terminology. The CAIS is an electronic database of cancer information that contains demographic, treatment, prognostic and outcome information on approximately 10,000 referred breast cancer cases. Every effort was made by Donna Mates to match patients as closely as possible with respect to these characteristics.

Dr. Abram Hoffer is a practicing physician who specializes in orthomolecular medicine in Victoria, British Colombia where he has treated more than 900 cancer patients with vitamins and minerals. The data provided by Dr. Hoffer's office included his entire population of breast cancer patients that had been given the vitamin and mineral treatment. Each patient had been given varying amounts of vitamin C, β-carotene, niacin (B_3), selenium, zinc and Coenzyme Q_{10}, depending on the patient's disease status. This treatment information as well

as the following information was provided by Dr. Hoffer's office: Hoffer ID number, name, date of birth, date of diagnosis of breast cancer, conventional treatment given (surgery, radiation, chemotherapy), whether the patient started taking the vitamins and minerals (compliance), date of first Hoffer visit, the number of years alive after starting Hoffer's treatment and date of death where applicable.

Donna Mates provided a detailed data dictionary of Data set I for all variables used in the study. The Hoffer patients were matched to controls according to the following criteria (in this order):

- *agedx* (within ± 2 years), which is the age at diagnosis. Calculated in the warehouse tables by subtracting birth date from diagnosis date (divide by 365). If day has been estimated for the birthday then subtract 15 days from the total days old to estimate the middle of the month. If month has been estimated for the birthday, then subtract 182 days from the total days, to estimate the middle of the year.

- *dxyear*(within ± 1 year), the year of diagnosis taken from the diagnosis date (*dxdate*). 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*, the TNM pathological M stage. This variable is an indicator of (distant) metastatic disease where a 1 denotes distant metastasis (includes metastasis to supra-clavicular lymph nodes), 0 represents no distant metastasis and 9 denotes a missing value (presence of distant metastasis cannot be assessed). If pathological is missing or cannot assess then the clinical M stage is entered in this variable.

- *staget*, TNM pathological T stage. This variable represents the categorical size of tumour. The information is coded from TNM pathological stage (*stagept*), if missing then coded from size of lesion (*sizecat*), if still missing then coded from TNM clinical T stage (*stageclt*). The categories are -1 for carcinoma in-situ: DCIS or LCIS, or Paget's disease of the nipple with no tumour (note: Paget's disease associated with tumour is classified according to the size of the tumour), 0 indicates no evidence of primary tumour, 1 for a tumour 2 cm or less in greatest dimension, 2 for a tumour more than 2 cm but not more than 5 cm in greatest dimension, 3 for a tumour more than 5 cm in greatest dimension, and 4 represents a tumour of any size with direct extension to chest wall or skin (note: chest wall includes ribs, intercostal muscles and serratus anterior muscle but not pectoral muscle).

- *stagepn*, 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. The categories are: -1 for cannot assess, or no axillary

dissection, 0 if there is no evidence of invasion of regional nodes, 1 when there is evidence of invasion of movable homolateral axillary lymph nodes, 2 when there is evidence of invasion of homolateral axillary lymph nodes fixed to one another or to other structures, 3 denotes evidence of invasion of homolateral internal mammary lymph nodes and 9 for an unknown or missing value.

- *bccasr*, 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. This is a 'yes/no' indicator variable.

- *bccard*, indicates if the patient had radiation treatment. This may refer to either initial treatment or subsequent treatment (yes/no).

- *bccach*, indicates if the patient received chemotherapy treatment where a 1 indicates initial chemotherapy given as part of the initial treatment plan, (includes continuous chemotherapy given when one treatment fails and another is given), a 2 represents subsequent chemotherapy given for residual disease or at relapse for recurrent or metastatic disease, a 3 denotes chemotherapy given as part of initial treatment and further subsequent chemotherapy given for residual disease or at relapse for recurrent or metastatic disease. Also, N represents no chemotherapy given and a Y indicates that chemotherapy was given but could not differentiate between inital or subsequent therapy (for non-referred cases). For this variable it was noted in the data dictionary that chemotherapy includes hormone

therapy (*bccahr*, below).

- *bccahr*, indicates whether or not the patient received hormone therapy. 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 denotes subsequent hormone therapy given for residual disease or at relapse for recurrent or metastatic disease, 3 for hormone therapy given as both initial and subsequent treatment, N for no hormone therapy given, and Y for hormone therapy given but cannot differentiate between initial or subsequent treatment. Hormone treatment includes: hormones given to inhibit tumour growth, and drugs given that will decrease natural hormone production within a patient in order to inhibit the tumour growth. It is also noted that hormone therapy is included with chemotherapy for non-referred cases. Also, hormones given for replacement purposes are not considered to be hormone treatment, (e.g., pituitary, thyroid, gyne), and hormones given for alleviation of symptoms are not considered hormone treatment (e.g., steroids given in brain patients to reduce swelling of the brain). In addition, note that NSABP Trial B21, B23, or B24 and NCIC MA 12, (double-blind study in which tamoxifen or placebo is given), coded as a 1 (initial treatment) and note is entered on *note* variable.

- *dxgrade*, designates the histopathological degree of dedifferentiation 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. The coding is as follows: 1 for well differentiated, fairly well differentiated (low grade), 2 for moderately/partially differentiated, moderately well differentiated, 3 for poorly differentiated (high grade), and 9 for unknown, grade or differentiation undetermined, not stated. Notes: Nuclear/histologic/cytologic grade or degree of differentiation is based on the final pathological diagnosis. Nuclear Grade takes precedence over histologic/cytologic grade. Type of grade is reported in variable *dxgradet*. The highest of the nuclear/histologic grade if multiple comments/specimens is coded (e.g. grade 2-3 = grade 3). Invasive grade is coded when there is a discrepancy between invasive and insitu. When grade is reported as 2/3, grade 2 is entered. If surgery was done after chemotherapy, hormone or radiation therapy, the grade of the Fine Needle Aspirate is coded. If a patient was diagnosed with breast cancer from another site, (i.e. the 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). If the front of the path review states moderately differentiated and the back of the review has a 2+, grade = 2. If the front of the path review states poorly differentiated and the back of the review has a 2+, grade = 3

- *dxer*, estrogen receptor status. Estrogen receptor status recoded from Provincial ER Report or Immunohistochemical stain (*er* for biochemical test or *immuno* for immunohistochemical staining). A 1 indicates ER negative status (1-14 units for *er* or 'Negative' immunostain), 2 represent ER positive (15+ units or, 'low positive' or 'moderately posi-

tive' or 'strongly positive' immunostain), and 9 for a missing, unknown or 'not sufficient quantity'. Nov 23/93 – not recorded for benign cases.

- *dxlvn*, indicates lymphatic, vascular or neural invasion. A 1 indicates positive invasion for either lymphatics or veins or nerves or any combination, a 0 denotes no invasion of lymphatics or veins or nerves, and a 9 for not applicable, no pathology or unknown status. Collected separately in 3 SPSS variables effective Jan 1/96 - *dxlymph*, *dxveins*, *dxnerves*. This information has been recoded into this variable, *dxlvn*. Effective Nov 23/93 - not recorded for benign cases.

The event and outcome variables considered are:

- *dthcsurv*, number of days from *dxdate* to death from breast cancer. If no death from breast cancer, then death any cause or last contact date with patient is recorded. Use with event variable *dthcevt*, an event variable that indicates death from breast cancer. Coding for the event variable is 1 for a death from breast cancer, 0 for no death from breast cancer and -1 denotes a patient who was lost to follow-up.

- *syssurv*, number of days from *dxdate* to systemic relapse (that is, any regional or distant relapse or death from breast cancer) {21}. If no regional or distant relapse or death from breast cancer then number of days to last contact or death from any cause reported. Use with event variable *sysevent*, an event variable that indicates systemic relapse where a 1 denotes a systemic relapse, a 0 indicates no systemic relapse,a and a -1 for lost to follow-up.

Other important variables considered in the analysis include:

- *treated*, an indicator variable with a 1 denoting a Hoffer patient and a 0 for a control patient.
- *pair*, pair number given to each treated patient and her corresponding matched control.
- *id2*, the identification number for a treated patient (Hoffer file patient number).
- *idtran*, unique identification number given to each patient by the BCCA Cancer agency.
- *day2hof*, number of days from *dxdate* to date Hoffer first seen (ie, to date treatment was first started).
- *dxdate*, the earliest date of the class IV/ V cytology, positive pathology, or autopsy report. If none of these confirms the malignancy/condition, the date of the first positive lab is used. If there is no positive lab, the date diagnosis confirmed by a physician clinically. The date of positive radiology is used for non-referred cases when no other information is available. Note: the only time a clinical diagnosis date would be recorded as the *dxdate* (when there is subsequent histologic confirmation), is if the patient begins treatment prior to the pathologic/cytologic diagnosis. That is, the patient has been clinically diagnosed with breast cancer by a physician, then begins treatment, then has an Fine Needle Aspirate {28}. In this case, the diagnosis date would be the date of clinical diagnosis, but this occurs very rarely. Most diagnosis dates are the first highly suspicious or positive tissue/cell results.

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

- *dxposnod*, categorical variable recoded from *posnod*. A 1 indicates positive lymph node involvement (N_+), a 0 indicates no lymph nodes infected (N_-), a -1 denotes no axillary dissection (N_x) or cannot assess, and 9 for unknown. Patients who are coded $N_x = -1$ are usually older breast cancer patients. For many elderly patients a delayed axillary dissection is offered to spare these patients who have a substantially higher risk of dying from other causes, the discomfort and complications of this surgery {22}. If *posnod* is missing, then coded from *stagepn*.

- *sizeles*, indicates the actual size of the lesion (in cm). If the lesion is greater than 9.9 cm, 9.9 is entered and the actual size of lesion is recorded in the *note* variable. Size 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 operations, 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.

- *hist1 (hist2, hist3)*, cell and tumour type, ICD-0 (M-0000/0). 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. The first four digits of the code describe the cell type and the fifth digit describes the behaviour code (*behvr*). The ICD-O Second Edition is consistently used for all diagnosis coding for patients having a diagnosis 01/Jan/92 to the present. Prior to that, either 1st or 2nd edition was used. Up to 3 histologies are entered (*hist1, hist2, hist3*) in the order they are listed on the pathology report. Infiltrating duct carcinoma and lobular carcinoma 8522/3 effective Jan 92 (previously 2 codes used 8500/3 and 8520/3). Infiltrating duct carcinoma mentioned with tubular carcinoma is coded as 8211/3 (tubular carcinoma) effective June 88. See ICD-O manual or on-line datadictionary. The codes for the variable *behvr*, (type of tumour) are the same as those given in the previous subsection: International Classification Codes.

- *histcat*, categorical variable for histology created from *hist1, hist2* and *hist3*. A coding of 1 denotes ductal, 2 denotes lobular, and 3 represents NOS. Tables 2.4 through 2.6 are frequency tables of D. Mates categorization of *hist1, hist2* and *hist3*. Note: histological

codes 85212 and 85213 should be coded as a 1 (not 2) (C. Speers, 2000), and it is unknown if *histcat* was coded only from *hist1*, or a combination of *hist1, hist2* and *hist3.*

HIST1 * HISTCAT Crosstabulation

Count

		HISTCAT			
		1	2	3	Total
HIST1	81403	1			1
	82113	5			5
	82302			1	1
	84803	5			5
	85002	2			2
	85003	129			129
	85012	3			3
	85013			6	6
	85103	1			1
	85203		15		15
	85213		7		7
	85303	1			1
Total		147	22	7	176

Table 2.4: Frequency Table of *hist1* by *histcat* (ductal, lobular, other)

HIST2 * Categorical Histology Crosstabulation

Count

		Categorical Histology			Total
		Ductal	Lobular	Other	
HIST2		119	18	5	142
	82012	5			5
	82302	3			3
	85002	3		1	4
	85003		2		2
	85012	15		1	16
	85202	2	1		3
	85203		1		1
Total		147	22	7	176

Table 2.5: Frequency Table of *hist2* by *histcat* (ductal, lobular, other)

HIST3 * Categorical Histology Crosstabulation

Count

		Categorical Histology			Total
		Ductal	Lobular	Other	
HIST3		142	21	7	170
	80502	1			1
	82012	1			1
	82302	2			2
	85012	1			1
	85212		1		1
Total		147	22	7	176

Table 2.6: Frequency Table of *hist3* by *histcat* (ductal, lobular, other)

- *site*, point of origin of the primary cancer, ICD-0 (**C**00.0). An International Classification of Diseases for Oncology Second Edition (ICD-O) code which indicates the point of

origin from which the primary cancer has arisen. Primary site is determined from the following sources: mammogram, staging diagram, surgeon's consultation report/breast questionnaire, and/or operation report/chart notes. Specifically for site codes for the breast: C500 nipple, C501 central portion of breast, C502 upper-inner quadrant of breast, C503 lower-inner quadrant breast, C504 upper-outer quadrant of breast, C505 lower-outer quadrant of breast, C506 axillary tail of breast, C508 breast: C5081 inner breast, C5082 outer breast, C5083 upper breast, C5084 lower breast, C5086 multiple/multifocal breast primaries; C509 Breast, NOS (see Table 2.1 section 2.1.3).

- *srmhist1 (srmhist2 - 4)*, site of relapse, ICD-0 (C00.0). For each relapse record, the International Classification of Diseases for Oncology, Second Edition (ICD-O) or Systemized Nomenclature of Medicine (SNOMED) code is recorded which indicates: i) site of recurrence, metastases, referred multiple skins, referred multiple lymph nodal regions, or ii) site of diagnosis when specimen is not taken from the primary site or is non-diagnostic, or iii) primary site is unknown, or iv) progressive disease (histology code).

- *srmdate1 (srmdate2 - 4)*, date of relapse according to ICD-0. The earliest date the relapse was confirmed for each relapse record. May be clinical, radiological, laboratory, cytological or pathological (MM DD YY).

- *srmdes1 (srmdes2 - 4)*, type of relapse (local, regional, or distant) according to ICD-0. 'A' represents a regional relapse, (a recurrence of tumour witthin the ipsilateral axillary,

supraclavicular or internal mammary nodes), 'B' denotes a distant relapse, (a recurrence of tumour in organs beyond the confines of the breast chest wall or regional lymph nodes), 'D' for NOS, 'L' denotes a local relapse, (a recurrence of tumour within the ipsilateral breast or chest wall), 'M' multiple (not used for breast), 'P' progressive (not used for breast), 'R' Met/EXT/Rec - Loc/Reg (not used for breast), and 'S' for site of Diagnosis when tissue (includes pathological, cytological, or laboratory specimen - as of August 1992 admit date) is taken from a secondary site ONLY to confirm malignancy. Notes: supraclavicular nodes at time of diagnosis are entered as distant. Only first distant metastasis is recorded. This variable, *srmdes*, is recorded for each relapse record.

2.3 Data Collection: Linking Hoffer Patients to CAIS and Finding Matched Controls

Donna Mates was given a list of 271 of Dr. Hoffer's breast cancer patients. The Hoffer data was linked to CAIS to retrieve medical records for all of Hoffer's patients from the BCCA database. A total of 256 of these cases were also treated at the BCCA and were considered in the study. There were some discrepancies between Hoffer's information and the information in CAIS for the following Hoffer patients which were used in the final analysis (by *id2*): 460, 502, 592, 606, 608, 641, 665, 747, 782, 866.

Patients with cancer in both breasts (bilateral cancer), were counted twice in the dataset. There were bilateral patients reported in both CAIS and Dr. Hoffer's medical records that were considered twice in the data set (by *id2*): 159, 285,668, 688, 736, 743, 897. There were some

discrepancies between bilateral patients recorded in CAIS and those provided by Dr. Hoffer: CAIS showed bilateral breast cancer for 11 patients, but Hoffer did not have this information recorded (by *id2*): 21, 129, 241, 254, 385, 418, 434, 481, 523, 758, 965), 4 patients in which Dr. Hoffer reported bilateral cancer but CAIS has only one cancer recorded (35, 274, 457, 660). Due to these deficiencies in the data as well as the fact that it would be extremely difficult to find a matched control with the same bilateral breast cancer characteristics, matched controls for bilateral patients were not used in the final analysis (note the 4 patients with Hoffer ID numbers 35, 274, 457, 660, that were not considered bilateral by CAIS, were still matched when possible) {27}. Matched controls were also excluded if they had previous cancer except non-melanoma skin cancer, in situ or benign disease, and one matched control was dropped (replaced) because she was a patient of a certain physician known to also prescribe large dose vitamins {27}. D. Mates also reported that 28 Hoffer patients could not be matched because there was no electronic record with respect to staging on CAIS, and she provided the Hoffer identitification number for those Hoffer cases that had matching problems but were still matched. Those treated patients with problems matching into CAIS that were included in the final survival analysis are 608, 683, 725 (for patients 608 and 725 there were problems matching TNM N stage, and for 683 the problem was matching conventional treatment received).

2.4 Data Cleaning

Only data for patients who were diagnosed between the years 1989 to 1996 was considered, as that data are most complete and accurate, leaving 131 pairs (D. Mates, 1999). It should also be noted that the BC breast screening program was started in 1988 and that there was a change in recommended screening practices in 1996/1997. There were 18 pairs for whom the information for *dthcsurv* was missing and therefore those pairs were dropped from the analysis (by *pair*): 57, 68, 81, 94, 122, 130, 134, 145, 156, 173, 183, 191, 193, 200, 204, 212, 230, 243. There was also one male Hoffer patient (*id2* is 728) that was dropped along with its matched control (*pair* 184),and one bilateral patient (*id2* is 241) and her pair match were dropped. There were a select group of Hoffer patients who elected not to take the vitamins and minerals (noncompliant patients). Five noncompliant patients were found in the data and removed (*pair* 35, 79, 90, 143, 241). This leaves 106 pairs in the dataset.

Those patients with 'advanced stage' breast cancer, or metastatic disease, should be considered separately from other breast cancer patients as they are much sicker and are therefore not comparable with nonmetastatic patients {27}. For the purposes of this study, those patients with metastatic disease (*stagem*=1) were dropped from the analysis as well as any pairs with unknown status for this variable (*stagem*=9). There were 18 pairs in total that were dropped, 16 of which had missing values: 32, 54, 91, 99, 102, 106, 123, 127, 163, 164, 176, 180, 185, 206,209, 213, 216, 232. Therefore 88 pairs were considered in total for the final analysis using Data set I.

2.5 Vitamin and Mineral Treatment Plan

Figures 2.1 through 2.6 shows the frequencies and dosages for the 88 treated patients for each vitamin and mineral administered by Dr. Hoffer. In most cases, patients were given amounts far greater than the RDA (ie, for β-carotene, selenium, vitamin C and zinc). Table 2.7 shows the recommended daily intakes (RDA) for these vitamins and minerals {33}. Also, it is important to note that most of the patients who were given the treatment did not start the Hoffer regime at the time of diagnosis. Figure 2.7 shows a histogram of the variable *day2hof*, the number of days from diagnosis to the date the treatment was first started.

	RDA
β - carotene	3300 (IU) per day
Selenium	200 (mcg) per day
Vitamin C	100 (mg) per day
Zinc	8 - 9 (mg) per day

Table 2.7: Recommended Daily Intakes

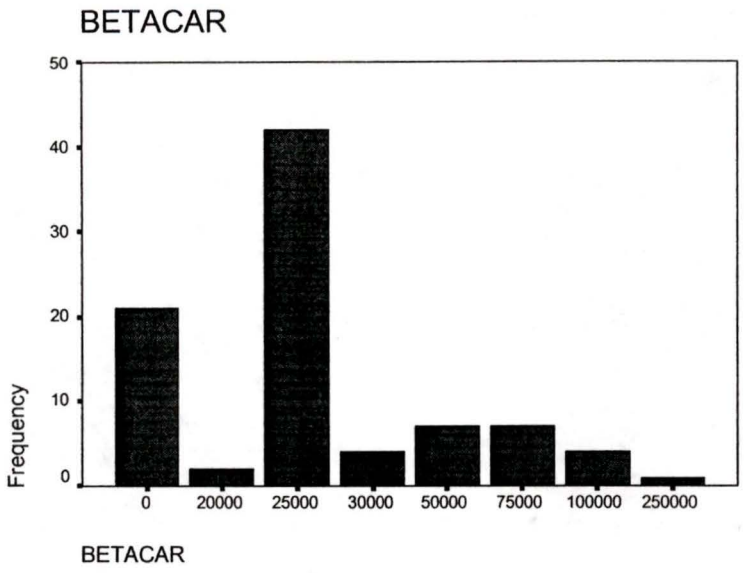

Figure 2.1: β-carotene doses (IU) per day for 88 Hoffer patients

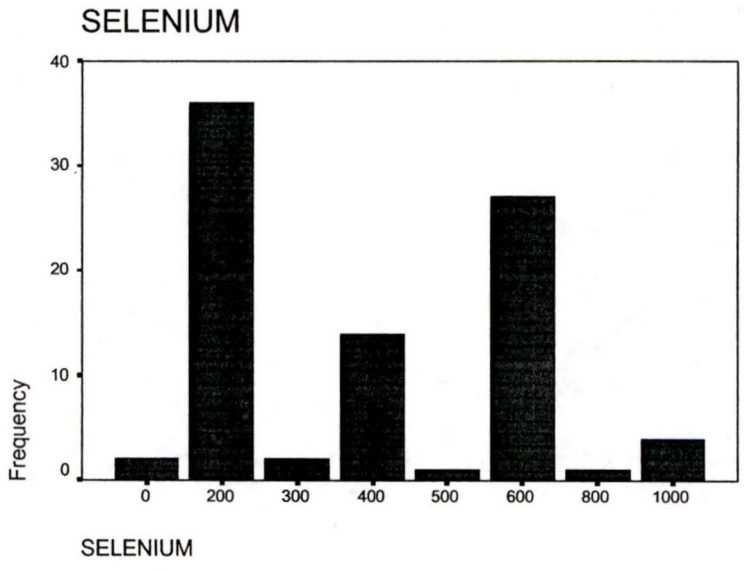

Figure 2.2: Selenium doses (in mcg) per day for 88 Hoffer patients

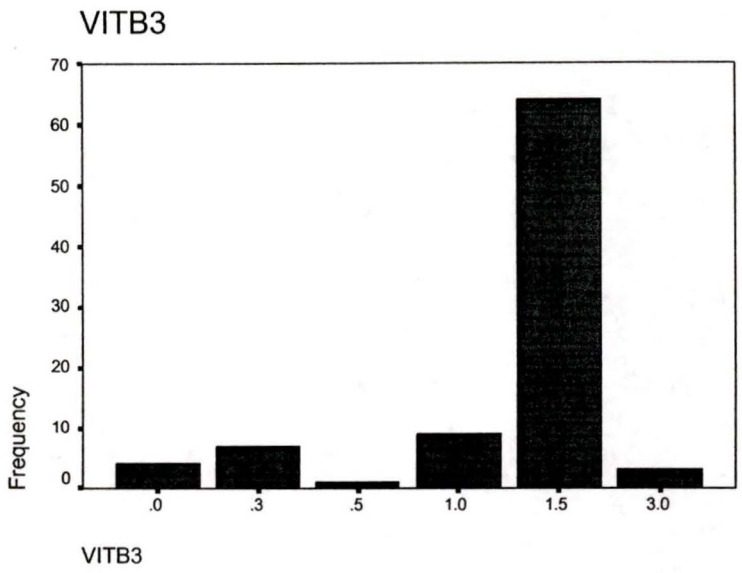

Figure 2.3: Vitamin B_3 doses (grams) per day for 88 Hoffer patients patients

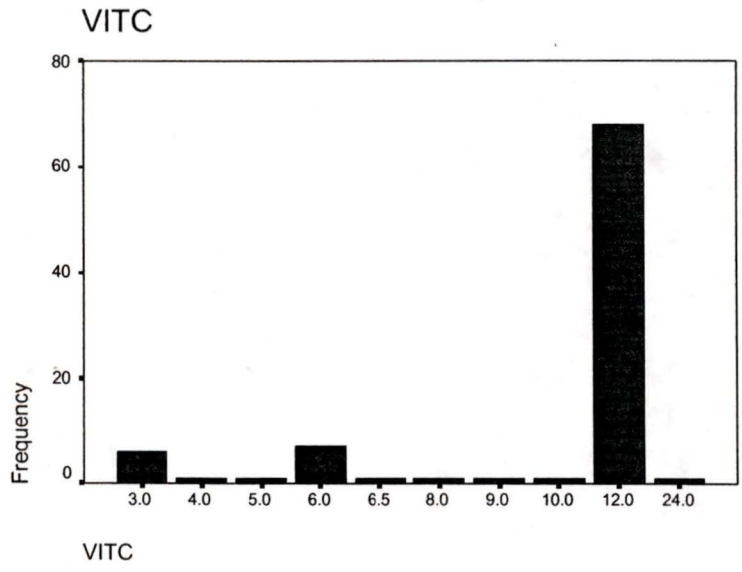

Figure 2.4: Vitamin C doses (grams) per day for 88 Hoffer patients

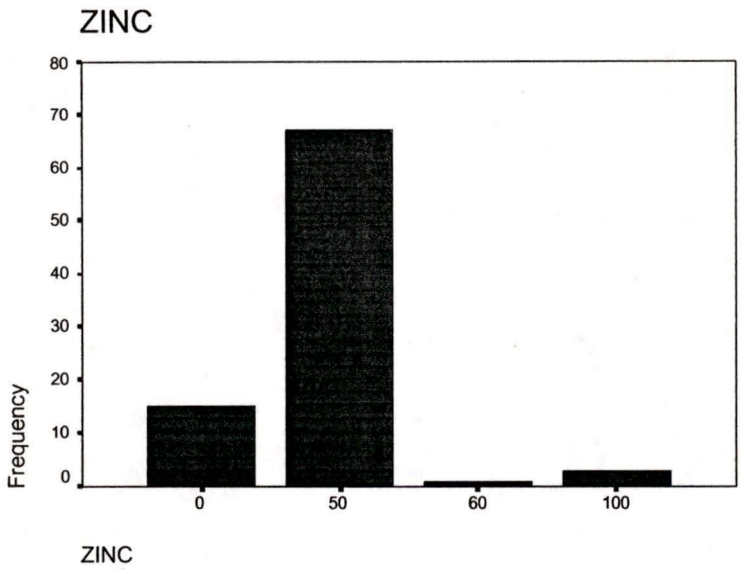

Figure 2.5: Zinc doses (in mg) per day for 88 Hoffer patients

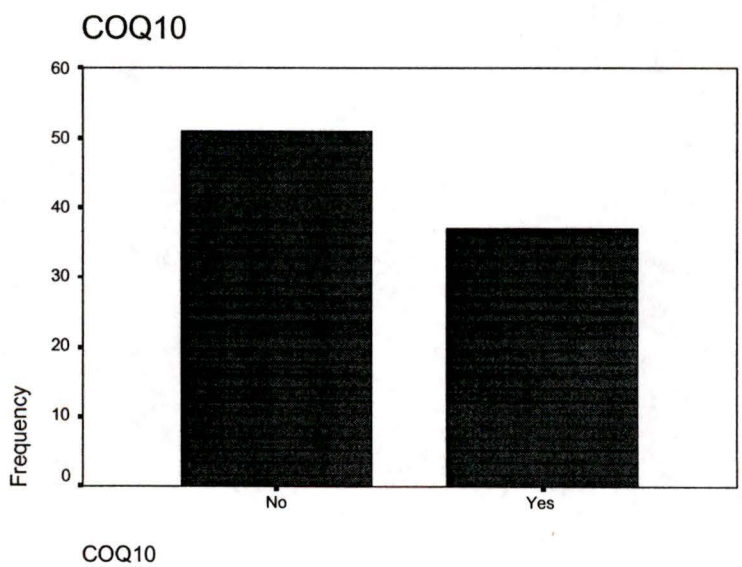

Table 2.6: Frequency Table for Coenzyme Q_{10} for 88 Hoffer patients

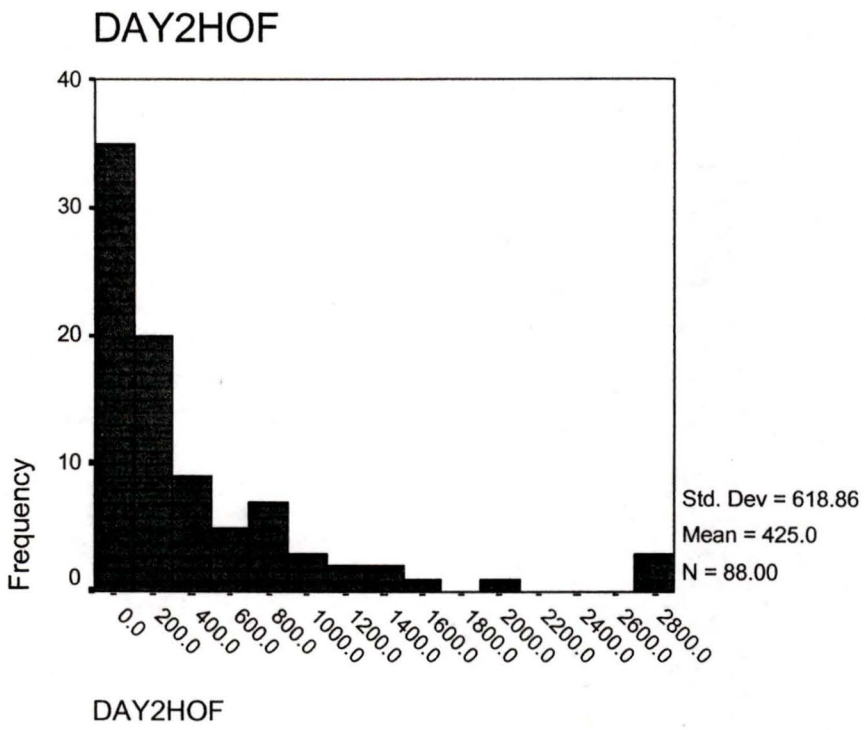

Figure 2.7: Histogram of *day2hof*

Chapter 3

Preliminary Analysis

3.1 Data Quality for Matched Variables

The matching criteria for the diagnosis, staging and treatment variables is given in section 2.2. Crosstabulations between the two treatment groups as well as Paired Difference Tests (for continuous variables), and Pearson's Chi-square Tests of Homogeneity (for categorical variables) were performed to check the quality of the matching. In order to accurately assess treatment effect, the two groups need to be as similar as possible in all other ways. We assume that the differences between the paired subjects for continuous variables is approximately normal for our sample size. The crosstabulations of categorical variables should show no off diagonal elements if the matching is perfect. For categorical variables that are not perfectly matched, chi-square tests to test for two identical multinomial populations were computed. Note: the treated group in some tables and figures may be represented as a 'case'.

3.1.1 Paired Difference Tests for Continuous Variables

The null hypothesis is that of zero mean paired difference. If the paired sample mean difference is approximately normally distributed, then the test statistic has a t-distribution under the null distribution. Table 3.1 shows the results for *agedx* and *dxyear*. The P-values are large for each test, therefore the data is matched well with respect to these two variables.

Paired Differences	Mean	Std. Err Mean	t	df	Sig. (2-tailed)
agedx/control - treated	0.17	0.16	1.098	87	0.275
dxyear/control - treated	-3.41E-02	6.94E-02	-0.491	87	0.625

Table 3.1: Paired Samples t-tests for *agedx* and *dxyear*

3.1.2 Crosstabulations for Treated Patients vs. Control Patients

Crossbulations by treated patients versus control patients are given in Tables 3.2 to 3.11 for variables *dxyear*, *staget*, *stagepn*, *bccasr*, *bccard*, *bccach*, *bccahr*, *dxgrade*, *dxer*, and *dxlvn* (*stagem* all 0). The treated group are displayed in the rows, and the control subjects are in the columns (denoted with a 'C' at the end of the variable name). The year of diagnosis was successfully matched within one year for the two groups, and the variable *bccasr* was perfectly.matched for all 88 pairs. The variable *staget* was also matched perfectly however one pair had missing values. All other variables matched by D. Mates had off diagonal elements therefore chi-square tests of the these variables for the two groups are performed and reported in section 3.1.3. The variables *bccach* and *bccahr* were changed to two categories (yes=1,2,3 or Y, and no=N).

Diagnosis Year * DXYEARC Crosstabulation

Count

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

Table 3.2: Crosstab for *dxyear*, treated by control

T-stage * STAGETC Crosstabulation

Count

		STAGETC					
		-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 3.3: Crosstab for *staget*,treated by control

Pathological N-stage * STAGEPNC Crosstabulation

Count

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

Table 3.4: Crosstab for *stagepn*, treated by control

BCCASR * BCCASRC Crosstabulation

Count

		BCCASRC		
		N	Y	Total
BCCASR	N	1		1
	Y		87	87
Total		1	87	88

Table 3.5: Crosstab for *bccasr*, treated by control

BCCARD * BCCARDC Crosstabulation

Count

		BCCARDC		
		N	Y	Total
BCCARD	N	20	4	24
	Y		64	64
Total		20	68	88

Table 3.6: Crosstab for *bccard*, treated by control

BCCACH * BCCACHC Crosstabulation

Count

		BCCACHC		
		N	Y	Total
BCCACH	N	35	2	37
	Y	5	46	51
Total		40	48	88

Table 3.7: Crosstab for *bccach*, treated by control

BCCAHR * BCCAHRC Crosstabulation

Count

		BCCAHRC		
		N	Y	Total
BCCAHR	N	49	5	54
	Y	8	26	34
Total		57	31	88

Table 3.8: Crosstab for *bccahr*, treated by control

Tumor Grade * DXGRADEC Crosstabulation

Count

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

Table 3.9: Crosstab for *dxgrade*, treated by control

Estrogen Receptor * DXERC Crosstabulation

Count

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

Table 3.10: Crosstab for *dxer*, treated by control

Invasive Ivn * DXLVNC Crosstabulation

Count

		DXLVNC		Total
		0	1	
Invasive Ivn	Negative	34	7	41
	Positive	5	32	37
Total		39	39	78

Table 3.11: Crosstab for *dxlvn*, treated by control

3.1.3 Pearson's Chi-Square Test for Categorical Variables

The null hypothesis for the Pearson's Chi-Square Test is that of homogeneity of a diagnosis variables for the treated and control group. This test is done for the categorical factors *stagepn*, *bccard*, *bccach*, *bccahr*, *dxgrade*, *dxer*, and *dxlvn* which were not perfectly matched according to the crosstabulations (see section 3.1.2). The results are shown in Tables 3.12 to 3.25. For *stagepn*, categories 1 and 2 were grouped (presence of axillary and fixed nodal mets) to represent 'presence of nodal mets'. There were no patients with category 3 (internal mammary lymph

nodes). Table 3.13 shows that the P-value is large (P=0.42) for the chi-sqared test for *stagepn*, therefore there is no evidence against the null hypothesis that the two groups are similar with respect to stage of regional lymph node metastasis. The variables *bccach* and *bccahr* recoded in the previous section have expected counts greater than 5 due to the change and the results in Tables 3.17 and 3.19 show that the two groups appear to be similar with respect to these two variables. For all other factors there is no evidence against the null hypothesis of homogeneity. Therefore, the two groups appears to be adequately matched overall with respect to the variables in the matching criteria for this study. It should be noted, however, that tumour grade, ER status and lymphatic invasion status all have missing values (see Table 3.26).

case or control * Pathological N-stage Crosstabulation

			Pathological N-stage			
			"No axil. disect"	"No Nodal Mets"	"Axillary Nodal Mets"	Total
case or control	matched control	Count	3	38	47	88
		Expected Count	5.0	36.5	46.5	88.0
	Hoffer patient	Count	7	35	46	88
		Expected Count	5.0	36.5	46.5	88.0
Total		Count	10	73	93	176
		Expected Count	10.0	73.0	93.0	176.0

Table 3.12: Frequency table with expected counts for *stagepn*

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.734[a]	2	.420
Likelihood Ratio	1.780	2	.411
Linear-by-Linear Association	.389	1	.533
N of Valid Cases	176		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.00.

Table 3.13: Chi-square test for *stagepn*

case or control * BCCARD Crosstabulation

			BCCARD		
			N	Y	Total
case or control	matched control	Count	20	68	88
		Expected Count	22.0	66.0	88.0
	Hoffer patient	Count	24	64	88
		Expected Count	22.0	66.0	88.0
Total		Count	44	132	176
		Expected Count	44.0	132.0	176.0

Table 3.14: Frequency Table with expected counts for *bccard*

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 3.15: Chi-square test for *bccard*

case or control * BCCACH Crosstabulation

			BCCACH		
			N	Y	Total
case or control	matched control	Count	40	48	88
		Expected Count	38.5	49.5	88.0
	Hoffer patient	Count	37	51	88
		Expected Count	38.5	49.5	88.0
Total		Count	77	99	176
		Expected Count	77.0	99.0	176.0

Table 3.16: Frequency Table with expected counts for *bccach*

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 3.17: Chi-square test for *bccach*

case or control * BCCAHR Crosstabulation

			BCCAHR		Total
			N	Y	
case or control	matched control	Count	57	31	88
		Expected Count	55.5	32.5	88.0
	Hoffer patient	Count	54	34	88
		Expected Count	55.5	32.5	88.0
Total		Count	111	65	176
		Expected Count	111.0	65.0	176.0

Table 3.18: Frequency Table with expected counts for *bccahr*

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 3.19: Chi-square test for *bccahr*

Crosstab

			Tumor Grade			
			Well Differentiated	Moderately Differentiated	Poorly Differentiated	Total
case or control	matched control	Count	8	33	43	84
		Expected Count	10.8	30.9	42.3	84.0
	Hoffer patient	Count	12	24	35	71
		Expected Count	9.2	26.1	35.7	71.0
Total		Count	20	57	78	155
		Expected Count	20.0	57.0	78.0	155.0

Table 3.20: Frequency Table with expected counts for *dxgrade*

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 3.21: Chi-square test for *dxgrade*

Crosstab

			Estrogen Receptor		
			Negative	Positive	Total
case or control	matched control	Count	25	48	73
		Expected Count	28.7	44.3	73.0
	Hoffer patient	Count	32	40	72
		Expected Count	28.3	43.7	72.0
Total		Count	57	88	145
		Expected Count	57.0	88.0	145.0

Table 3.22: Frequency Table with expected counts for *dxer*

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 3.23: Chi-square test for *dxer*

Crosstab

			Invasive lvn		
			Negative	Positive	Total
case or control	matched control	Count	41	43	84
		Expected Count	42.8	41.2	84.0
	Hoffer patient	Count	42	37	79
		Expected Count	40.2	38.8	79.0
Total		Count	83	80	163
		Expected Count	83.0	80.0	163.0

Table 3.24: Frequency Table with expected counts for *dxlvn*

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 3.25: Chi-square test for *dxlvn*

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
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 * Invasive lvn	163	92.6%	13	7.4%	176	100.0%

Table 3.26: Summary of Missing Observations for *dxgrade*, *dxer*, *dxlvn*

3.2 General Data Checks for Dataset 1

To check if the two treatment groups are similar with respect to type of breast cancer (*histcat*; ductal, lobular or other), a Pearson's Chi-square Test of Homogeneity was done (note that the

expected counts were less than 5 for 2 categories and therefore the P-values are not accurate). The frequency table and chi-square test are given in Tables 3.27 and 3.28. There is strong evidence against the null hypothesis for *histcat* (P=0.023), therefore the two groups are different with respect to this variable. Note that there is some question as to the accuracy of this variable (see the description in section 2.2).

Table 3.29 shows a table of the *site* categories for the 88 pairs. All patients have primary site codes that relate to the breast. In addition, the primary or secondary classification information is given in the last digit of the histology codes (*behvr*) for malignant cancer (see Table 2.3). If *behvr*=3, the patient has a malignant tumour in the primary site, if *behvr*=6, the malignant tumour is a secondary tumour, and if *behvr*=2, it is nonmalignant (in situ). It is important that the two groups are matched well with respect to malignancy of the tumour. In situ patients should not be matched with malignant patients in any of the pairs, and all subjects in the study who have malignant cancer should have last digit equal to 3 (no secondary cancers). Table 3.30 shows that there are only 6 in situ patients in the data set, and the remaining have malignant tumours in the breast. This table also verifies that there are no secondary malignant cancer patients in the data set, and that the two groups are matched perfectly (no chi-square test needed).

It is preferable in regression analysis to use continuous prognostic factors rather than categorizing them, as the categorization of continuous factors may limit the investigation of factor effects and interactions {23}. The continuous variables *sizeles* (tumour size in cm) and *posnod*

(actual number of cancerous lymph nodes) could be used in the model to represent tumour size and nodal status, two important prognostic factors related to the risk of death from breast cancer. Table 3.31 shows the status of the missing values of the categorical (*staget*, *dxposnod*) versus the continuous variables for tumour size and nodal status. There are no missing values for the categorical variables however there are for the continuous ones. Also, to test if the continuous variables are matched well, paired-t tests for *sizeles* and *posnod* are given in Table 3.32. We have evidence against the null hypothesis of equal means for *posnod* and only moderate evidence that the two treatment groups are the same with respect to *sizeles*. The treated group has a larger mean value for *posnod*, and smaller mean for *sizeles*. However, a crosstabulation of *staget* (Table 3.3) shows the groups were perfectly matched with respect to this variable, and Tables 3.33 and 3.34 show the results of a chi-squared test of homogeneity for *dxposnod*, indicating this variable appears to be homogeneous with respect to the two treatment groups (P=0.728, however 2 cells with expected counts equal to 4).

case or control * Categorical Histology Crosstabulation

			Categorical 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 3.27: Frequency Table for *histcat*

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 3.28: Chi-square test for *histcat*

SITE * case or control Crosstabulation

Count

		case or control		
		matched control	Hoffer patient	Total
SITE	C501	8	7	15
	C502	10	8	18
	C503	7	4	11
	C504	37	40	77
	C505	10	8	18
	C506	1	1	2
	C508		2	2
	C5081		1	1
	C5082	6	2	8
	C5083	5	3	8
	C5084	3	2	5
	C5086	1	7	8
	C509		3	3
Total		88	88	176

Table 3.29: *Site* codes for the two treatment groups

BEHVR * case or control Crosstabulation

			case or control		
			matched control	Hoffer patient	Total
BEHVR	2	Count	3	3	6
		Expected Count	3.0	3.0	6.0
	3	Count	85	85	170
		Expected Count	85.0	85.0	170.0
Total		Count	88	88	176
		Expected Count	88.0	88.0	176.0

Table 3.30: Frequency Table for *behvr*

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
case or control * # of +ive Nodes Examined (posnod)	163	92.6%	13	7.4%	176	100.0%
case or control * Tumor size in cm (sizeles)	164	93.2%	12	6.8%	176	100.0%
case or control * Nodal Status (dxposnod)	176	100.0%	0	.0%	176	100.0%
case or control * T-stage (staget)	176	100.0%	0	.0%	176	100.0%

Table 3.31: Summary of Missing Observations for *posnod, sizeles, dxposnod, staget*

Paired Differences	Mean	Std. Err Mean	t	df	Sig. (2-tailed)
posnod/control - treated	-1.2533	0.4769	-2.628	74	0.010
sizeles/control - treated	0.191	0.118	1.625	77	0.108

Table 3.32: Paired Samples t-tests for *posnod* and *sizeles*

case or control * Nodal Status Crosstabulation

			Nodal Status			
			No Axillary Dissection	No pos nodes	Pos nodes	Total
case or control	matched control	Count	3	38	47	88
		Expected Count	4.0	36.5	47.5	88.0
	Hoffer patient	Count	5	35	48	88
		Expected Count	4.0	36.5	47.5	88.0
Total		Count	8	73	95	176
		Expected Count	8.0	73.0	95.0	176.0

Table 3.33: Frequency Table for *dxposnod*

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.634[a]	2	.728
Likelihood Ratio	.639	2	.726
Linear-by-Linear Association	.017	1	.898
N of Valid Cases	176		

a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 4.00.

Table 3.34: Chi-square test for *dxposnod*

The goal of this study is to determine if the vitamin therapy has a positive or negative effect on survival. The outcome variable of primary interest is disease-free survival from breast cancer (*dthcsurv*, with event indicator *dthcevt*). To see any preliminary differences in survival time between the treatment groups, the survival curves for the two groups can be plotted. The standard estimator of the survival function for right-censored data is the Product-Limit estimator

proposed by Kaplan and Meier {26}. If events occur at D distinct times $x_1 < x_2 < \ldots < x_D$, and at x_i there are d_i possible events, and Y_i individuals at risk just prior to x_i, the Product-Limit estimator is represented by,

$$\hat{S}(x) = \begin{cases} 1 & if\ x < x_1 \\ \Pi_{x_i \leq x} \left[1 - \frac{d_i}{Y_i}\right] & if\ x_1 \leq x \end{cases},$$

where $\frac{d_i}{Y_i}$ is an estimate of the hazard rate, $\lambda(x)$, and $\hat{S}(x)$ is a decreasing step function with jumps at the event times {26}. Note that no distributional assumptions are required for estimating the survival function. The log-rank test can also be performed to test if the treatments are equally effective. To test $H_o : S_1(x) = S_0(x)$, where a 1 represents a treated patient, and a 0 for a matched control, the test statistic for for the log-rank test when there are only two groups is:

$$Z = \frac{\sum_{i=1}^{D} [d_{i1} - Y_{i1}\left(\frac{d_i}{Y_i}\right)]}{\sqrt{\sum_{i=1} \frac{Y_{i1}}{Y_i}\left(1 - \frac{Y_{i1}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i}}, \text{ for } x_1 < x_2 < \cdots x_D,\ j = 1, 2$$

The test statistic is approximately normally distributed for large samples when H_o is true {26}. Figure 3.1 shows the survival curves for the two groups with the controls showing a more favourable survival from breast cancer. For data set I the log-rank test results is:

dthcsurv	Statistic	df	Significance
Log Rank	12.89	1	.0003

Test Statistic for Equality of Survival Distributions for TREATED

Therefore there is evidence against the null hypothesis, the two groups are different with respect to survival.

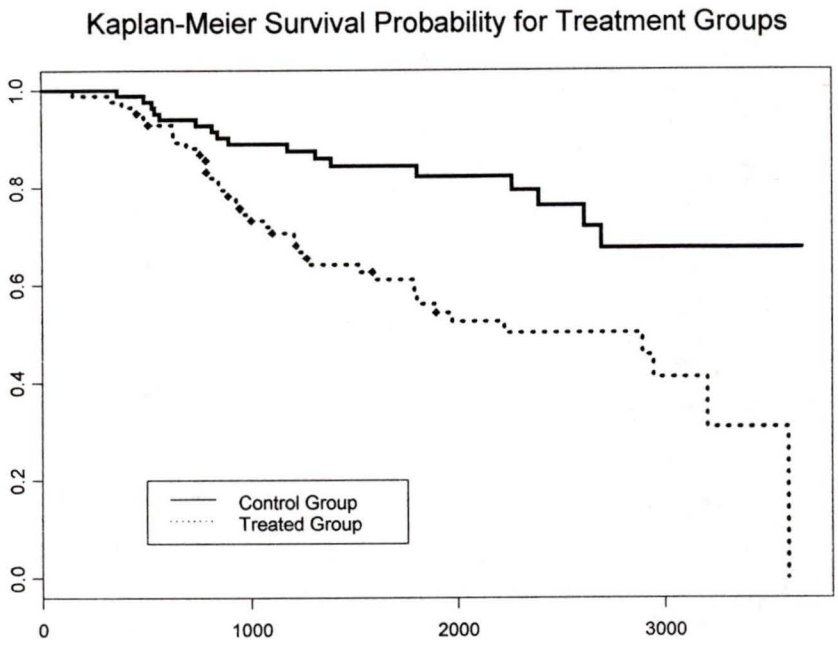

Figure 3.1: Survival Curves for *treated*

3.2.1 Final Covariates Considered for Survival Modelling

The final pairs all have *stagem* equal to zero, therefore this variable in not included. It would be preferable to use *sizeles*, and *posnod* instead of *staget* and *dxposnod*, however due to missing values in the data set, we chose to use the categorical variables. Therefore, *staget* was used instead of *sizeles*, and *dxposnod* instead of *posnod*. *Stagepn* was dropped as this information is contained in *dxposnod* (nodal status). A continuous variable for ER status was not given in the data set, therefore *dxer* was considered although there are many missing values. In practice, this is often the case as the ER tests are not always performed. No variables were given in the

data set to assess the effects of progesterone on survival. In terms of treatment variables to consider, *bccasr* is not used as all pairs received surgery (primary therapy) but one (this pair would be dropped from the final analysis due to missing values). *Bccahr* was not used because it is included in *bccach* (see data dictionary), and *bccard* is not considered because radiaton therapy is usually given in combination with surgery, not on its own. This leaves *bccach* as the only treatment variable to be considered. The variable *histcat* was used instead of *hist1* and *behvr*, and *dxyear* instead of *dxdate* (the variables *id2*, *idtrans*, *site*, *srmhist1*, *srmdate*, *srmdes* are not factors to be considered for covariate effects on survival).

Therefore the final variables considered in the model are *treated*, *agedx*, *dxyear*, *dxlvn*, *dxgrade*, *dxposnod*, *staget*, *histcat*, *dxer*, *bccach*. The variable *pair* was also included in the modelling section. For the purposes of this study, we are interested in whether the patients are alive or dead (or relapsed), as well as determining the risk of these possible outcomes for Hoffer's patients as compared to the controls, while adjusting for relevant confounding variables.

Chapter 4

Concepts of Parametric Survival Analysis

4.1 Introduction to Survival Analysis

Let X represent the time until some specified event for an individual. Assume X is a *continuous* random variable where $X > 0$ for time. If there are n subjects in the study, $j = 1, \ldots, n$, assume all subjects are identical with probability density function $f(x)$, survivor function $S(x)$, and hazard function $\lambda(x)$. The distribution of X can be characterized by the above three functions. The survival function $S(x)$ is defined as $P(X > x)$, or $1 - F(x)$, where $F(x)$ is the cumulative density function *(cdf)* of X. In terms of survival analysis, $F(x)$ is the 'risk' or probability of experiencing the event between time 0 and x. The probability density function *(pdf)*, $f(x)$, represents the unconditional probability the event occurs in the next small interval of time,

$$f(x) = \lim_{dx \to 0} \frac{P\{X \in [x, x + dx)\}}{dx} = F'(x) = -S'(x)$$

The hazard rate, or hazard function, represents the conditional probability that the event occurs in the next small instant given the individual has survived up to time x,

$$\lambda(x) = \lim_{dx \to 0} \frac{P\{X \in [x, x+dx) \mid x \leq X\}}{dx} = \frac{f(x)}{S(x)}$$

Therefore we can obtain an instantaneous rate of disease, relapse or death, depending on the event of interest for $\lambda(x) \geq 0$. To obtain one measure for the rate of disease (or death or relapse) from time 0 to time x, the *cumulative hazard,* $\Lambda(x) = \int_0^x \lambda(u)\, du = -\ln S(x)$, can be used to accumulate the instantaneous hazards up to time x. Any one of these four functions determines the other three.

There are many possible shapes for the hazard rate and it is particularly useful in modelling survival time data. One estimator of the cumulative hazard function is the Nelson-Aalen, (N-A) estimator denoted $\hat{\Lambda}_{N-A}(x)$. If events occur at D distinct times $x_1 < x_2 < \ldots < x_D$, and at x_i there are d_i possible events, and Y_i individuals still under study just prior to x_i, the Nelson-Aalen estimator is defined up to the largest observed time on study as follows:

$$\hat{\Lambda}_{N-A}(x) = \begin{cases} 0 & \text{if } x < x_1 \\ \sum_{x_i < x} \frac{d_i}{Y_i} & \text{if } x_1 \leq x \end{cases}$$

The Nelson -Aalen estimator is an increasing step function with jumps of size $\frac{d_i}{Y_i}$ at each event time x_i. The quantity $\frac{d_i}{Y_i}$ can be used to provide crude estimates of the hazard rate $\lambda(x)$ {26}.

When using a parametric survival model, the hazard function for the proposed distribution must describe the data set in a reasonable way. Therefore an appropriate parametric hazard is

necessary that reflects the shape of the data. Plots of $\hat{\Lambda}_{N-A}(x)$ versus event times can be used to aid in the choice of a suitable parametric hazard, where the tangents to the plot of $\hat{\Lambda}_{N-A}(x)$ will give us an idea of the shape of the hazard function.

The events of interest for this study are death from breast cancer for disease-specific survival, (DSS), and regional or distant relapse for disease-free survival (DFS). The breast cancer survival data for this study involves *right-censoring.* For those subjects where the exact survival time is not known at the end of study, let C_j represent the censoring time for the jth individual. The observed censoring time tells us only that the survival time for that individual is at least C_j, and $P(X > C_j) = S(C_j)$ {25}. The data for a subject can then be represented by (T, δ), where $T = min(X, C)$ and δ is an indicator of whether the lifetime is an *exact* lifetime ($\delta = 1$) or a censored time ($\delta = 0$),

$$\delta = \begin{cases} 1 \text{ if } C \geq X \\ 0 \text{ if } C < X \end{cases}$$

Accidental deaths or deaths due to causes other than the event of interest as well as dropouts ('lost to follow-up') can occur at any time during the course of a study period. We assume these occur randomly and therefore fall under the heading of *random censoring.* A censoring time C for an individual is thus a random variable with pdf $g(c)$ and survivor function $G(c)$. A critical assumption when there is random censoring is that the distribution of the censored times is *independent* of the distribution of the observed survival times {25}.

The likelihood function for the data (T_j, δ_j), $j = 1, \ldots, n$, when there is right-censored data

can be written as

$$L \propto \Pi_{j \in D} f(x_j) \Pi_{j \in R} S(C_j)$$

where D represents the set of death times and R the set of right-censored observations {26}, or equivalently as,

$$L \propto \Pi_{j=1}^{n} [f(t_j)]^{\delta_j} [S(t_j)]^{1-\delta_j} = \Pi_{j=1}^{n} [\lambda(t_j)]^{\delta_j} [S(t_j)]$$

since the distribution of the censoring times does not depend on the distribution of the failure times or any parameters associated with the failure time distribution.

4.2 Parametric Regression Models

4.2.1 The Weibull Parametic Survival Model

Let the lifetimes in the study follow the Weibull distribution:

$$f(x) = \rho \alpha x^{\alpha - 1} \exp[-\rho x^{\alpha}], \quad \rho, \alpha > 0, \quad x > 0$$

In terms of survival, since $f(x) = \lambda(x) S(x)$,

$$\lambda(x) = \rho \alpha x^{\alpha - 1}$$

$$S(x) = \exp[-\rho x^{\alpha}]$$

The two-parameter Weibull distribution is a very flexible model for modelling survival data. It has a hazard rate, $\lambda(x)$, that is either monotone increasing (for $\alpha > 1$), decreasing (for $\alpha < 1$)

or constant (for $\alpha = 1$). When $\alpha = 1$, the Weibull distribution reduces to the exponential distribution.

Thus far, we have assumed a Weibull model for a homogeneous population with no covariates. To incorporate covariates into the model, let $\mathbf{Z}_{px1} = (z_1, z_2, \ldots, z_p)^T$ be the *risk* vector of p covariates associated with an individual. To model covariate effects, we use an approach that is similar to the Classical Linear Regression Approach. In this approach the natural logarithm of the survival time, $Y = \ln(X)$, is modelled. This is the usual transformation made in linear models to convert positive variables to observations on the real line {26}.

We assume a linear model of the form

$$Y = \mu + \gamma^T \mathbf{Z} + \sigma W$$

where $\gamma^{T=}(\gamma_1, \gamma_2, \ldots, \gamma_p)$, a vector of regression coefficients, $\mu_o = \mu + \gamma^T \mathbf{Z}$ is the log-linear location parameter, and $\sigma =$ the log linear scale parameter. The distribution for $Y = \ln(X)$ where X follows the Weibull distribution becomes

$$f_Y(y \mid \mathbf{Z}) = \frac{1}{\sigma} \exp\left\{\left[\frac{y - \mu - \gamma^T \mathbf{Z}}{\sigma}\right] - \exp\left[\frac{y - \mu - \gamma^T \mathbf{Z}}{\sigma}\right]\right\}, \quad -\infty < y < \infty$$

where $\alpha = \frac{1}{\sigma}$ and $\rho = \exp\left(\frac{-\mu_o}{\sigma}\right)$,which represents the extreme value distribution. This can be rewritten as,

$$f_Y(y \mid \mathbf{Z}) = \frac{1}{\sigma} f_W\left[\frac{y - \mu - \gamma^T \mathbf{Z}}{\sigma}\right]$$

where W has the *standard extreme value* distribution $(\mu = 0,\ \sigma = 1,\ \mathbf{Z} = \mathbf{0})$.

This model can equivalently be expressed as an *accelerated failure-time model* such that

$$S_X(x \mid \mathbf{Z}) = S_0\left[x \exp\left(-\gamma^T \mathbf{Z}\right)\right]$$

If $S_0(x)$ is the survival time for $X = \exp(Y) = \exp(\mu + \sigma W)$ when $\mathbf{Z} = \mathbf{0}$, then

$$\begin{aligned} S_X(x \mid \mathbf{Z}) &= P(X > x \mid \mathbf{Z}) \\ &= P(Y > \ln x \mid \mathbf{Z}) \\ &= P(\mu + \sigma W > \ln x - \gamma^T \mathbf{Z} \mid \mathbf{Z}) \\ &= P\left[\exp(\mu + \sigma W) > x \exp(-\gamma^T \mathbf{Z}) \mid \mathbf{Z}\right] \\ &= S_0\left[x \exp(-\gamma^T \mathbf{Z})\right] \end{aligned}$$

If the covariate vector is a scalar, with $z_1 = 1$ representing a treated patient, $z_1 = 0$ for a control, and $\theta_1 = -\gamma_1$, then $\exp(\theta_1)$ represents the *acceleration factor.* This implies that the median lifetime for a control is $\exp(\theta_1)$ times that of a treated patient {26}. That is, if $\tilde{x}_{z_1}$ and $\tilde{x}_0$ represents the median time for the treated group and control group respectively, then $S_{z_1}(\tilde{x}_{z_1} \mid z_1 = 1) = S_0\left[\tilde{x}_{z_1} \exp(\theta_1)\right] = 0.5$ and $S_0(\tilde{x}_0) = 0.5$, and therefore $\tilde{x}_0 = \tilde{x}_{z_1} \exp(\theta_1)$.

A third equivalent model for the Weibull distribution has a *proprotional hazards* interpretation on the original time scale. That is,

$$\lambda_X(x \mid \mathbf{Z}) = \lambda_0(x) \exp\left(\beta^T \mathbf{Z}\right) = \rho \alpha x^{\alpha - 1} \exp\left(\beta^T \mathbf{Z}\right) = \frac{1}{\sigma} \exp\left(\frac{-\mu}{\sigma}\right) x^{\frac{1}{\sigma} - 1} \exp\left(\beta^T \mathbf{Z}\right),$$

where $\beta = \frac{-\gamma}{\sigma}$, $\rho = \exp\left(\frac{-\mu_0}{\sigma}\right)$, and the baseline hazard is $\lambda_0(x) = \rho \alpha x^{\alpha - 1} = \frac{1}{\sigma} \exp\left(\frac{-\mu}{\sigma}\right) x^{\frac{1}{\sigma} - 1}$. A proportional hazards model impies that the ratio of the hazards for a treated patient as compared to a control is constant and the survivor functions of the two groups will not cross.

Therefore, if X represents time to death from breast cancer,

$$\frac{\lambda(x \mid z_1 = 1)}{\lambda(x \mid z_1 = 0)} = \frac{\lambda_0(x)\exp(\beta_1)}{\lambda_0(x)} = \exp(\beta_1),$$

and $\exp(\beta_1)$ is called the *relative risk* of death due to breast cancer for a treated patient ($z_1 = 1$), as compared to a control. If the relative risk is a constant that does not change over time, then the hazard rates are considered proportional {26}.

4.2.2 The Log logistic Survival Model

The log logistic distribution has the density function:

$$f(x) = \frac{\rho\alpha x^{\alpha-1}}{(1+\rho x^{\alpha})^2}$$

and in terms of survival where $f(x) = \lambda(x)S(x)$,

$$\lambda(x) = \frac{\rho\alpha x^{\alpha-1}}{1+\rho x^{\alpha}}$$

$$S(x) = \frac{1}{1+\rho x^{\alpha}}$$

The numerator of the hazard function is the same as the Weibull hazard, but the denominator causes the hazard to be hump-shaped (that is, the hazard increases initially to a maximum and then decreases to zero as time approaches infinity). If we incorporate (fixed-time) covariates into the model, in the linear model format, $Y = \ln(X)$ is modelled where

$$Y = \mu + \gamma^T \mathbf{Z} + \sigma W = \mu_o + \sigma W$$

with $\gamma^{T=}(\gamma_1,\gamma_2,\ldots,\gamma_p)$, as the vector of regression coefficients, μ_o, and σ are respectively the location and scale parameter of Y.

The distribution for Y is the *logistic* distribution, a distribution closely ressembling the normal distribution. The pdf for Y can also be written in the form,

$$f_Y(y \mid \mathbf{Z}) = \frac{1}{\sigma} f_W\left[\frac{y-\mu-\gamma^T\mathbf{Z}}{\sigma}\right] = \frac{1}{\sigma}\frac{\exp\left(\frac{y-\mu-\gamma^T\mathbf{Z}}{\sigma}\right)}{\left[1+\exp\left(\frac{y-\mu-\gamma^T\mathbf{Z}}{\sigma}\right)\right]^2}\ ,\ -\infty < y < \infty,$$

where $\alpha = \frac{1}{\sigma}$, $\rho = \exp\left(\frac{-\mu_o}{\sigma}\right)$, and $f(w)$ is the density of a *standard logistic* distribution ($\mu = 0,\ \sigma = 1,\ \mathbf{Z} = \mathbf{0}$). The hazard function for the logistic distribution is very similar to the log normal (see 4.2.3), providing a good approximation to it except in the extreme tails {25}. This model has the advantage of having simple algebraic expressions for $S(x)$ and $\lambda(x)$, and is therefore more convenient in handling censored observations than the log normal {25}.

This model can equivalently be expressed as an *accelerated failure-time model* such that

$$S_X(x \mid \mathbf{Z}) = S_0\left[x\exp\left(\theta^T\mathbf{Z}\right)\right]$$

where $S_0(x)$ is the baseline log logistic survival function of $X = \exp(\mu + \sigma W)$ when $\mathbf{Z} = 0$. If the covariate vector is a scalar with $z_1 = 1$ representing a treated patient, and $z_1 = 0$ for a control patient, the median life for a baseline patient is $\exp(\theta_1)$ that of a treated patient.

A third equivalent model has a *proportional odds* interpretation on the original time scale. The conditional survival function for an individual with exact lifetime X is given by

$$S_X(x \mid \mathbf{Z}) = \frac{1}{1+\exp\left(\frac{-\mu}{\sigma}\right)\exp(\beta^T\mathbf{Z})x^{\frac{1}{\sigma}}}$$

The odds of survival beyond time t is given by,

$$\frac{S_X(x \mid \mathbf{Z})}{1 - S_X(x \mid \mathbf{Z})} = \exp\left(-\beta^T \mathbf{Z}\right) \frac{S_0(x \mid \mathbf{Z} = 0)}{1 - S_0(x \mid \mathbf{Z} = 0)}$$

Therefore the relative odds of survival for an individual with covariate vector $\mathbf{Z}$ compared to an individual with the baseline characteristics is represented by $\exp\left(-\beta^T \mathbf{Z}\right)$. The log logistic model is the only parametric model with both a proportional odds and an accelerated failure-time representation {26}.

4.2.3 The Log Normal Survival Model

An alternate accelarated failure time model is the log normal distribution,

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu - \gamma^T \mathbf{Z}}{\sigma}\right)^2\right], \quad -\infty < x < \infty$$

and

$$S(x) = 1 - \Phi\left[\frac{\ln x - \mu - \gamma^T \mathbf{Z}}{\sigma}\right]$$

where $\Phi(.)$ is the *cdf* of a standard normal random variable. Given a set of covariates $\mathbf{Z}_{px1} = (z_1, z_2, \ldots, z_p)$, the log normal survival model

$$Y = \mu + \gamma^T \mathbf{Z} + \sigma W = \mu_o + \sigma W$$

assumes that the logarithm of the time to some event follows the usual normal regression model,

$$f_Y(y \mid \mathbf{Z}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y - \mu - \gamma^T \mathbf{Z}}{\sigma}\right)^2\right], \quad -\infty < y < \infty$$

where μ_o and σ are the location and scale parameters respectively and $f(w)$ is the *standard normal* density {26}. This model also has an *accelerated failure-time representation,* $S_X(x \mid \mathbf{Z}) = S_0\left[x \exp\left(\theta^T \mathbf{Z}\right)\right]$.

The general shape of the hazard function increases and then eventually begins declining ('hump-shaped' hazard), which is similar to the log logistic model. Studies have shown that the log normal model is an appropriate choice to model breast cancer data {23}. The log normal hazard in terms of risk of breast cancer would suggest an initially increasing hazard from time zero to some maximum, followed by an eventual decline in risk after some point in time {23}. This assumption is consistent with clinical practice as breast cancer rates tend to decrease after a certain period of time, depending on the characteristics of the patients {23}.

Likelihood for γ for the Log-linear Survival Model

The likelihood (for right-censored observations), with corresponding risk vector $\mathbf{Z}_{px1}$for the response variables Y_j $\{j = 1, \ldots, n\}$, can be written as

$$L \propto \Pi_{j=1}^{n} \left[f_Y(y_j \mid \mathbf{Z})\right]^{\delta_j} \left[S_Y(y_j \mid \mathbf{Z})\right] = \Pi_{j=1}^{n} \left[\frac{1}{\sigma} f_W\left(\frac{y_j - \mu_0}{\sigma}\right)\right]^{\delta_j} \left[S_W\left(\frac{y_j - \mu_0}{\sigma}\right)\right]$$

4.3 Residual Analysis

Graphical checks can be done to reject inappropriate parametric models that are used to study the effects of covariates on survival. The Cox-Snell residual for each subject, r_j, provides a

check of the overall fit of the model, denoted,

$$r_j = -\ln\left[\hat{S}_X(X_j \mid \mathbf{Z}_j)\right] = \hat{\Lambda}\left(X_j \mid \mathbf{Z}_j\right)$$

where $\hat{\Lambda}$ is the fitted model. If the model is correct, a probability integral transformation on the true death time X yields a uniform distribution for $F(X_j)$ and $S(X_j)$, which leads to a standard exponential distribution for the cumulative hazard function $(-\ln S = \Lambda(X))$. Therefore, a hazard plot of r_j versus the Nelson-Aalen estimator of the cumulative hazard of the r_j's should be a straight line through the origin with slope 1. For the log normal model the Cox-Snell residuals are

$$r_j = -\ln[1 - \Phi\left(\frac{\ln X_j - \hat{\mu} - \hat{\gamma}^T \mathbf{Z}_j}{\hat{\sigma}}\right)] \ , \quad j = 1, \ldots, n$$

Regression diagnostics using the Cox-Snell residuals is equivalent to the 'standardized' residuals based on the log-linear model representation, where $s_j = \frac{\ln X_j - \hat{\mu} - \hat{\gamma}^T \mathbf{Z}_j}{\hat{\sigma}}$, represents a standardized residual for the j^{th} subject. The log normal model assumption holds if the s_j's for all j, look like a censored sample from a standard normal distribution.

The martingale residual for a parametric model is defined as $\hat{M}_j = \delta_j - r_j$, where δ_j is the indicator of an observed event and r_j represents the Cox-Snell residual above. The martingale residuals can be thought of as the observed number of events minus the expected number of events under the assumed model. The maximum value of the residual is $+1$ and the minimum value is $-\infty$, therefore the martingale residuals have a skewed distribution. To obtain a more

'Normal-shaped' residual, a transformation can be done of the martingale residual, called the deviance residual, which is defined as

$$D_j = sign\left[\hat{M}_j\right]\left\{-2\left[\hat{M}_j + \delta_j \ln\left(\delta_j - \hat{M}_j\right)\right]\right\}^{\frac{1}{2}}$$

If the model is correct, then a plot of the deviance residuals against the observation number j should look like random noise. These residuals are motivated by the deviance in the theory of general linear models. Given n observations, a model containing n parameters (saturated model) provides a perfect fit to the data and can be used as a baseline for measuring the discrespancy for a fitted model with p parameters ($p < n$). The deviance of a model is denoted by twice the difference between the mazimized log likelihood under the saturated model and the fitted model.

Chapter 5

Parametric Regression Models for Data set I

5.1 Model 1 : Weibull Model

A Weibull model using backward elimination was fitted to find the prognostic factors to include in the model. Factors considered include *treated, agedx, dxyear, dxgrade* (1, 2, 3; see grading chart Section 2.1.1), *dxlvn* (yes = 1, no = 0), *dxposnod* ($N_x = -1$, $N_- = 0$, $N_+ = 1$), *staget* (<2cm = 1, 2-5 cm = 2, >5 cm = 3), *histcat* (ductal = 1, lobular = 2, NOS = 3), *bccach* (yes=1, no=0), *dxer* (ER -ive = 1, ER +ive = 2). All categorical variables with a code of 9, were recoded as missing variables. With regard to tumour size, only three levels out of 5 were used in the models. There were 3 pairs (42, 158, 207) of in-situ patients that were all dropped due to missing data, and there was only one pair (83) in the 'extended' category that was dropped as the parameter estimate for this category would be unreliable. Therefore, 84 pairs in total (n=168) were considered in all parametric survival models. The variable of interest, *treated*, is included in all models as a fixed-time covariate. The event of interest is

defined as time to death due to breast cancer or disease-specific survival (DSS). A patient who did not die from breast cancer, was still alive at the end of the study period, died from other causes, or was lost to follow-up was considered a *censored* observation.

Indicator Parameter Coding for the Class Variables (**baseline variables in bold**)

Class	Levels	Values
TREATED	2	1 **0**
DXLVN	2	1 **0**
DXGRADE	3	2 3 **1**
DXPOSNOD	3	1 0 **-1**
STAGET	3	2 1 **3**
HISTCAT	3	1 2 **3**
DXER	2	2 **1**
BCCACH	2	Y **N**

The results are shown in Table 5.1 and the final model selected is shown in Table 5.2. For the final model, there were 57 events (deaths due to breast cancer) and 110 right-censored values. One observation was dropped due to missing values. Chi-square tests were computed to test the restriction that each parameter is zero given all other covariates in the model. For class variables, overall chi-squared tests were used to test all levels of the factor are zero given the other covariates in the model. The results of the stepwise regression show that along with the treated variable, tumour size and diagnosis year are the only covariates to be kept in the final model.

A negative value of the coefficient $\hat{\gamma}_k$, $k = 1, \ldots, 10$, indicates worst survival as compared to the baseline. The results in the final model show an estimated treatment effect of $\hat{\gamma}_1 = -.59$ (P=0.0017) which implies a treated patient does significantly worse than a patient who did not receive the Hoffer treatment. There was a significant negative association between diagnosis

year and DSS. That is, the more recent the diagnosis year, the worse the survival rate. This is unexpected as one would expect the survival rate should remain constant based on a patient's year of diagnosis. In terms of tumour size and association with DSS, the estimate for breast cancer patients with medium sized tumours (2-5 cm) is $\hat{\gamma}_3 = 0.695$ (P=0.0001), which indicates significantly better survival as compared to patients with large tumours (>5 cm), and patients with small tumours do better than both groups. From Table 5.1, there appears to be a positive association between adjuvant chemotherapy and/or hormone therapy as all coefficients were positive until this factor was dropped from the model. Increasing age, high tumour grade, and lymphatic invasion were also associated with worse survival. These estimates are consistent with prognosis factors for breast cancer patients in clinical practice. Recall that the variable *dxer* has many missing observations. Therefore the significant drop in likelihood from step 1 to step 2 in Table 5.1 is due to the number of observations available (n = 126) before the variable *dxer* is dropped, as compared to 149 observations available after this variable is removed, and 167 available for the final model. The deviance residuals are shown in Figure 5.1. They show a fairly random pattern around zero, therefore the final Weibull model appears adequate.

Weibull	step1	step 2	step 3	step 4	step 5	step 6	step 7	Level
INT	20.450	19.199	18.908	18.611	19.102	18.864	19.867	
treated	-0.881	-0.725	-0.731	-0.718	-0.717	-0.665	-0.584	1
agedx	-0.005	-0.008	-0.008	-0.010	drop	-	-	
dxyear	-0.122	-0.107	-0.107	-0.106	-0.108	-0.114	-0.131	
dxlvn	-0.372	-0.381	-0.391	-0.430	-0.423	-0.366	drop	1
dxgrade	-0.384	-0.079	-0.064	-0.066	-0.080	drop	-	2
	-0.698	-0.524	-0.512	-0.518	-0.522	drop	-	3
dxposnod	-0.431	-0.380	drop	-	-	-	-	1
	-0.301	-0.286	drop	-	-	-	-	0
staget	0.268	0.422	0.422	0.4285	0.414	0.413	0.547	2
	0.848	0.726	0.721	0.7414	0.707	0.738	1.153	1
histcat	-0.017	drop	-	-	-	-	-	1
	-0.318	drop	-	-	-	-	-	2
dxer	drop	-	-	-	-	-	-	2
bccach	0.365	0.204	0.186	drop	-	-	-	Y
loglik	-72.20	-98.83	-99.05	-99.34	-99.48	-101.09	-114.83	
n	126	149					167	

Table 5.1: Stepwise (Backward Elimination) Regression for Weibull Model

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Level
$\hat{\mu}$	1	19.1437627	3.952712	23.45656	0.0001	
$\hat{\sigma}$	1	0.6069267	0.072144			
treated ($\hat{\gamma}_1$)	1	-0.5900317	0.188435	9.804517	0.0017	1
dxyear ($\hat{\gamma}_2$)	1	-0.1262582	0.042388	8.872137	0.0029	
staget	2			19.94926	0.0001	
($\hat{\gamma}_3$)	1	0.69509788	0.298724	5.414404	0.0200	2
($\hat{\gamma}_4$)	1	1.30935546	0.3225	16.48373	0.0001	1
Log likelihood		-125.4521364				

Table 5.2: Final Model for Weibull Model

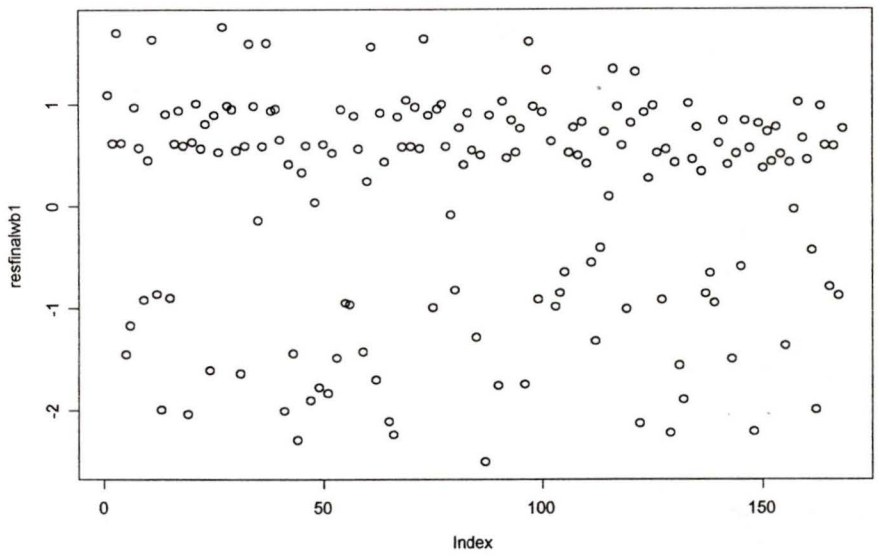

Figure 5.1: Deviance Residuals for Final Weibull Model

If we apply the transformation $\beta = \frac{-\gamma}{\sigma}$ on the original time scale for X, we obtain parameter estimates for the proportional hazards model. The relative risk of death due to breast cancer for a treated patient as compared to a control is $\exp\left(\frac{0.59}{0.61}\right) = \exp(0.97) = 2.64$, assuming the hazard rates are proportional and everything else held fixed. The relative risks for *staget* $= 1, 2,$ and 3, respectively are, $\exp\left(\frac{-1.31}{0.61}\right) = \exp(-2.16) = 0.12$, $\exp\left(\frac{-0.70}{0.61}\right) = \exp(-1.15) = 0.32$, and $\exp(0) = 1$. Since the relative risks are smaller than 1 for the two lower stages as compared to the baseline, a stage 3 patient with a tumour larger than 5 cm has a larger chance of dying from breast cancer than a stage 2 patient, and an even greater chance of dying than a stage 1 patient.

A rough check of the proportionality assumption for *treated* and *staget*, can be done using a plot of the estimated survival curves for each factor. If the survival curves do not cross, then the hazard rates are proportional. Figures 3.1 and 5.2 show the survival curves for the two treatment groups and for each of the stages, respectively. In Figure 3.1, the curves do not appear to cross, and for the three tumour sizes, there is some intital crossing at the very beginning, but in general the curves do not cross, with stage 1 patients showing the best survival probability as compared to both higher stages. If a Cox proportional hazards model was fit to this data, one could also plot survival curves taking into account other significant factors in the model at their average values to check for proportionality.

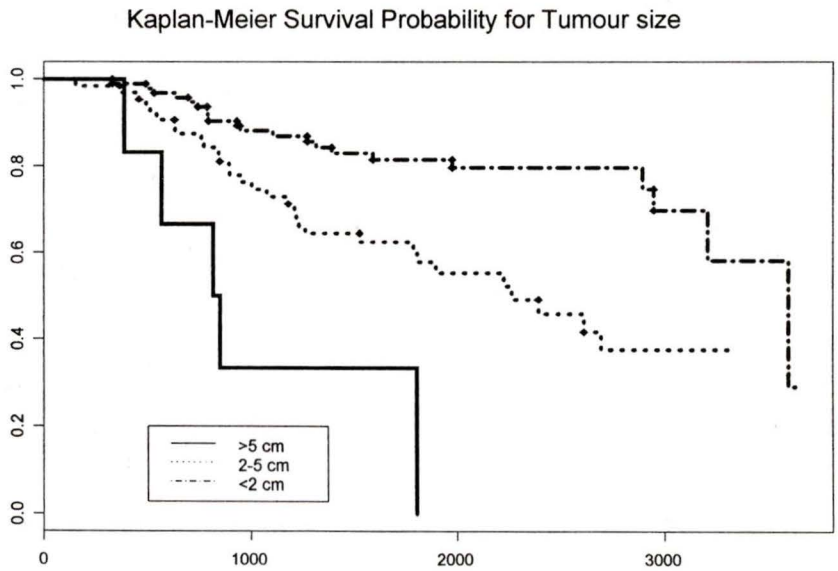

Figure 5.2: Survival Curves for *staget*

Using the accelerated failure-time representation, recall that the acceleration factor is $\theta = -\gamma$, and therefore the acceleration factor for a treated patient as compared to a control is $\exp(0.59) = 1.80$, which implies the median lifetime for a control is approximately 1.8 times longer than that of a treated patient, assuming all other variables remain constant. The median lifetime for a stage 3 patient as compared to a stage 1 patient is $\exp(-1.31) = 0.27$, and similarly for a stage 3 patient as compared to a stage 2 is $\exp(-0.70) = 0.50$. Therefore the median disease-specific survival time for stage 3 patients is approximately a quarter that of stage 1 patients and only half that of stage 2 patients when everything else is held fixed.

5.2 Model 2: Log Logistic Model

The log logistic stepwise results are shown in Table 5.3 and the final model selected is shown in Table 5.4. The same variables were considered and the same parameter coding was used as above. For the final model, there were 50 events (deaths due to breast cancer) and 100 right-censored values as well as 18 missing observatons. The results of the stepwise regression show that tumour size, diagnosis year and grade are to be kept in the final model with the treated variable. A significant drop in the likelihood is seen in Table 5.3 when *dxer* is dropped (the number of observations available changes from 126 to 149) with 150 available for the final model.

The final model had an estimated treatment effect of $\hat{\gamma}_1 = -0.69$ which indicates worse survival for the treated group as compared to the controls. This result for treatment effect is

worse than was observed in the Weibull model after adjusting for *staget, dxgrade*, and *dxyear*. The coefficient for diagnosis year was negative indicating worse survival the more recent the year a patient was diagnosed with breast cancer. Tumour grade showed a significant overall effect on survival with a higher grade associated with more breast cancer death. From Table 5.3, there appears to be a positive association between *bccach* and DSS before this variable was dropped. The age of diagnosis did not appear to have an effect on death due to breast cancer. One would expect a negative association between age and DSS, that is, as you get older, your chances of dying of the disease would increase, however studies have shown that older women experience more non-breast cancer related deaths {22}. Patients with lymphatic invasion showed worse survival rates as compared to those patients with no lymphatic invasion as did ER negative patient as compared to ER positive patients. The deviance residuals are shown in Figure 5.3, which appear to show a random pattern about zero.

Log-log	step1	step 2	step 3	step 4	step 5	step 6	Level
INT	19.456	19.416	18.052	18.114	17.479	17.762	
treated	-0.864	-0.864	-0.693	-0.705	-0.678	-0.671	1
agedx	drop	-	-	-	-	-	
dxyear	-0.114	-0.114	-0.101	-0.106	-0.103	-0.106	
dxlvn	-0.242	-0.239	-0.189	-0.221	-0.268	drop	1
dxgrade	-0.424	-0.421	-0.102	-0.087	-0.080	-0.059	2
	-0.710	-0.706	-0.553	-0.513	-0.511	-0.472	3
dxposnod	-0.614	-0.618	-0.526	drop	-	-	1
	-0.418	-0.420	-0.345	drop	-	-	0
staget	0.273	0.275	0.477	0.486	0.509	0.471	2
	0.900	0.906	0.853	0.866	0.915	0.883	1
histcat	-0.113	-0.113	drop	-	-	-	1
	-0.317	-0.320	drop	-	-	-	2
dxer	0.0292	drop	-	-	-	-	2
bccach	0.387	0.396	0.226	0.201	drop	-	Y
loglik	-75.17	-75.17	-100.57	-100.87	-101.48	-101.75	
n	126		149			150	

Table 5.3: Stepwise (Backward Elimination) Regression for Log logistic Model

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Level
$\hat{\mu}$	1	17.8393283	4.056479	19.34009	0.0001	
$\hat{\sigma}$	1	0.4771989	0.059054			
treated ($\hat{\gamma}_1$)	1	-0.6921738	0.195987	12.47315	0.0004	1
dxyear ($\hat{\gamma}_2$)	1	-0.107234	0.043707	6.019406	0.0141	
dxgrade	2			7.769273	0.0206	
($\hat{\gamma}_3$)	1	-0.1219968	0.309836	0.155037	0.6938	2
($\hat{\gamma}_4$)	1	-0.6374343	0.291945	4.767248	0.0290	3
staget	2			8.018464	0.0181	
($\hat{\gamma}_5$)	1	0.51275297	0.366203	1.96053	0.1615	2
($\hat{\gamma}_6$)	1	0.94026149	0.384329	5.985357	0.0144	1
Log likelihood		-104.5711557				

Table 5.4: Final Model for Log Logistic Model

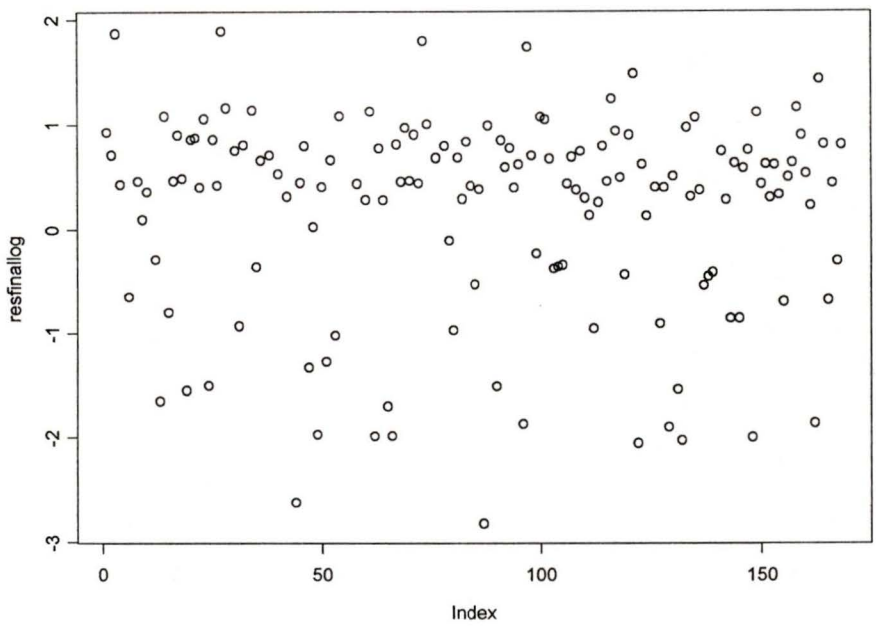

Figure 5.3: Deviance Residuals for Final Log Logistic Model

In terms of a proportional odds representation, where $\exp\left(-\beta^T \mathbf{Z}\right)$ represents the relative odds of survival, the relative odds for a patient given the vitamin treatment is $\exp(-1.45) = 0.23$, therefore Hoffer-treated patients have 0.23 times lesser odds of surviving than controls (everything else held fixed). The relative odds of surviving for a stage 1 patient compared to a stage 3 patient are $\exp(1.97) = 7.17$, and for a stage 2 patient are $\exp(1.07) = 2.93$, which implies that a stage 1 patient has approximately 7 times more odds of surviving from breast cancer than a baseline patient (large tumour), and a stage 2 patient has almost 3 times better odds than the baseline, when all other variables remain constant. In terms of *dxgrade*, the

baseline grade is the least severe one (grade 1), and the relative odds of survival for a patient with a grade of 2 as compared to the baseline are $\exp(-.256) = 0.77$, and for a patient with the highest grade (most severe disease) are $\exp(-1.34) = 0.26$. Therefore assuming all other variables remain fixed, a grade 2 patient has 0.77 lesser odds of surviving (or $\frac{1}{0.77} = 1.29$ greater odds of dying from breast cancer), and a grade 3 patient has 0.26 lesser odds of surviving (or $\frac{1}{0.26} = 3.80$ greater odds of dying) than a baseline patient. Also, the acceleration factor for a treated patient as compared to a control is $\exp(0.69) = 1.99$ which indicates that a control patient has approximately double the median lifetime than a treated patient (all else held fixed).

5.3 Model 3: Log Normal Model

The log normal stepwise results are shown in Table 5.5 and the final model selected is shown in Table 5.6. The same variables were considered and the same parameter coding was used as above. For the final model, there were 50 events (deaths due to breast cancer) and 100 right-censored values as well as 18 missing observatons. The final model includes the same covariates as the log logistic model: tumour size, diagnosis year and grade in addition to the treatment variable.

The estimated treatment effect for the log normal model, $\hat{\gamma}_1 = -0.71$, indicates worse survival for the treated group than the other two models, after adjusting for the final covariates. Table 5.6 shows the same associations with death due to breast cancer as the previous (log logistic) model. The deviance residuals are shown in Figure 5.4. There is no pattern seen, therefore

the model seems reasonable for the data.

Using the accelerated failure-time for the log nomal, the acceleration factor for a treated patient as compared to a control is exp(.71) = 2.04, which is similar to the result in the log logistic model.

Log-norm	step1	step 2	step 3	step 4	step 5	step 6	Level
INT	19.385	18.891	18.842	18.833	18.834	18.934	
treated	-0.892	-0.711	-0.708	-0.713	-0.7052	-0.734	1
agedx	0.003	drop	-	-	-	-	
dxyear	-0.111	-0.105	-0.106	-0.111	-0.111	-0.112	
dxlvn	-0.155	-0.123	-0.118	-0.148	drop	-	1
dxgrade	-0.586	-0.276	-0.268	-0.246	-0.220	-0.236	2
	-0.770	-0.644	-0.633	-0.588	-0.554	-0.626	3
dxposnod	-0.835	-0.688	-0.695	-0.654	-0.568	drop	1
	-0.555	-0.455	-0.459	-0.436	-0.365	drop	0
staget	0.239	0.443	0.447	0.456	0.431	0.432	2
	0.956	0.889	0.899	0.914	0.894	0.875	1
histcat	-0.227	-0.488	drop	-	-	-	1
	-0.443	-0.437	drop	-	-	-	2
dxer	drop	-	-	-	-	-	2
bccach	0.396	0.135	0.155	drop	-	-	Y
loglik	-76.32	-100.85	-100.86	-101.24	-101.41	-103.38	
n	126	149				150	

Table 5.5: Stepwise (Backward Elimination)Regression for Log Normal Model

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Level
$\hat{\mu}$	1	18.21491	4.281175	18.10206	0.0001	
$\hat{\sigma}$	1	0.8802737	0.097484			
treated ($\hat{\gamma}_1$)	1	-0.7105756	0.199869	12.6395	0.0004	1
dxyear ($\hat{\gamma}_2$)	1	-0.1102474	0.046199	5.694775	0.0170	
dxgrade	2			6.723767	0.0347	
($\hat{\gamma}_3$)	1	-0.2962208	0.340534	0.756676	0.3844	2
($\hat{\gamma}_4$)	1	-0.7190539	0.326095	4.862233	0.0275	3
staget	2			10.11217	0.0064	
($\hat{\gamma}_5$)	1	0.5083362	0.401599	1.602199	0.2056	2
($\hat{\gamma}_6$)	1	1.0362069	0.414601	6.246427	0.0124	1
Log likelihood		-104.7186885				

Table 5.6: Final Model for Log Normal Model

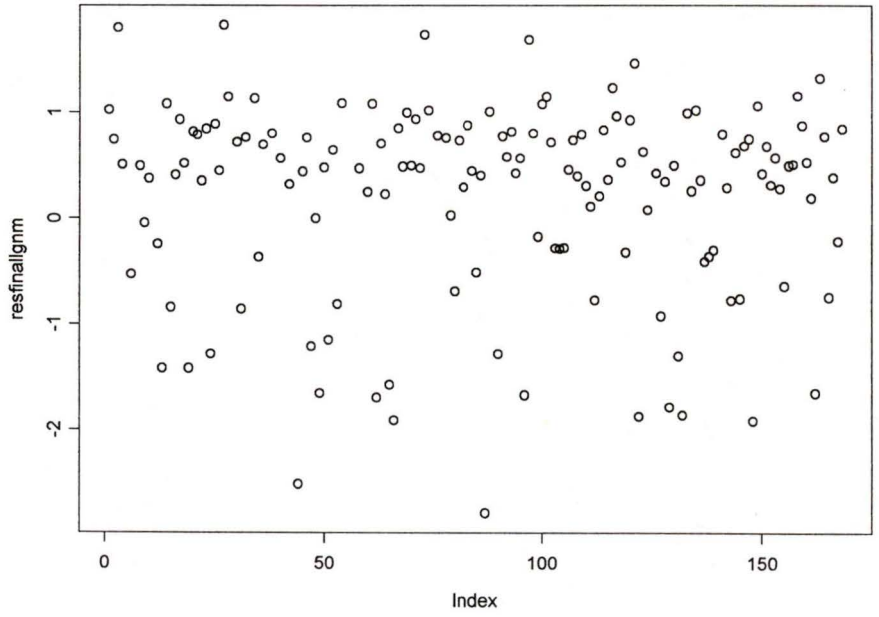

Figure 5.4: Deviance Residuals for Final Log Normal Model

5.4 Choosing an Appropriate Model: Generalized Gamma Distribution and Hazard Plots

The Generalized Gamma model is most often used to discriminate among alternative parametric models rather than as a final model for survival data {26}. The Gamma distribution for a lifetime X is

$$f(x) = \frac{\rho(\rho x)^{\alpha-1}\exp(-\rho x)}{\Gamma(\alpha)}, \quad \rho, \alpha > 0, \; x > 0$$

where $\Gamma(\alpha)$ represents the gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp(-x)\,dx$. For the log-linear model $Y = \mu_0 + \sigma W$, as before, and W has the density function,

$$f(w) = \frac{|\theta| \left[\frac{\exp(\theta w)}{\theta^2}\right]^{\left(\frac{1}{\theta^2}\right)} \exp\left[\frac{-\exp(\theta w)}{\theta^2}\right]}{\Gamma\left(\frac{1}{\theta^2}\right)}, \quad -\infty < w < \infty,$$

the 3-parameter *generalized gamma* distribution. This model reduces to the Weibull model when $\theta = 1$, the exponential when $\theta = \sigma = 1$, and the log normal when $\theta = 0$ {26}.

Wald tests for $H_o : \theta = 1$, or $H_o : \theta = 0$ can be done to check the assumption of a Weibull or log normal regression model respectively. To test the constraints $H_o : \theta = 1$ and $H_o : \theta = 0$, Lagrange Multiplier test statistics are used.

The generalized gamma was fit to the data to test the appropriateness of the final Weibull model given in Table 5.2. The *treated* variable as well as *dxyear* and *staget* are included along with the restriction that $\theta = 1$. The parameter estimates for this generalized gamma model were the same as in Table 5.2. The Lagrange Chi-Square test in Table 5.7 shows a small P-value

and therefore there is strong evidence against the null hypothesis. The Weibull model does not appear to be an appropriate model.

Variable	DF	Estimate	Std. Error	ChiSquare	Pr>Chi
SCALE	1	0.6069267	0.072144		
SHAPE	0	1	0		
Lagrange Multiplier ChiSquare for Shape1 29.51569 Pr>Chi is **0.0001**					

Table 5.7: Test $H_o : \theta = 1$, Weibull model is adequate

The generalized gamma is fit to the data along with the restriction that $\theta = 0$ to test the adequacy of the log normal model. The final covariates in the log normal model, (*treated, dxyear, dxgrade* and *staget*), were used and the same parameter estimates were reported as in Table 5.6. The LaGrange Multiplier Chi-Square to test the adequacy of the log normal model in Table 5.8 shows a large P-value, therefore, the log normal model appears to be adequate relative to the generalized gamma.

Variable	DF	Estimate	Std. Error	ChiSquare	Pr>Chi
SCALE	1	0.8802737	0.097484		
SHAPE	0	0	0		
Lagrange Multiplier ChiSquare for Shape1 0.529853 Pr>Chi is **0.4667**					

Table 5.8: Test $H_o : \theta = 0$, Log Normal model is adequate

It would appear that the log normal distribution may provide the best fit to the data. To examine the appropriateness of the assumption of the log normal distribution for the risk of death of breast cancer, plots of the estimated cumulative hazard rates given certain patient characteristics can be plotted using the Nelson-Aalen estimator of the cumulative hazard $\hat{\Lambda}_{N-A}(x)$. Recall, that the slope of the Nelson-Aalen estimator provides a crude estimate of the hazard rate $\lambda(x)$, where the risk of death for the log normal hazard increases steadily, then decreases after a certain point in time {22}. If the hazard function for the data follows this underlying

assumption for the risk of breast cancer death, the choice of the log normal model is reasonable. The plots for baseline patient characteristics considered in the study are given in Figures 5.5 through 5.13. The new variable *age4bin* was created to separate the continuous variable *agedx*, into four categories. The binned age groups are: 28 - 40 (premenopausal), 40 - 50 (menopausal), 50-58 (postmenopausal), and 58-88 (oldest age group). The plots of the estimated cumulative hazard rates show that the highest death rates correspond to the levels of the factors with worse survival in the parametric models, (except *bccach*). The slopes of the Nelson-Aalen estimators show a slow increase at the beginning of the study, with increasing slope, and then eventually decreasing slope at later times. This is consistent with the log normal model assumption.

Figure 5.14 shows the hazard plot of the Cox-Snell residuals. There is some deviation from a straight line but the differences are small. To check for interaction effects, a log normal model was fit which included interaction terms between *treated* and the other factors in the model, however no significant interactions were found in this model.

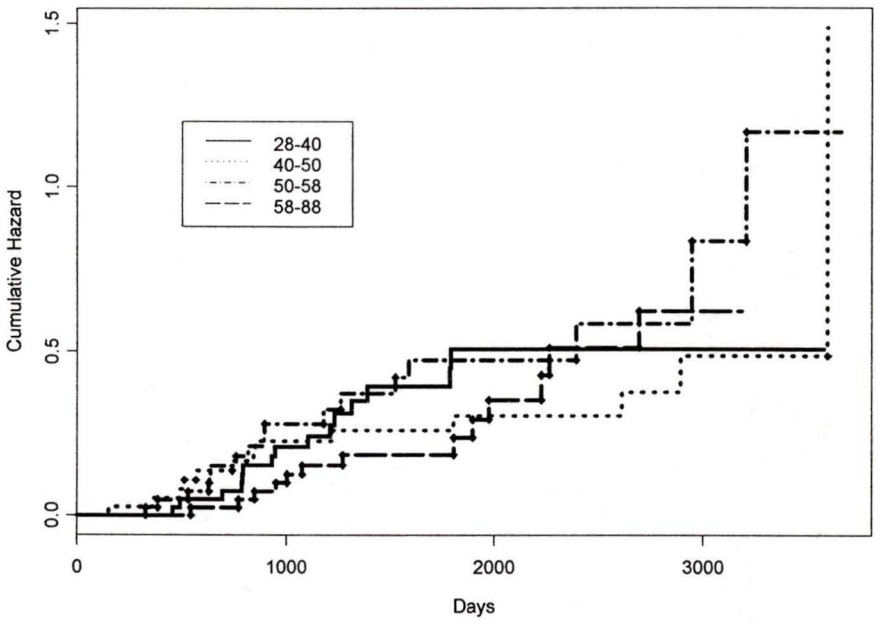

Figure 5.5: N-A Estimator for *age4bin*

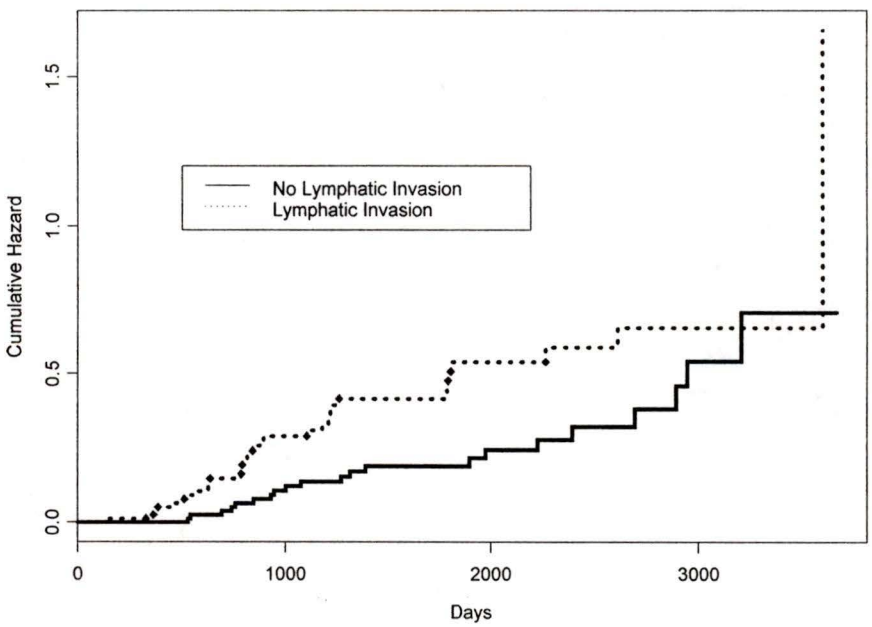

Figure 5.6: N-A Estimator for *dxlvn*

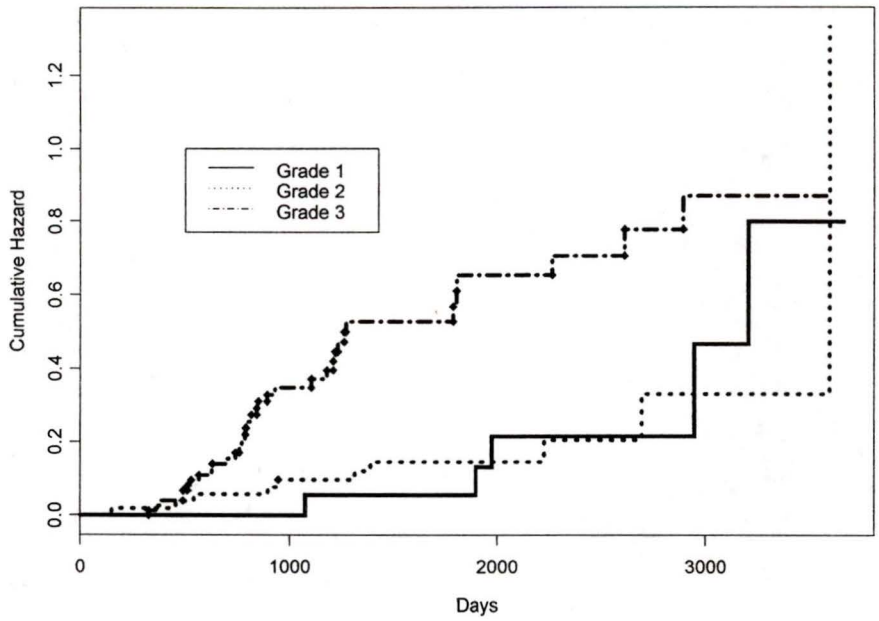

Figure 5.7: N-A Estimator for *dxgrade*

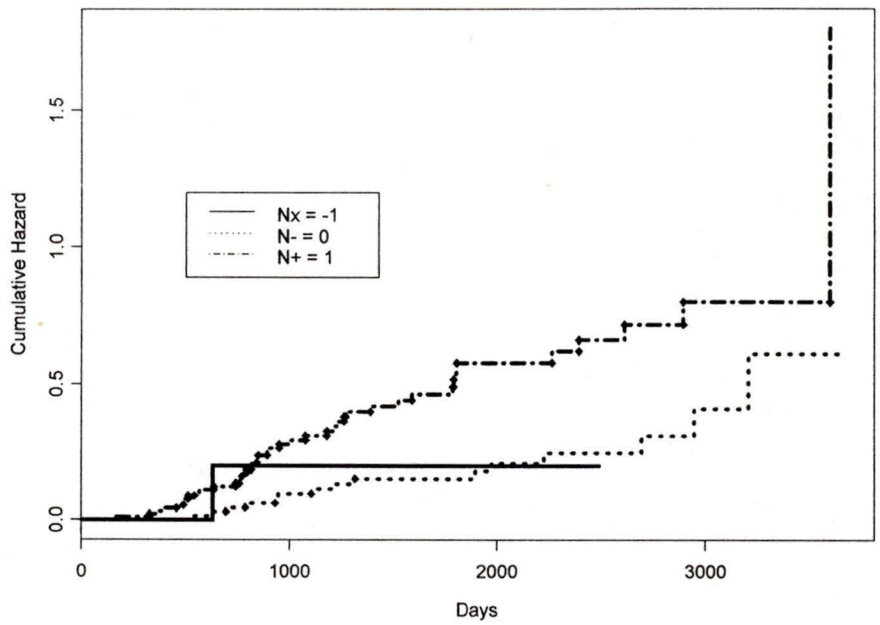

Figure 5.8: N-A Estimator for *dxposnod*

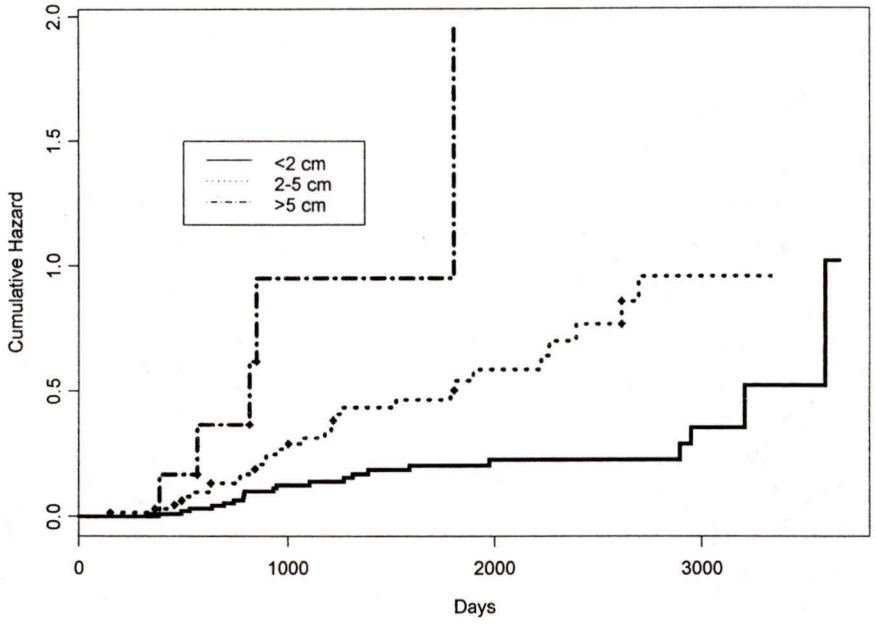

Figure 5.9: N-A Estimator for *staget*

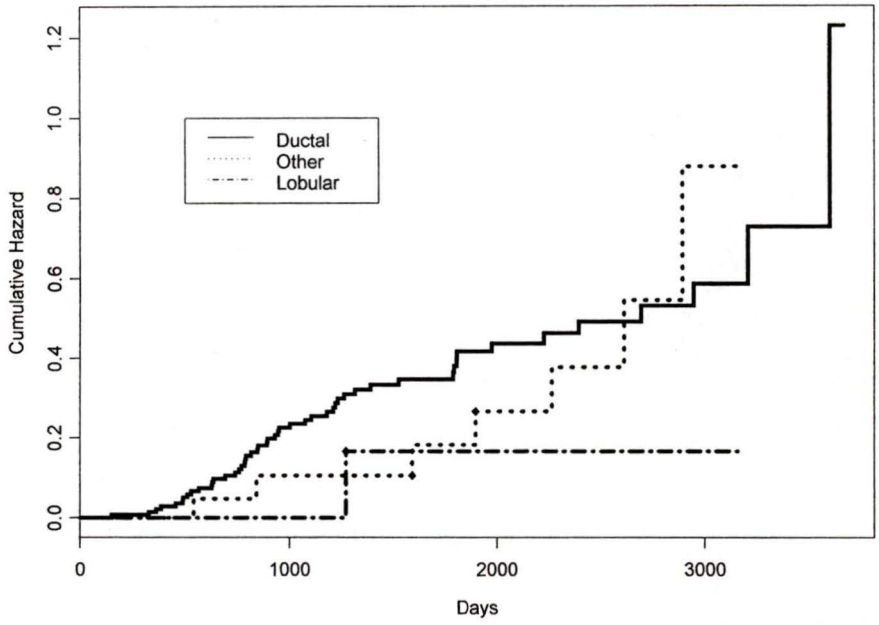

Figure 5.10: N-A Estimator for *histcat*

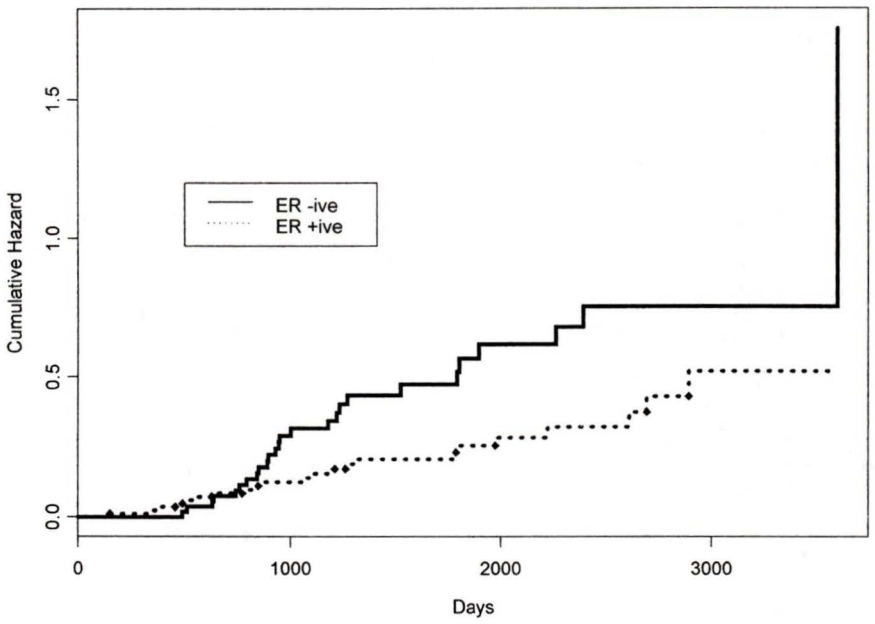

Figure 5.11: N-A Estimator for *dxer*

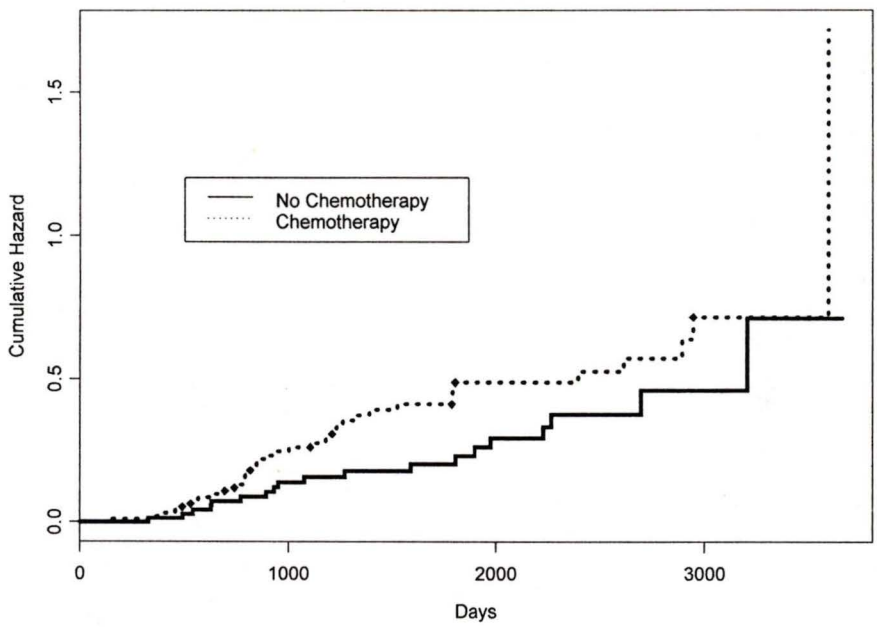

Figure 5.12: N-A Estimator for *bccach*

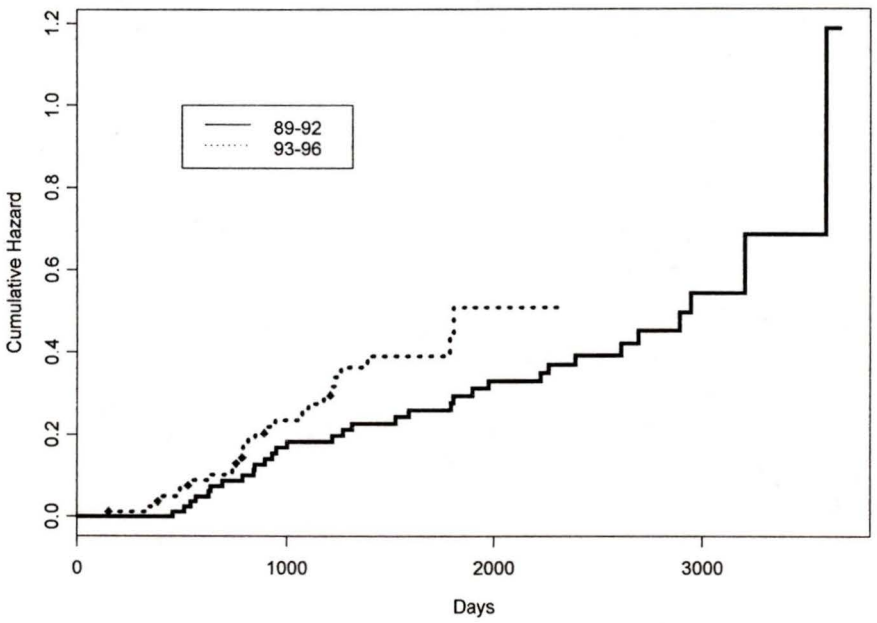

Figure 5.13: N-A Estimator for *dxyear*

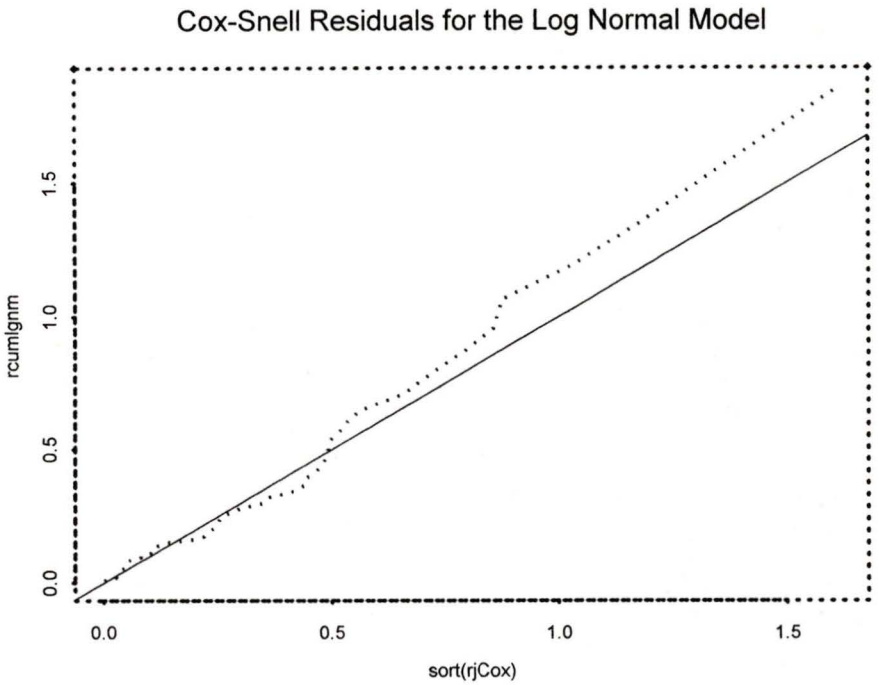

Figure 5.14: Hazard Plot of Cox-Snell Residuals vs. $\hat{\Lambda}_{N-A}(r_j)$

5.5 Model 4: Log-linear Models for paired data

Up to now all models have been used to assess treatment effect by adding prognostic factors to account for differences within groups. To incorporate the matching (or stratification), the *pair* factor can be included in the model to account for pair variation and increase the similarity between treatment groups. A parameter is given in the model for each pair to describe failure time properties of the stratum defined by that pair {25}. A log-linear survival model for paired

failure times can be repesented as:

$$Y_{sj} = \log(X_{sj}) = \mu + \tau_s + \gamma^T \mathbf{Z}_{sj} + \sigma W_{sj} \ , \ j = 1, 2$$

where X_{sj} represents the failure time for an individual in the sth pair (stratum), τ_s represents the effect of the s^{th} pair, γ^T is the vector of risk coefficients and $\mathbf{Z}_{sj}$ is the time independent risk vector for an individual {25}. The random variable W_{sj} has error distribution $f(w_{sj})$. In the paired model with just the treated variable and a factor variable for each pair, Z_{sj} is a scalar that equals one for a treated patient and 0 for a control patient. Three log-linear models were fit (Weibull, Log logistic, Log normal) and the parameter estimates for treatment effect are reported in Tables 5.9 to 5.11 (the parameters for each pair effect are not shown).

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Ch	Label/Value
TREATED	1			38.4844	0.0001	TREATED
	1	**-0.7204**	0.1161	38.4844	0.0001	1
	0	0	0			0

Table 5.9: Weibull Paired Model

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
TREATED	1			30.89769	0.0001	TREATED
	1	**-0.6592**	0.1186	30.89769	0.0001	1
	0	0	0			0

Table 5.10: Log Logistic Paired Model

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
TREATED	1			27.9229	0.0001	TREATED
	1	**-0.6395**	0.1210	27.9229	0.0001	1
	0	0	0			0

Table 5.11: Log Normal Paired Model

Taking into account the matched pairs in the model, the coefficient for treatment effect indicates decreased survival for the treated group in all models, with the Weibull model showing the least favourable prognosis. The residuals for the log normal model are shown in Figure 5.15, the model appears to fit the data well as most of the observations are close to zero.

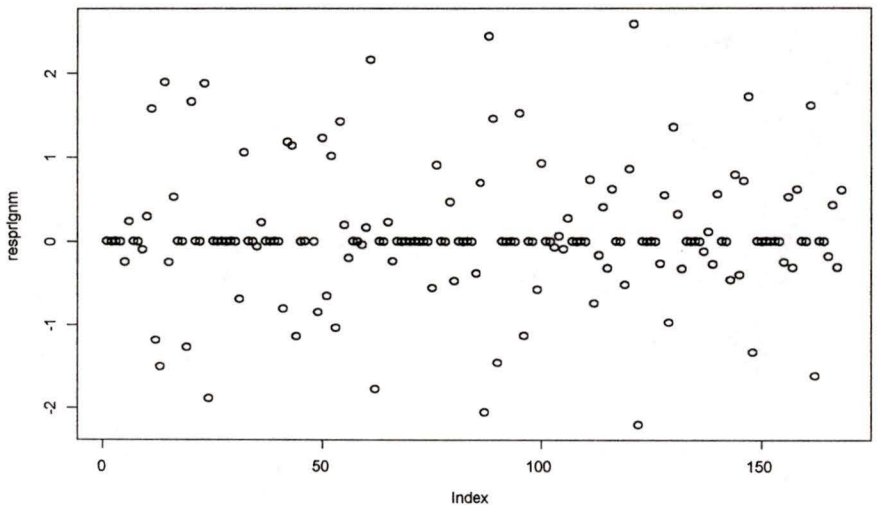

Figure 5.15: Deviance Residuals for Log Normal Paired Model

Chapter 6

Time to Relapse for Data set I: Log Normal Survival Model

6.1 Preliminary Analysis

Up to now we have investigated the effects of covariates on disease-specific survival. In this chapter, the event of interest is regional or distant relapse (or death) for disease-free survival (DFS) {21}. Any relapse found in regional lymph nodes as well as distant sites is considered 'treatment failure' {27}. An individual is said to be 'disease-free' if she is alive without recurrence of the disease {26}. The 88 pairs of Data set I were used in the analysis with the outcome variable *syssurv*, to compare the time to systemic relapse between the two treatment groups. The event variable *sysevent* was used with *syssurv*, where a failure is defined as any regional or distant relapse or death from breast cancer.

A systemic recurrence occurs when breast cancer cells travel to other parts of the body (metastasize). Most recurrences occur between 2 to 6 years after the initial diagnosis, if at all {17}. Therefore, after a certain point in time the risk of relapse (and breast cancer death) is

thought to decrease {22}. This assumption is consistent with the log-normal hazard function. Prognosis may depend on risk factors such as age, grade, histology, ER status, lymphatic involvement, and chemotherapy. In addition, the risk of relapse is related to tumour size, the presence and number of cancer-involved lymph nodes, and whether the cancer is in situ or invasive {17}. For invasive cancers, the risk of relapse is related to the stage {17}.

6.1.1 Data Quality

The variables *srmhist*, *srmdate*, and *srmdes* were used with *dxdate* to check the accuracy of the event variable *sysevent* in the data dictionary. In addition to relapses, the *srmdate* and *srmhist* have also been used to capture extent of disease at the time of diagnosis {28}. For example, if the patient was diagnosed with positive lymph nodes at the time of initial diagnosis, the date of positive confirmation of positive lymph nodes would be entered in the *srmdate1* field along with the appropriate topography code (**C**00.0) in *srmhist*. Therefore a *srmdate1* that is equal to or within 30 days of *dxdate* should not be coded as a relapse (ie, the value of *sysevent* should be zero in all of these cases). Of the 88 pairs, no initial diagnoses were incorrectly recorded as relapses when the two dates were within 30 days, however two pairs (the control subjects in pairs 201 and 218) had dates recorded in *srmdate* that were considerably more than 30 days from initial diagnosis (implies a relapse), however their event variables *sysevent* were coded as nonrelapse patients (*sysevent* = 0). These subjects were kept in the model unchanged. In addition, there were three observations where *syssurv* was bigger than *dthcsurv* for the matched controls:

pair	*dthcsurv*	*syssurv*
40	1761	3041
117	2065	2205
153	1288	1606

Since the values above for *dthcsurv* were used to model disease-specific survival in chapter 5, the values of *syssurv* were changed to 1761, 2065, and 1288 respectively for consistency (new variable *syssurv1*).

All women who have been diagnosed with breast cancer need follow-up visits to check for a local or systemic recurrence. At the BC Cancer Agency, follow-up of patients is done either through a BCCA clinic or through letter follow-up for those patients no longer attending a BCCA clinic {28}. For those patients no longer at a BCCA clinic, annual follow-up forms are sent to the patient's physician to update information on survival status, recurrence information as well as date of last contact {28}. There were no variables in the database to check the accuracy of the follow-up (recurrence) information, however C. Speers reported that in most cases the physician will return the forms annually, and patients in urban centres are more likely to return to a BCCA clinic at the time of recurrence, therefore their information would remain current in the CAIS database.

6.1.2 Data Matching Checks

Figure 6.1 shows the Kaplan-Meier survival curves for the two groups with respect to systemic relapse. The figure indicates larger DFS for the control group as compared to the treated group. The log-rank test for the original groups is given below. There is strong evidence against the null hypothesis therefore the survival distributions for the two treatments are different.

syssurv1	Statistic	df	Significance
Log Rank	16.25	1	.0001

Test Statistic for Equality of Survival Distributions for TREATED

The data was recoded such that a treated patient who relapsed before starting the Hoffer treatment was instead considered a control (new treatment variable *treateds*). There were 18 treated patients who were changed to controls (by *pair*: 73, 98, 103, 109, 112, 144, 147, 149, 152, 168, 174, 198, 201, 220, 222, 224, 228, 234). These 18 out of 88 patients in the treated group may have been referred to Dr. Hoffer after treatment relapse because they felt that the conventional therapy had not worked for them.

Tables 6.1 and 6.2 shows a summary of the events (deaths due to breast cancer) and relapses before and after the treatment groups were changed. In both tables 6.1 and 6.2, the first column indicates that 107 of the patients (out of 176) did not have a systemic relapse. Of these, 67 were control patients and 40 were treated patients and none of these patients died of breast cancer. For *sysevent* = 1, 69 of the 176 patients did have a systemic relapse. Table 6.1 shows the breakdown of relapses for the original two groups is 21 relapses for the controls and 48 for the treated group. Table 6.2 indicates that when the 18 new controls are taken into account, the new grouping, is more balanced, with 39 relapses for the controls as compared to 30 for the treated patients. Although there were no deaths for those who did not have a systemic relapse, 59 out of 69 (85.6 %) total relapse patients have so far died from breast cancer. Therefore the proportion of patients who die shortly after a regional or distant recurrence is high. For the old data set, there were more deaths reported for the treated group than controls (42 versus 17),

but for the new groups less deaths were seen for the treated group (25 versus 34).

Figure 6.2 shows the Kaplan-Meier survival curves for new treatment groups. There was an apparent difference between the survival curves for the original data set, with the controls showing worse survival, however, the groups appear to have very similar survival curves for the new treatment groups. The log-rank test below indicates the same result. With the new controls, a log normal survival model was used to assess whether the vitamin treatment prolongs time to relapse.

syssurv1	Statistic	df	Significance
Log Rank	.45	1	.5002

Test Statistic for Equality of Survival Distributions for TREATEDS

	Total	# of Events (Deaths)	# Censored	% Censored
Sysevent = 0 (no relapse)	**107**	**0**	**107**	100
control	67	0	67	100
treated	40	0	40	100
Sysevent = 1 (relapse)	**69**	**59**	**10**	14.49
control	21	17	4	19.05
treated	48	42	6	12.50
Overall	**176**	**59**	**117**	66.48

Table 6.1: Status of Observations for *sysevent* for Original Treatment Groups

	Total	# of Events (Deaths)	# Censored	% Censored
Sysevent = 0 (no relapse)	**107**	**0**	**107**	100
control	67	0	67	100
treated	40	0	40	100
Sysevent = 1 (relapse)	**69**	**59**	**10**	14.49
control	39	34	5	12.82
treated	30	25	5	16.67
Overall	**176**	**59**	**117**	66.48

Table 6.2: Status of Observations for *sysevent* for New Treatment Groups

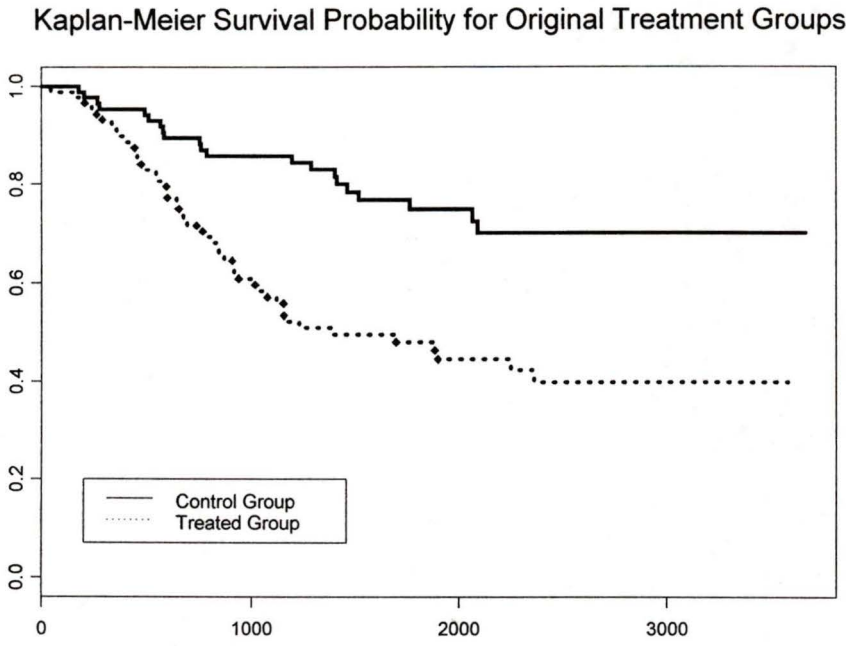

Figure 6.1: Survival Curves for the Time to Regional/Distant Relapse or Death for the Original Treatment Groups

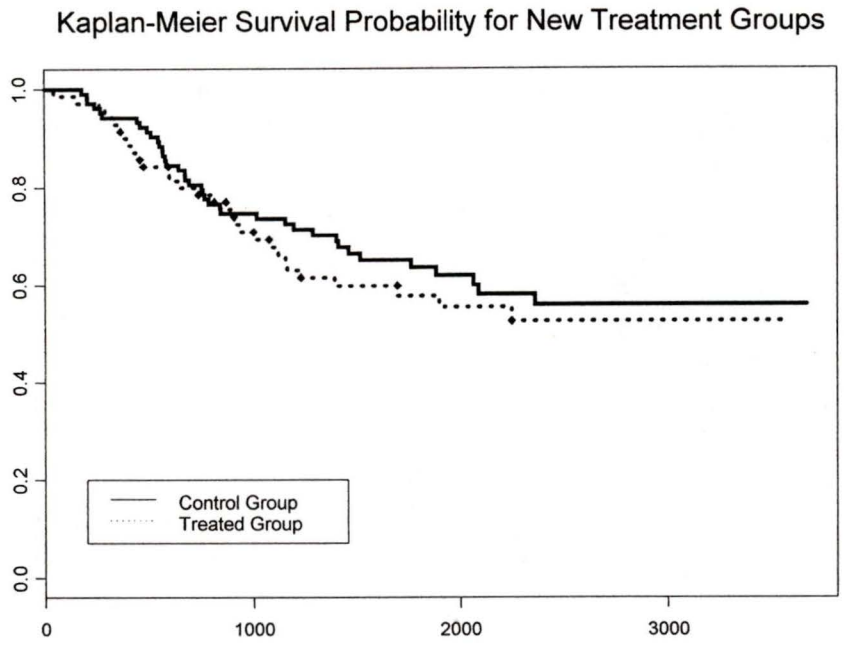

Figure 6.2: Survival Curves for the Time to Regional/Distant Relapse or Death for the New Treatment Groups

6.2 Model Fit: Log Normal Model

The final covariates considered for the log normal model are *treateds*, *agedx*, *dxyear*, *dxlvn*, *dxgrade*, *dxposnod staget*, *histcat*, *dxer*, *bccach*, and *behvr*, with the outcome variable *syssurv1* and event indicator *sysevent*. The final model results for the log normal model are given in Table 6.3. The final covariates after using backward elimination are: *treateds, dxyear, dxgrade, and dxposnod.* The levels of each variable are shown in the last column of Table 6.3, with the same

baseline for each variable as in chapter 5. Out of 168 patients considered (84 pairs, no in-situ or extended stages), there were 18 observations dropped due to missing values, 59 events and 91 censored observations included in the model. The results show a negative value of the coefficient for treatment effect which indicates decreased survival for the treated group considering the other covariates in the model, however this estimate is not significant (P=0.1673). Therefore, there appears to be no significant difference between the time to relapse survival rates for the new treatment and control groups. The deviance residuals and Cox-Snell residuals are given in Figures 6.3 and 6.4, indicating the model appears to be adequate.

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Level
$\hat{\mu}$	1	19.3029185	5.554607	12.07645	0.0005	
$\hat{\sigma}$	1	1.22893437	0.127242			
treateds ($\hat{\gamma}_1$)	1	-0.3535007	0.256014	1.906564	0.1673	1
dxyear ($\hat{\gamma}_2$)	1	-0.1128942	0.059115	3.64712	0.0562	
dxgrade	2	10.88324		10.88324	0.0043	
($\hat{\gamma}_3$)	1	0.19308706	0.432316	0.199482	0.6551	2
($\hat{\gamma}_4$)	1	-0.6976142	0.412976	2.853515	0.0912	3
dxposnod	2			8.554616	0.0139	
($\hat{\gamma}_5$)	1	-0.9278911	0.810215	1.311575	0.2521	1
($\hat{\gamma}_6$)	1	-0.1667105	0.824518	0.040881	0.8398	0
Log likelihood		-142.897750				

Table 6.3: Final Model for Time to Systemic Relapse

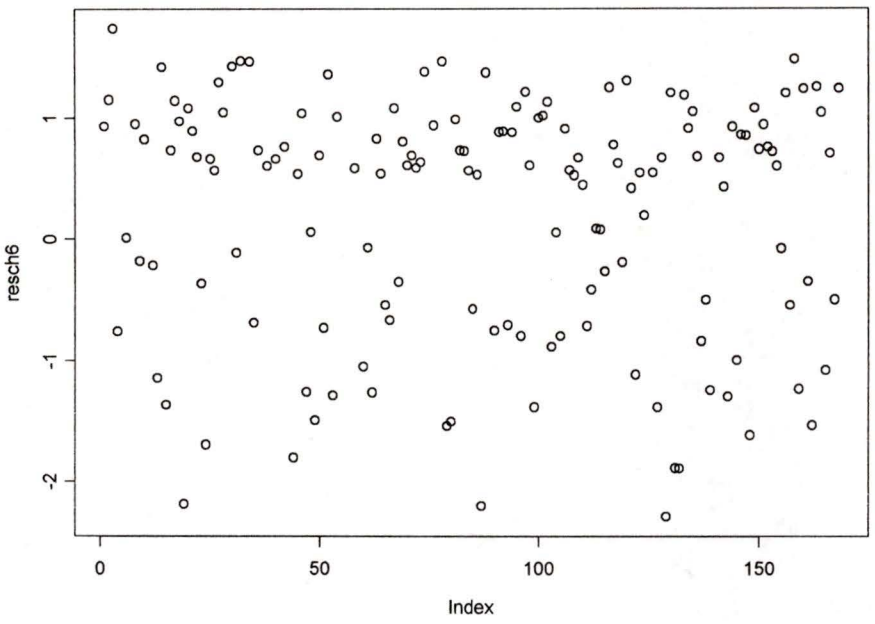

Figure 6.3: Deviance Residuals for Time to Systemic Relapse Model

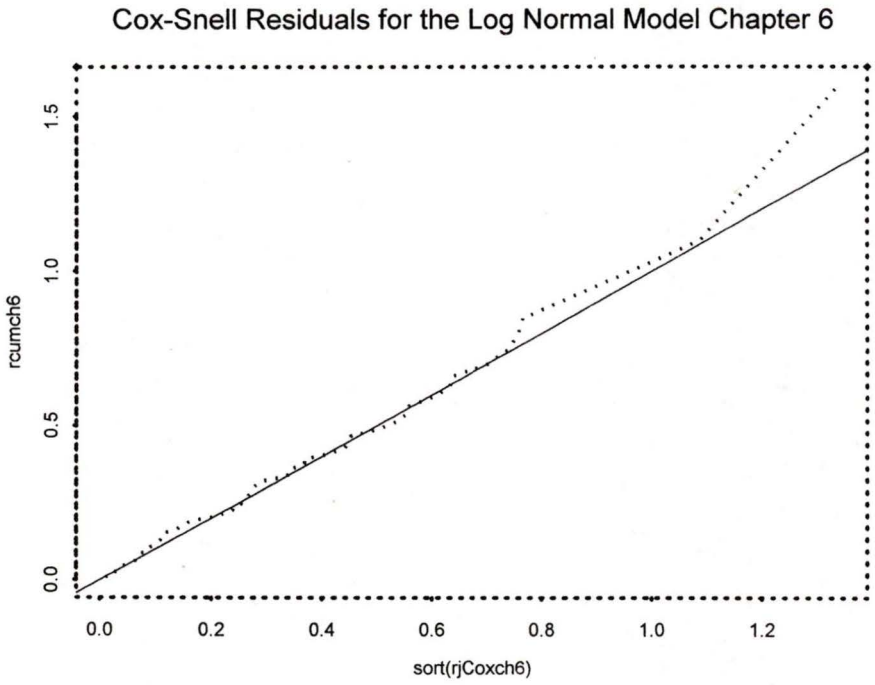

Figure 6.4: Hazard Plot of Cox-Snell Residuals vs. $\hat{\Lambda}_{N-A}(r_j)$

Chapter 7

Log Normal Survival Model for Disease-specific Survival for Data set II

7.1 Description of Data set II

Data set II consists of 2560 women who have been diagnosed with nonmetastatic breast cancer between the years 1989 to 1996 inclusive at the Victoria office of the BCCA. This data set contains all diagnosis, staging and treatment variables as before, but the patients were not matched by pair, and the Hoffer patients were not identified (no *treated* variable). Therefore it was necessary to identify which of these individuals were actually Hoffer patients who received the vitamin treatment. Hoffer patients were extracted by comparing the BCCA patient identification number, *idtran*, of Data set I with Dataset II. The *idtran* number is a unique number given to each breast cancer patient that would be the same in both data sets. There were 99 Hoffer entries that were identified in the new data set of which 6 had double entries (bilateral cancer patients have two entries in the database). These were all deleted as well as

106 bilateral entries from the control population. This left 2355 patients as our control group which will be compared to treated patients from Data set I with the following criteria: diagnosis year between 1989 and 1996, nonmetastatic breast cancer, complete survival time information, no male patients, no noncompliant patients and no bilateral patients. There were a total of 100 treated patients that met this criteria.

7.2 Data Checking and Preliminary Analysis

The means and standard deviations for the variables *agedx* and *dxyear* are given in Table 7.1. The average age for the control group is much higher than the treated group for this data set, however the average diagnosis year is similar for both. Tests to compare the differences in means between the two treatment groups for the variables *agedx* and *dxyear* are given in Table 7.2. The results in Table 7.2 indicate that there is a significant difference in age at diagnosis between the groups, but the means for year of diagnosis appear to be the same. Tables 7.3 through 7.26 displays the Pearson's Chi-square tests of homogeneity between the two groups for the categorical variables *staget*, *stagepn*, *bccasr*, *bccard*, *bccach*, *bccahr*, *dxgrade*, *dxer*, *dxlvn*, *dxposnod*, *behvr*, and *histcat*. Note that the treatment variables were all recoded into two categories (yes/no), except for *bccasr* which has a third category (D) which represents 'diagnostic surgery only'. There is evidence against homogeneity for the variables *stagepn*, *bccach*, *dxgrade* and *dxposnod*, and all other variables appear to be similar with respect to the treatment groups. Note that there were cells with expected counts less than 5 for *staget*, *stagepn*, *bccasr*, and *histcat*, therefore the

P-values are not precise for these variables. A summary of the number of missing observations is given in Table 7.27. The variables *staget* and *stagepn* had very few missing variables, whereas *dxgrade*, *dxer*, and *dxlvn* had substantially more.

Descriptive Statistics

	N	Mean	Std. Deviation
age at diagnosis (control)	2355	62.58	13.10
age at diagnosis (treated)	100	49.68	11.75
diagnosis year (contro)	2355	92.6556	2.2953
diagnosis year (treated)	100	92.52	2.22
Valid N (listwise)	0		

Table 7.1: Mean and standard deviation for *agedx* and *dxyear*

Paired Differences (Diff.)	Mean Diff.	Std. Error Diff.	t	df	Sig. (2-tailed)
agedx (equal σ^2 assumed)	-12.90	1.33	-9.680	2453	0.000
agedx (equal σ^2 not assumed)	-12.90	1.21	-10.693	109.705	0.000
dxyear (equal σ^2 assumed)	-.14	0.23	-0.579	2453	0.562
dxyear (equal σ^2 not assumed)	-.14	0.23	-0.597	108.162	0.552

Table 7.2: Independent Samples Test for *agedx* and *dxyear*

treated or control * T-stage Crosstabulation

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

Table 7.3: Frequency Table for *staget*

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 7.4: Chi-square test for *staget*

treated or control * Pathological N-stage Crosstabulation

			Pathological N-stage					
			No axil disect	No Nodal Mets	Axillary Nodal Mets	Fixed Nodal Mets	Mammary Nodal Mets	Total
treated or control	matched control	Count	270	1353	701	29	1	2354
		Expected Count	266.7	1336.2	721.4	28.8	1.0	2354.0
	Hoffer patient	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 7.5: Frequency Table for *stagepn*

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 7.6: Chi-square test for *stagepn*

treated or control * BCCASR Crosstabulation

			BCCASR			Total
			D	N	Y	
treated or control	matched control	Count	2	40	2313	2355
		Expected Count	1.9	39.3	2313.8	2355.0
	Hoffer patient	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 7.7: Frequency Table for *bccasr*

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 7.8: Chi-square test for *bccasr*

treated or control * BCCARD Crosstabulation

			BCCARD		
			N	Y	Total
treated or control	matched control	Count	844	1511	2355
		Expected Count	835.5	1519.5	2355.0
	Hoffer patient	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 7.9: Frequency Table for *bccard*

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 7.10: Chi-square test for *bccard*

treated or control * BCCACH Crosstabulation

			BCCACH		
			N	Y	Total
treated or control	matched control	Count	1920	435	2355
		Expected Count	1883.0	472.0	2355.0
	Hoffer patient	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 7.11: Frequency Table for *bccach*

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 7.12: Chi-square test for *bccach*

treated or control * BCCAHR Crosstabulation

			BCCAHR		
			N	Y	Total
treated or control	matched control	Count	1268	1087	2355
		Expected Count	1276.8	1078.2	2355.0
	Hoffer patient	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 7.13: Frequency Table for *bccahr*

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 7.14: Chi-square test for *bccahr*

treated or control * Tumor Grade Crosstabulation

			Tumor Grade			
			Well Differentiated	Moderately Differentiated	Poorly Differentiated	Total
treated or control	matched control	Count	360	950	543	1853
		Expected Count	359.1	937.5	556.4	1853.0
	Hoffer patient	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 7.15: Frequency Table for *dxgrade*

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 7.16: Chi-square test for *dxgrade*

treated or control * Estrogen Receptor Crosstabulation

			Estrogen Receptor		
			Negative	Positive	Total
treated or control	matched control	Count	730	1202	1932
		Expected Count	733.9	1198.1	1932.0
	Hoffer patient	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 7.17: Frequency Table for *dxer*

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 7.18: Chi-square test for *dxer*

treated or control * Invasive lvn Crosstabulation

			Invasive lvn		
			Negative	Positive	Total
treated or control	matched control	Count	1352	683	2035
		Expected Count	1344.9	690.1	2035.0
	Hoffer patient	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 7.19: Frequency Table for *dxlvn*

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 7.20: Chi-square test for *dxlvn*

treated or control * Nodal Status Crosstabulation

			Nodal Status			
			No Axillary Dissection	No pos nodes	Pos nodes	Total
treated or control	matched control	Count	265	1350	740	2355
		Expected Count	260.0	1333.4	761.7	2355.0
	Hoffer patient	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 7.21: Frequency Table for *dxposnod*

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 7.22: Chi-square test for *dxposnod*

treated or control * BEHVR Crosstabulation

			BEHVR		
			2	3	Total
treated or control	matched control	Count	141	2214	2355
		Expected Count	141.0	2214.0	2355.0
	Hoffer patient	Count	6	94	100
		Expected Count	6.0	94.0	100.0
Total		Count	147	2308	2455
		Expected Count	147.0	2308.0	2455.0

Table 7.23: Frequency Table for *behvr*

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.000[b]	1	.996		
Continuity Correction[a]	.000	1	1.000		
Likelihood Ratio	.000	1	.996		
Fisher's Exact Test				1.000	.560
Linear-by-Linear Association	.000	1	.996		
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 5.99.

Table 7.24: Chi-square test for *behvr*

treated or control * Categorical Histology Crosstabulation

			Categorical Histology			
			Ductal	Lobular	Other	Total
treated or control	matched control	Count	2030	267	58	2355
		Expected Count	2033.6	261.9	59.5	2355.0
	Hoffer patient	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 7.25: Frequency Table for *histcat*

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 7.26: Chi-square test for *histcat*

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
treated or control * T-stage	2435	99.2%	20	.8%	2455	100.0%
treated or control * Pathological N-stage	2454	100.0%	1	.0%	2455	100.0%
treated or control * BCCASR	2455	100.0%	0	.0%	2455	100.0%
treated or control * BCCARD	2455	100.0%	0	.0%	2455	100.0%
treated or control * BCCACH	2455	100.0%	0	.0%	2455	100.0%
treated or control * BCCAHR	2455	100.0%	0	.0%	2455	100.0%
treated or control * Tumor Grade	1935	78.8%	520	21.2%	2455	100.0%
treated or control * Estrogen Receptor	2014	82.0%	441	18.0%	2455	100.0%
treated or control * Invasive lvn	2126	86.6%	329	13.4%	2455	100.0%
treated or control * Nodal Status	2455	100.0%	0	.0%	2455	100.0%
treated or control * BEHVR	2455	100.0%	0	.0%	2455	100.0%
treated or control * Categorical Histology	2455	100.0%	0	.0%	2455	100.0%

Table 7.27: Summary of Missing Observations

7.3 Log Normal Survival Model

The log normal was fit using the 10 covariates considered in Chapter 5: *treated, agedx, dxyear, dxlvn, dxgrade, dxposnod, staget, histcat, dxer,* and *bccach* with outcome variable *dthcsurv* and

event variable *dthcevt*. The baseline values are the same as in Chapter 5 except for *staget*. There were only 2 subjects (both controls) in the category *staget* = 0 (see Table 7.3), therefore these subjects were dropped, and the baseline changed to *staget* = 4 (extended category). The final model using backward elimination is given in Table 7.28. Out of 2455 total entries in the study, 890 observations were dropped due to missing values, leaving 1565 observations for the analysis. There were 233 events (deaths due to breast cancer) and 1332 censored observations. The final covariates included in the model along with the *treated* variables include *agedx*, *dxlvn*, *dxgrade*, *dxposnod*, *dxer* and *staget*. The coefficient for treatment effect is $\hat{\gamma}_1 = -0.86$ as compared to the estimate for the final log normal model for Data set I, $\hat{\gamma}_1 = -0.71$ (see Table 5.6). These estimates are relatively close, which indicates that the matched pairs analysis in Chapter 5 was reasonable. The coefficients for the risk factors included in the model are consistent with diagnostic factors for breast cancer in practice. The deviance residuals and Cox-Snell residuals are given in Figures 7.1 and 7.2. Most of the deviance residuals are close to zero and the differences in the Cox-Snell residuals are not large except in the upper tail, therefore the model seems reasonable.

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Level
$\hat{\mu}$	1	8.9197459	0.43991	411.1281	0.0001	
$\hat{\sigma}$	1	1.07968658	0.053631			
treated ($\hat{\gamma}_1$)	1	-0.8602788	0.175854	23.93181	0.0001	1
agedx ($\hat{\gamma}_2$)	1	-0.0077863	0.003588	4.71013	0.0300	
dxlvn ($\hat{\gamma}_3$)		-0.5298169	0.102741	26.59297	0.0001	1
dxgrade	2			30.91413	0.0001	
($\hat{\gamma}_4$)	1	-0.608058	0.189435	10.30316	0.0013	2
($\hat{\gamma}_5$)	1	-0.9873728	0.194756	25.70285	0.0001	3
dxposnod	2			31.5684	0.0001	
($\hat{\gamma}_6$)	1	-0.0066961	0.197677	0.001147	0.9730	1
($\hat{\gamma}7$)	1	0.5854128	0.202296	8.374349	0.0038	0
staget	4			17.70794	0.0014	
($\hat{\gamma}_8$)	1	0.62107936	0.323639	3.682741	0.0550	3
($\hat{\gamma}_9$)	1	0.38985311	0.247124	2.488691	0.1147	2
($\hat{\gamma}_{10}$)	1	0.74273336	0.247486	9.006687	0.0027	1
($\hat{\gamma}_{11}$)	1	0.73774278	0.694859	1.127239	0.2884	-1
dxer ($\hat{\gamma}_{12}$)	1	0.42648009	0.097412	19.16785	0.0001	2

Table 7.28: Final Log normal Model for Data set II

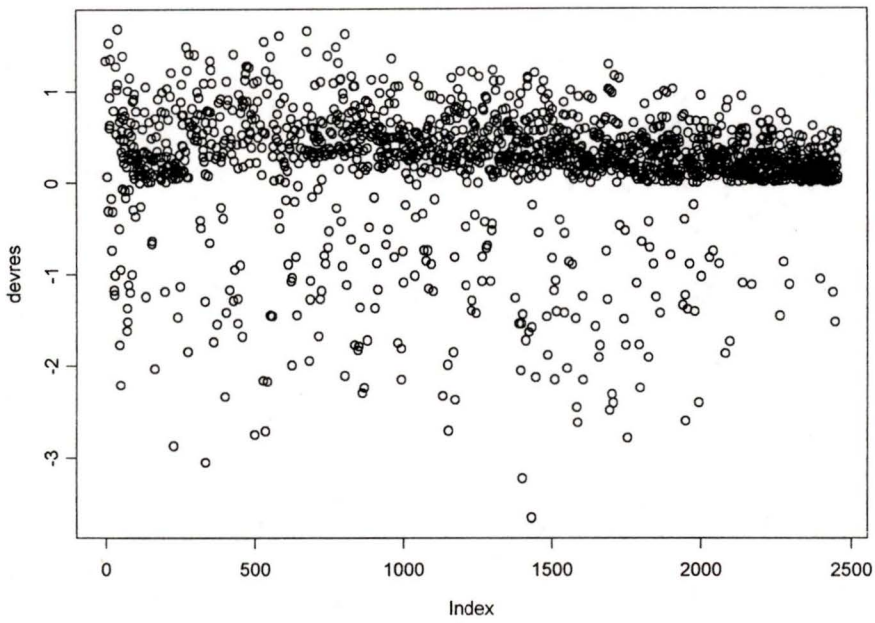

Figure 7.1: Deviance Residuals for Final Log Normal Model for Data set II

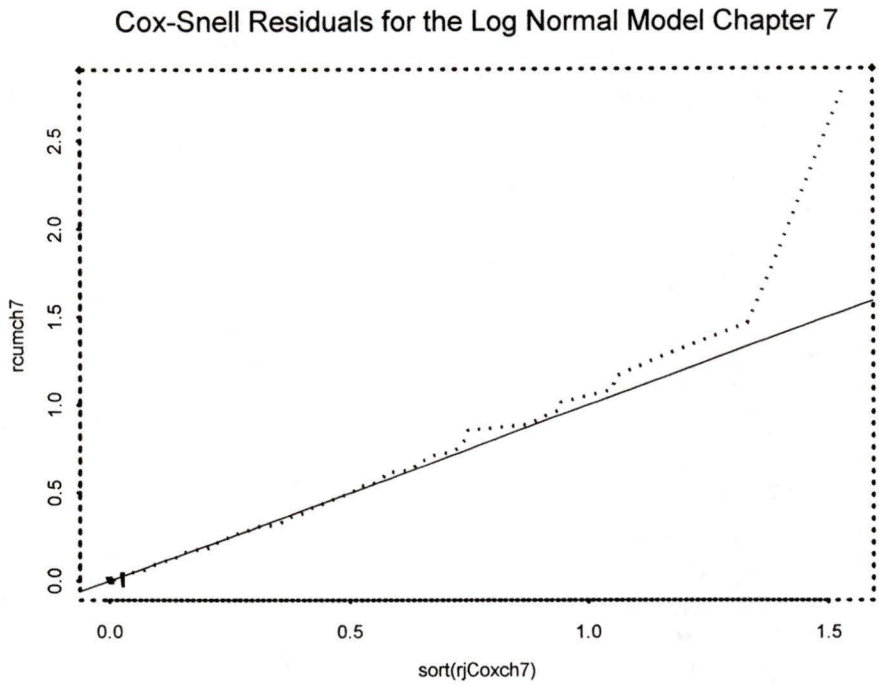

Figure 7.2: Hazard Plot of Cox-Snell Residuals vs. $\hat{\Lambda}_{N-A}(r_j)$

Plots of the Nelson-Aalen estimator of the cumulative hazard $\hat{\Lambda}_{N-A}(x)$ are given in Figures 7.3 through 7.8, for the final covariates *dxlvn*, *dxgrade*, *dxposnod*, *staget*, *dxer* and categorical age at diagnosis, *age4bin*. The continuous variable age at diagnosis was separated into the following categories: 25 - 40, 40 - 50, 50 - 60, and 60+. It should be noted that there are many more older subjects in Data set II, with 150 subjects in the first age group, 423 in the age range 40 - 50, 478 between 50 - 60, and 1404 over 60 years of age. The average age at diagnosis for the control group was much higher than that of the treated group (see Table 7.1), therefore most of the older patients are controls. The slope of the Nelson-Aalen estimator in these figures suggests

that the hazards follow the shape of the log normal hazard given the risk factors considered in the model. In Figure 7.3, the youngest age group is showing the highest combined death and relapse rate. A possible explanation for this is that younger women who develop breast cancer are usually patients with a family history of breast disease. They may have a more aggressive form of cancer due to their genetic disposition as well as a higher exposure to estrogen which may lead to more recurrences and more deaths from breast cancer.

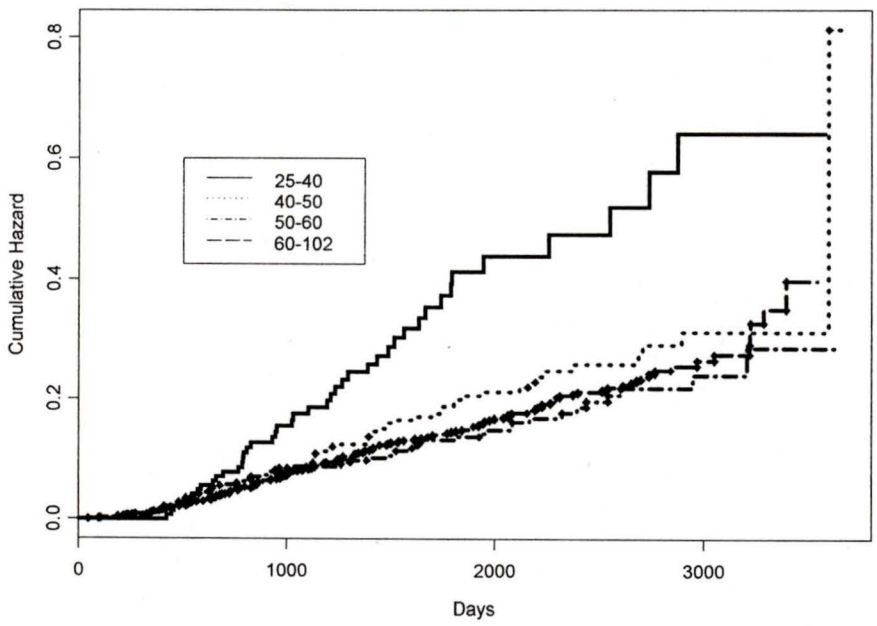

Figure 7.3: N-A Estimator for *age4bin*

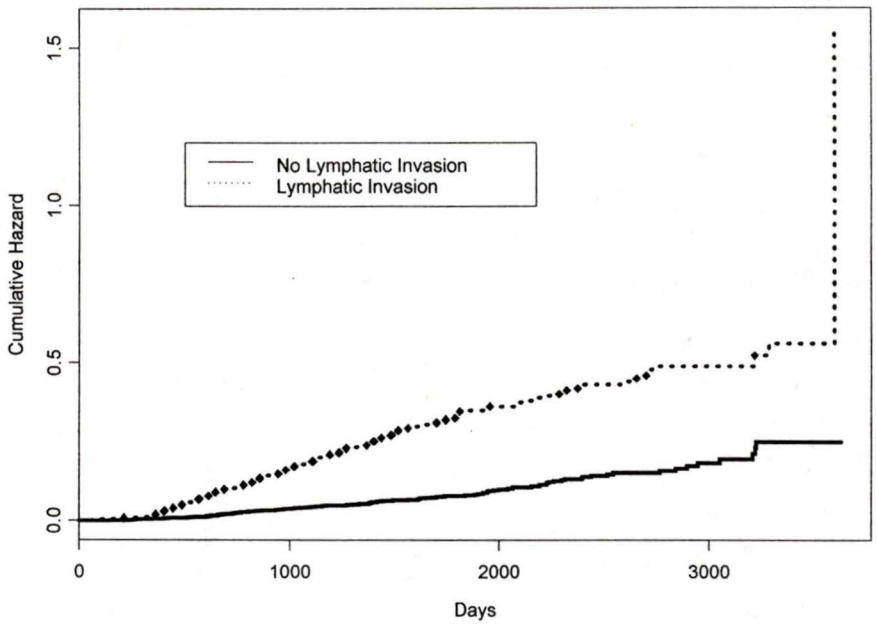

Figure 7.4: N-A Estimator for *dxlvn*

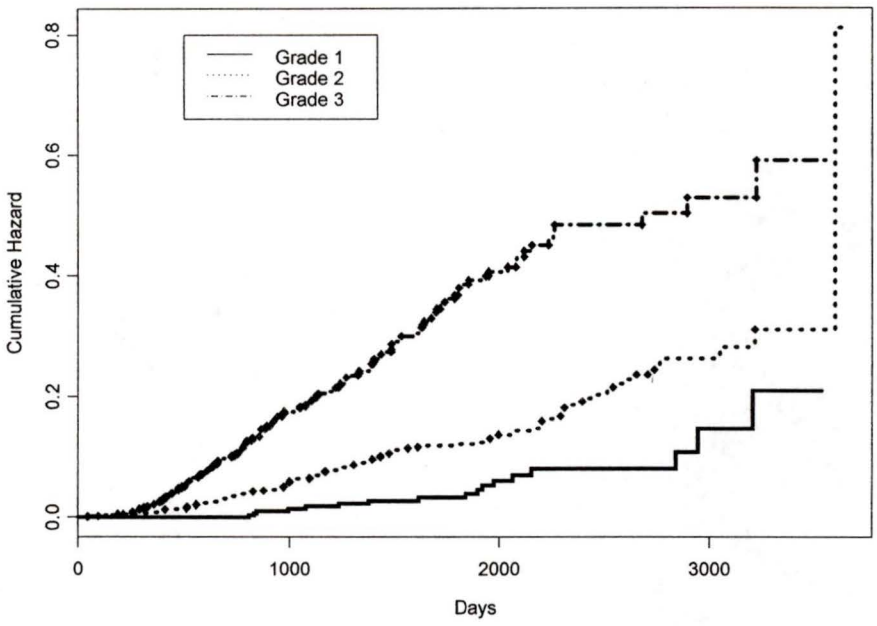

Figure 7.5: N-A Estimator for *dxgrade*

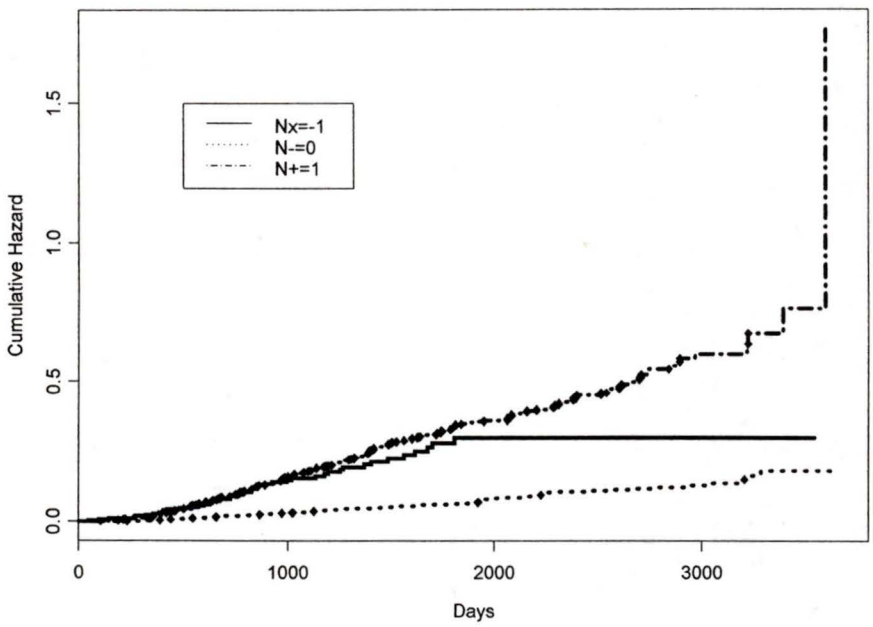

Figure 7.6: N-A Estimator for *dxposnod*

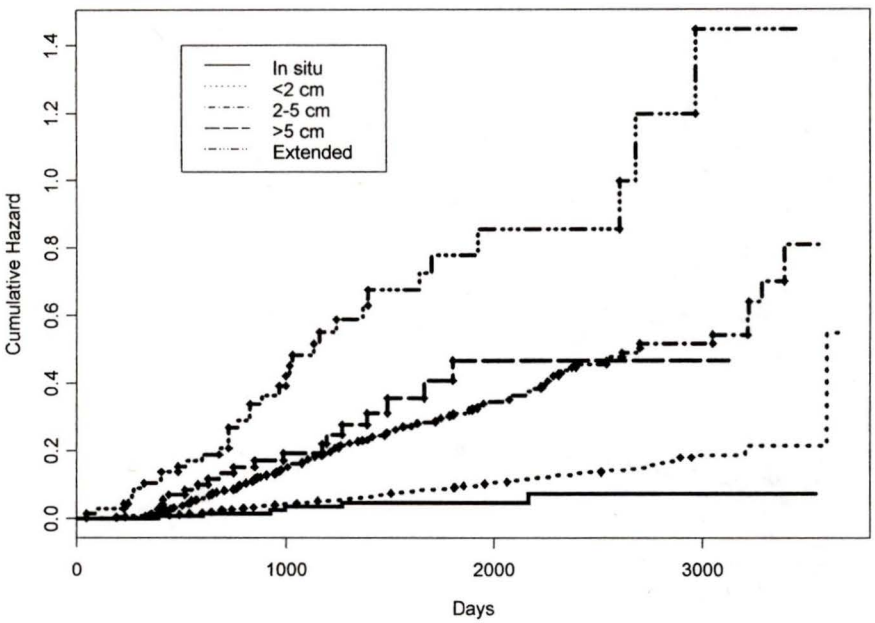

Figure 7.7: N-A Estimator for *staget*

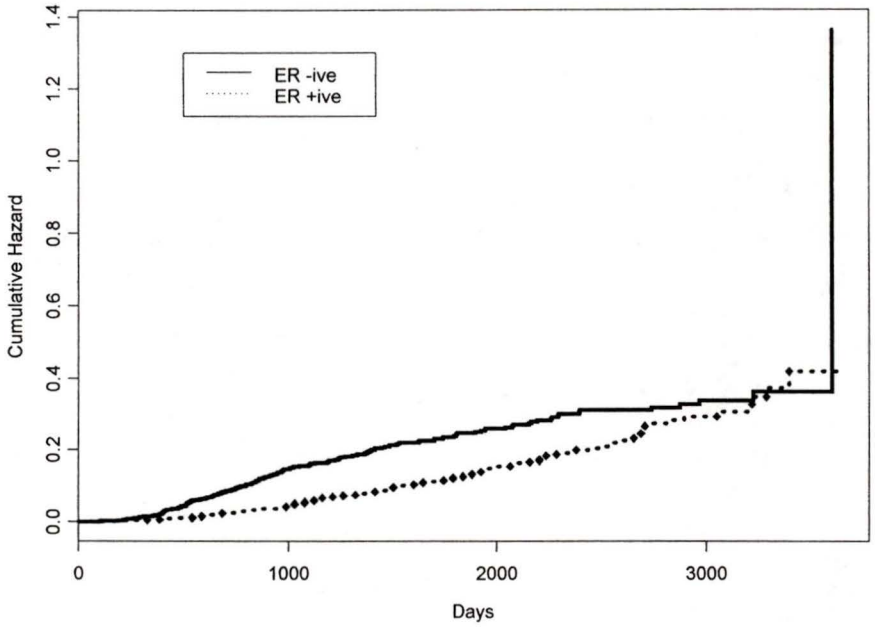

Figure 7.8: N-A Estimator for *dxer*

Chapter 8

Conclusions

Based on our results, the women with nonmetastatic breast cancer who took Dr. Abram Hoffer's vitamin and mineral regime have significantly decreased survival rates than the control group with respect to disease-specific survival. In terms of systemic relapse, there was no significant difference in the treament groups with respect to disease-free survival.

We have examined the effects of all available standard prognostic factors for breast cancer. We did not have information about tumour progesterone status, accurate delineation of the therapeutic modalities (radiotherapy, chemotherapy, hormonal therapy) as occurring in the adjuvant and metastatic setting, nor information on the diet and body weight of patients. In terms of follow-up, it is important to know how well the patients were followed after diagnosis, especially when considering model selection. The choice of a parametric model is dependent on the hazard of an event (recurrence or death) for a particular cancer with the length of follow-up for the study {23}. In general, with longer follow-up (more events), the results of the analysis would be more accurate. In data set I, 117 out of 176 subjects were censored, (see Table 6.1

and 6.2), which indicates there was heavy censoring for this data, therefore a longer follow-up may have allowed for more reliable modelling results.

In terms of diagnosis year for the matched-pair study, worse survival was seen in later years. A possible explanation for this is that the referral pattern to the BCCA for early diagnosis years has changed as compared to later years, possibly due to the growing number of medical oncologists in the community {28}. For newly diagnosed breast cancer patients, not all cases are referred to the BCCA (approximate referral rate only 74%) {30}. Of those who are referred to the BCCA, they tend to have higher risk histology (ie, worst cases only referred in later years) {30}. Therefore, the patients included in the matched-pair study in the later years may have been a subgroup of the BCCA-referred cases.

When dealing with an observational study where the two groups were not randomly assigned, the groups may not be comparable. There may be other lurking variables which may be causing failures. The treated group may have been self-medicating with additional vitamins and minerals or with other alternative treatments, since they appeared to be unsatisfied with conventional cancer therapy alone. The doses administered by Dr. Hoffer were extremely large as compared to the RDA and since it is still uncertain whether megadoses of vitamins and minerals may be toxic or even fatal in large doses, a combination of conventional cancer therapy with the Hoffer treatment as well as other unknown nonconventional treatments may be harmful. Therefore, more evidence is necessary to conclude that vitamin and minerals administered to nonmetastatic breast cancer patients is associated with decreased survival. In an observational study it is nearly

impossible to control for all confounders. Stratification or matching, used in combination with randomization would increase the similarity between the treated and control groups.

Only fixed time covariates were considered in this analyis. A time-dependent covariate to represent the time on the Hoffer treatment could not be used as parametric modelling methods in the statistical software packages currently do not offer this option. Therefore, the treatment effect may not be accurate as most of the patients who were given the treatment did not start the Hoffer regime at the time of diagnosis.

The log normal model has been shown to be clinically reasonable for modelling breast cancer survival data, particularly when there is strong evidence against proportional hazards {23}. This model provided adequate results for both datasets considered. Although there appears to be no apparent benefit from the use of mega-doses of vitamins and minerals, further study using accelerated failure-time models and more controlled experimental design would be advisable.

Bibliography

[1] Boik, J. 1996. Cancer and Natural Medicine: A Textbook of Basic Science and Clinical Research, Princeton, Minnesota: Oregon Medical Press, pp. 142-147; 233-235.

[2] Benode, L., Howard, T. and Buck, D. 1969. Synergistic Killing of Ehrlich Ascites Carcinoma Cells by Ascorbate and 3 - Amino - 1,2,4, - triazole. Oncology 23(1), pp. 33-43.

[3] Bendich, A. and Langseth, L. 1995. The Health Effects of Vitamin C Supplementation: A Review. Journal of the American College of Nutrition 14 (2), pp. 124-136.

[4] Cameron, E. and Pauling, L. 1993. Cancer and Vitamin C. Philadelphia: Camino Books, Inc., Revised and Expanded, pp. xv-xxiv; 149, 160, 162-163, 171, 217-255.

[5] Frei B., England, L. and Ames, B.N. 1989. Ascorbate is an Outstanding Antioxidant in Human Blood Plasma. Proceedings of the National Academy of Sciences USA 86(16), pp. 6377-6381.

[6] Frey, J.R., Peck, R. and Bollag, W. 1991. Antiproliferative Activity of Tetinoids, Interferon x and their Combination in Five Human Transformed Cell Lines. Cancer Letters 57, pp. 223-227.

[7] Fugh-Berman, A. 1996. Dietary Supplements and Nutrition. Alternative Medicine: what works, Tucson, Arizona. Odonian Press, pp.61-93.

[8] Gershoff, S.N. 1993. Vitamin C (Ascorbic Acid): New Roles, New Requirements? Nutrition Reviews 51(11), pp. 313-326.

[9] Lerner, M. 1994. Choices in Healing. Cambridge, Massachusetts: The MIT Press, pp. 195-259; 353; 397-404.

[10] Noto, V., Taper, H.S., Yi-hua, J., Janssens, J., Bonte, J. and De Loecker, W. 1989. Effects of Sodium Ascorbate (Vitamin C) and 2-Methyl-1,4-Naphthoquinone (Vitamin K3) Treatment on Human Tumor Cell Growth in Vitro. Cancer 63(5), pp.901-906.

Vitamin C. Life Sciences 27(4), pp. 275-280.

[11] Riordan, N.H., Riordan, H.D., Meng, X., Li, Y. and Jackson, J.A. 1995. Intravenous Ascorbate as a Tumor Cytotoxic Chemotherapeutic Agent. Medical Hypotheses 44, pp. 207-213..

[12] Roberts, H.J. 1981. Perspective on Vitamin E as Therapy. Journal of the American Medical Association 246(2), pp. 129-131.

[13] Rustin, G.J.S.. 1993. Therapy of Solid Tumour with Retinoids, Monotherapy and Combination Therapy. European Journal of Cancer 29A (Suppl 6 Abstract 202), p. S42.

[14] Stahelin, H.B. 1993. Critical Reappraisal of Vitamins and Trace Minerals in Nutritional Support of Cancer Patients. Supportive Care in Cancer 1, pp. 295-297.

[15] Weisburger, J.H. 1991. Nutritional Approach to Cancer Prevention with Emphasis on Vitamins, Antioxidants, and Carotenoids. American Journal of Clinical Nutrition 53(1), pp. 226S-237S

[16] Prasad, K.N., Kumar, A., Kochupillai, V. and Cole, W.C. 1999. High Doses of Multiple Antioxidant Vitamins: Essential Ingredients in Improving the Efficacy of Standard Cancer Therapy. Journal of the American College of Nutrition 18(1), pp. 13-25.

[17] Olivotto, Ivo A. 1996. The Intelligent Patient Guide to Breast Cancer. Intelligent Patient Guide Ltd.

[18] http://onhealth.com/ch1/in-depth/item/item, 2280.

[19] http://onhealth.com/ch1/in-depth/item/item, 1974.

[20] Percy, C., Van Holten V. and Muir, C. 1990. International Classification of Diseases for Oncology, Second Edition. World Health Organization (ICD-0).

[21] Chapman, J.W., Mobbs, B.G., Hanna, W., Kahn, H.J., Murray, D., Lickley, H.L.A., Fish, E.B., Trudeau, M.E., Miller, N.A. and McCready, D.R. (2000). A Changing Role for Steriod Hormone Receptors in Primary Breast Cancer. Annals of Surgical Oncology 7, pp. 416-426.

[22] Fish, E.B., Chapman, J.W. and Link, M.A. 1998. Competing Causes of Death for Primary Breast Cancer. Annals of Surgical Oncology 5(4), pp. 368-375.

[23] McCready, D.R., Chapman, J.W., Hanna, W.M., Kahn, H.J., Murray, D., Fish, E.B., Trudeau, M.E., Andrulis, I.L., Lavina, H. and Lickley, A. (in press). Factors Affecting Distant Disease-free Survival for Primary Invasive Breast Cancer: use of a Log-normal Survival Model.

[24] Chapman, J.W., Wolman, E., Wolman, S.R., Remvikos, Y., Shackney, S., Axelrod, D.E., Baisch, H., Christensen, I.J., White, R.A., Liebovitch, L.S., Moore, D.H., Waldman, F.M.,

Cornelisse, C.J. and Shankey, T.V. 1998. Assessing Genetic Markers of Tumour Progression in the Context of Intratumour Heterogeneity. Cytometry 31, pp. 67-73.

[25] Kalbfleisch, J.D. and Prentice, R.L. 1980. The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Mathematical Statistics.

[26] Klein, J.P. and Moeschberger, M.L. 1997. Survival Analysis: Techniques for Censored and Truncated Data. Springer.

[27] Mates, Donna 1999. Personal Communication, BC Cancer Agency, Vancouver.

[28] Speers, Caroline 2000. Personal Communications, BC Cancer Agency, Vancouver.

[29] Duchesne, T. 1999. Multiple Time Scales in Survival Analysis. Doctoral Thesis, University of Waterloo.

[30] Olivotto, Ivo A., Mates, D., Kan, L., Fung, J., Samant, R. and Warren Burhenne, L.J. 1999. Prognosis, treatment, and recurrence of breast cancer for women attending or not attending the Screening Mammography Program of British Colombia. Breast Cancer Research and Treatment 54, pp. 73-81.

[31] Olivotto, Ivo A., Coldman, A.J., Hislop, G., Trevison C.H., Kula J., Goel, V. and Sawka, C. 1997. Compliance With Practice Guidelines for Node-Negative Breast Cancer. Journal of Clinical Oncology 15(1), pp. 216-222.

[32] Ontario Dietetic Association 1989. Nutritional Care Manual. The Ontario Hospital Association.

[33] Committee for the Revision of the Dietary Standard for Canada 1983. Recommended Nutrient Intakes for Canadians. Minister of National Health and Welfare.

Appendix A

Computer codes

A.1 SAS codes

- Fit Weibull, Log logistic and Log normal Parametric Survival Models (Full Model: first step of Stepwise Procedure)

```
libname nat 'c:\SAS\mydoc';
proc lifereg data=nat.par87db order=data;
class treated dxlvnp dxgradep dxposnop stagetp histcat dxer bccach;
model dthcsurp*dthcevp(0,-1)=treated agedxp dxyearp dxlvnp dxgradep dxposnop
stagetp histcat dxer bccach;
model dthcsurp*dthcevp(0,-1)=treated agedxp dxyearp dxlvnp dxgradep dxposnop
stagetp histcat dxer bccach/dist=llogistic;
model dthcsurp*dthcevp(0,-1)=treated agedxp dxyearp dxlvnp dxgradep dxposnop
stagetp histcat dxer bccach/dist=lnormal;
run;
```

- Log-linear Survival Models: Levels and Information

```
Data Set =NAT.PAR87DB
Dependent Variable=Log(DTHCSURP)
Censoring Variable=DTHCEVP
Censoring Value(s)= 0 -1
Noncensored Values= 40 Right Censored Values= 86
Left Censored Values= 0 Interval Censored Values= 0
For A: Log Likelihood for WEIBULL -72.1956385
For B: Log Likelihood for LLOGISTC -75.1679328
For C: Log Likelihood for LNORMAL -76.31481285
```

- Fit Final Models in SAS

```
libname nat 'c:\SAS\mydoc';
proc lifereg data=nat.par87db order =data;
class treated dxlvnp dxgradep dxposnop stagetp histcat dxer bccach;
model dthcsurp*dthcevp(0,-1)=treated dxyearp stagetp;
model dthcsurp*dthcevp(0,-1)=treated dxyearp dxgradep stagetp/dist=llogistic;
model dthcsurp*dthcevp(0,-1)=treated dxyearp dxgradep stagetp/dist=lnormal;
run;
```

- Fit Log-linear Survival Models for Paired Data

```
libname nat 'c:\SAS\mydoc';
proc lifereg data=nat.par87db order=data;
class treated pairp;
A: model dthcsurp*dthcevp(0,-1)=treated pairp;
B: model dthcsurp*dthcevp(0,-1)=treated pairp/dist=llogistic;
C: model dthcsurp*dthcevp(0,-1)=treated pairp/dist=lnormal;
run;
```

- Paired Models: Levels and Information

```
Lifereg Procedure
Class Level Information
Number of observations used = 168
Data Set =NAT.PAR87DB
Dependent Variable=Log(DTHCSURP)
Censoring Variable=DTHCEVP
Censoring Value(s)= 0 -1
Noncensored Values= 58 Right Censored Values= 110
Left Censored Values= 0 Interval Censored Values= 0
For A: Log Likelihood for WEIBULL -53.41198871
For B: Log Likelihood for LLOGISTC -57.97940037
For C: Log Likelihood for LNORMAL -58.24068871
```

Class	Levels	Values
TREATED	2	1 0
PAIRP	84	26 40 43 44 45 56 67 69 72 73 75 77 78 82 85 87 88 92 93 96 97 98 100 103 104 107 108 109 111 112 113 114 115 117 118 119 121 124 125 126 132 133 136 137 138 139 140 141 142 144 146 147 149 150 151 152 153 155 157 160 161 162 165 168 169 174 175 178 182 192 194 198 201 202 205 208 210 218 220 222 224 226 228 234

- SAS Code to Check Appropriateness of the Weibull model

```
proc lifereg data=nat.par87db order =data;
class treated dxlvnp dxgradep dxposnop stagetp histcat dxer bccach;
model dthcsurp*dthcevp(0,-1)=treated dxyearp stagetp/dist=gamma
shape1=1 noshape1;
run;
```

- SAS Code to Check Appropriateness of the Log normal model

```
proc lifereg data=nat.par87db order =data;
class treated dxlvnp dxgradep dxposnop stagetp histcat dxer bccach;
model dthcsurp*dthcevp(0,-1)=treated dxyearp dxgradep stagetp/dist=gamma
shape1=0 noshape1;
run;
```

- Fit Log Normal Model for *syssurv1*

```
libname nat 'c:\SAS\mydoc';
proc lifereg data=nat.par84sys order = data;
class treateds dxlvn dxgrade dxposnod staget histcat dxer bccach behvr;
model syssurv1*sysevent(0)=treateds agedx dxyear dxlvn dxgrade dxposnod
staget histcat dxer bccach behvr/dist=lnormal;
run;
```

- Fit Log Normal Model for Dataset II

```
libname nat 'c:\SAS\mydoc';
proc lifereg data=nat.parch7 order = data;
class treated dxlvn dxgrade dxposnod staget histcat dxer bccach;
model dthcsurv*dthcevt(0,-1)=treated agedx dxyear dxlvn dxgrade dxposnod
staget histcat dxer bccach/dist=lnormal;
run;
```

A.2 S-PLUS Codes

- Deviance Residuals

A) For Weibull:
>resfinalwb1_residuals.survReg(finalwb1,type='deviance')
>plot(resfinalwb1)
B) For Log logistic:
>resfinallog_residuals.survReg(finallog,type='deviance')
>plot(resfinallog)
C) For Log Normal:
>resfinallgnm_residuals.survReg(finallgnm, type='deviance')
>plot(resfinallgnm)
D) For Paired Log normal:
>resprlgnm_residuals.survReg(pairlgnm,type='deviance')
>plot(resprlgnm)
E) For Log Normal (Time to Systemic Relapse)
>resch6_residuals.survReg(finalch6new,type='deviance')
>plot(resch6)
F) For Log Normal (Dataset II):
>devres_residuals.survReg(final2455, type='deviance')
>plot(devres)

- Cox-Snell Residuals

A) For Log Normal Chapter 5:
Coxsnell_data.frame(Coxsnelllgnm)
finallgnm.pred_predict(finallgnm,type='lp')
Coxsnell_data.frame(Coxsnell, finallgnm.pred)
timerj_para87.dat$DTHCSURP
censorj_para87.dat$CEVTP
sigmalgnm_ 0.8802737
rjlgnm_-log(1-pnorm((log(timerj)-finallgnm.pred)/sigmalgnm))
rcumlgnm_-log(summary(survfit(Surv(rjlgnm,censorj),na.action=na.exclude,
type="fh"))$surv)
rjplot_(1:168)[1==Coxsnell$V5]
rjCox_rjlgnm[rjplot]
plot(sort(rjCox),rcumlgnm, type="l", lty=2, lwd=4)
abline(0,1)
title("Cox-Snell Residuals for the Log Normal Model")
B) For Log Normal Chapter 6:
ch6.pred_predict(finalch6new,type='lp')
Coxsnellch6new_data.frame(Coxsnellch6new,ch6.pred)
systimerj_Coxsnellch6new$SYSSURV

syscensorj_Coxsnellch6new$SYSEVENT
sigmach6_1.22893437
rjch6_-log(1-pnorm((log(systimerj)-ch6.pred)/sigmach6))
rcumch6_-log(summary(survfit(Surv(rjch6,syscensorj),na.action=na.exclude,
type='fh'))$surv)
rjplotch6_(1:168)[1==Coxsnellch6new$SYSEVENT]
rjCoxch6_rjch6[rjplotch6]
plot(sort(rjCoxch6),rcumch6, type="l",lty=2,lwd=4)
abline(0,1)
title('Cox-Snell Residuals for the Log Normal Model Chapter 6")
C) For Log Normal Chapter 7:
final2455.pred_predict(final2455,type='lp')
CoxSnell2455_data.frame(CoxSnell2455,final2455.pred)
timerjch7_CoxSnell2455$DTHCSURV
censorjch7_CoxSnell2455$DTHCEVT
sigmach7_1.07968658
rjch7_-log(1-pnorm((log(timerjch7)-final2455.pred)/sigmach7))
rcumch7_-log(summary(survfit(Surv(rjch7,censorjch7),na.action=na.exclude,
type='fh'))$surv)
rjplotch7_(1:2455)[1==CoxSnell2455$V5]
rjCoxch7_rjch7[rjplotch7]
plot(sort(rjCoxch7),rcumch7, type="l", lty=2, lwd=4)
abline(0,1)
title("Cox-Snell Residuals for the Log Normal Model Chapter 7")

- Kaplan-Meier Curves for Treatment Groups

>KMtrmt_survfit(formula = Surv(DTHCSURP, CEVTP, type = "right")
~TREATED,data =para87.dat1, na.action = na.exclude,
conf.int = 0.95, se.fit = T, type= "kaplan-meier", error = "greenwood",
conf.type ="log", conf.lower ="usual")
>legend(500,0.2,c("Control Group","Treated Group"),lty=2:3)
>title("Kaplan-Meier Survival Probability for Treatment Groups")

- Kaplan-Meier Curves for Tumour Size

>KMstage_survfit(formula = Surv(DTHCSURP, CEVTP, type = "right")
~STAGET, data = para87.dat1, na.action = na.exclude,
conf.int =0.95, se.fit = T, type ="kaplan-meier",error = "greenwood",
conf.type = "log", conf.lower = "usual")

>legend(500,0.2,c(">5 cm","2-5 cm","Expansion <2cm"),lty=1:3)
>title("Kaplan-Meier Survival Probability for Tumour Size")

- Kaplan-Meier Curves for Original Treatment Groups Chapter 6

>KM.oldtrmt_survfit(formula = Surv(SYSSURV1, SYSEVENT, type = "right")
~TREATED,data = systreateds, na.action = na.exclude,
conf.int = 0.95,se.fit = T, type= "kaplan-meier", error = "greenwood",
conf.type = "log", conf.lower = "usual")
>legend(200,0.2,c("Control Group","Treated Group"),lty=1:2)
>title("Kaplan-Meier Survival Probability for Original Treatment Groups")

- Kaplan-Meier Curves for New Treatment Groups Chapter 6

>KM.newtrmt_survfit(formula = Surv(SYSSURV1, SYSEVENT, type = "right")
~TREATEDS,data = systreateds,na.action = na.exclude,
conf.int = 0.95, se.fit = T, type = "kaplan-meier",error = "greenwood",
conf.type = "log", conf.lower = "usual")
>legend(200,0.2,c("Control Group","Treated Group"),lty=1:2)
>title("Kaplan-Meier Survival Probability for New Treatment Groups")

- Chapter 5 N-A Hazard Plots:

>NA.agedx_survfit(Surv(DTHCSURP,CEVTP)~AGE4BIN, para87.dat,
type='fleming')
>plot(NA.agedx, mark.time=FALSE, lty=1:4, lwd=4, fun="cumhaz",
xlab="Days", ylab="Cumulative Hazard")
>legend(500,1.2,c("28-40","40-50","50-58","58-88"), lty=1:4)
>NA.dxlvnp_survfit(Surv(DTHCSURP,CEVTP)~DXLVNP, para87.dat,
na.action = na.exclude, type='fleming')
>plot(NA.dxlvnp, mark.time=FALSE, lty=1:2, lwd=4, fun="cumhaz",
xlab="Days", ylab="Cumulative Hazard")
>legend(500,1.2, c("No lymphatic Invasion","Lymphatic Invasion"), lty=1:2)
>NA.dxgradep_survfit(Surv(DTHCSURP,CEVTP)~DXGRADEP, para87.dat,
na.action = na.exclude,type='fleming')
>plot(NA.dxgradep, mark.time=FALSE, lty=1:3, lwd=4, fun="cumhaz",
xlab="Days",ylab="Cumulative Hazard")
>legend(500,1.0,c("Grade 1","Grade 2","Grade 3"),lty=1:3)
>NA.dxposnop_survfit(Surv(DTHCSURP,CEVTP)~DXPOSNOP, para87.dat,
na.action = na.exclude,type='fleming')
>plot(NA.dxposnop, mark.time=FALSE, lty=1:3, lwd=4, fun="cumhaz",

xlab="Days", ylab="Cumulative Hazard")
>legend(500,1.2,c("Nx = -1","N- = 0","N+ = 1"),lty=1:3)
>NA.stagetp_survfit(Surv(DTHCSURP,CEVTP)~STAGETP, para87.dat,
na.action = na.exclude,type='fleming')
>plot(NA.stagetp, mark.time=FALSE,lty=1:3, lwd=4, fun="cumhaz",
xlab="Days", ylab="Cumulative Hazard")
>legend(500,1.5, c("<2 cm","2-5 cm",">5 cm"),lty=1:3)
>NA.histcat_survfit(Surv(DTHCSURP,CEVTP)~histcat, para87.dat,
na.action=na.exclude,type='fleming')
>plot(NA.histcat, mark.time=FALSE, lty=1:3, lwd=4, fun="cumhaz",
lab="Days", ylab="Cumulative Hazard")
>legend(500,1.0,c("Ductal","Other", "Lobular"), lty=1:3)
>NA.dxer_survfit(Surv(DTHCSURP,CEVTP)~DXER, para87.dat,
na.action = na.exclude,type='fleming')
>plot(NA.dxer, mark.time=FALSE, lty=1:2, lwd=4, fun="cumhaz",
xlab="Days", ylab="Cumulative Hazard")
>legend(500,1.2, c("ER -ive","ER +ive"), lty=1:2)
>NA.bccach_survfit(Surv(DTHCSURP,CEVTP)~BCCACH, para87.dat,
na.action = na.exclude,type='fleming')
>plot(NA.bccach, mark.time=FALSE, lty=1:2, lwd=4, fun="cumhaz",
xlab="Days", ylab="Cumulative Hazard")
>legend(500,1.2, c("No Chemotherapy=N","Chemotherapy=Y"), lty=1:2)
>NA.dxyear_survfit(Surv(DTHCSURV,CEVT)~DXYEAR3, parch6,
na.action=na.exclude, type='fleming')
>plot(NA.dxyear,mark.time=FALSE,lty=1:2, lwd=4, fun="cumhaz",
xlab="Days", ylab="Cumulative Hazard")
>legend(500,1,c("89-92","93-96"),lty=1:2)

- Chapter 7 N-A Hazard Plots:

>NAch7.agedx_survfit(Surv(DTHCSURV,DTHCEVT)~age4bin, par2455ch7,
na.action=na.exclude,type='fleming')
>plot(NAch7.agedx,mark.time=FALSE,lty=1:4,lwd=4,fun='cumhaz',
xlab="Days",ylab="Cumulative Hazard")
>legend(500,0.6,c("25-40","40-50","50-60","60-102"),lty=1:4)
>NAch7.dxlvn_survfit(Surv(DTHCSURV,DTHCEVT)~DXLVN, par2455ch7,
na.action=na.exclude,type='fleming')
>plot(NAch7.dxlvn,mark.time=FALSE,lty=1:2,lwd=4,fun='cumhaz',
xlab="Days",ylab="Cumulative Hazard")
>legend(500,1.2,c("No Lymphatic Invasion","Lymphatic Invasion"),lty=1:2)

```
>NAch7.dxgrade_survfit(Surv(DTHCSURV,DTHCEVT)~DXGRADE,par2455ch7,
na.action=na.exclude,type='fleming')
>plot(NAch7.dxgrade,mark.time=FALSE,lty=1:3,lwd=4,fun='cumhaz',
xlab="Days",ylab="Cumulative" Hazard")
>legend(500,0.8,c("Grade 1","Grade 2", "Grade 3"),lty=1:3)
>NAch7.dxposnod_survfit(Surv(DTHCSURV,DTHCEVT)~DXPOSNOD,par2455ch7,
na.action=na.exclude,type='fleming')
>plot(NAch7.dxposnod,mark.time=FALSE,lty=1:3,lwd=4,fun='cumhaz',
xlab="Days",ylab="Cumulative Hazard")
>legend(500,1.5,c("Nx=-1","N-=0", "N+=1"),lty=1:3)
>NAch7.staget_survfit(Surv(DTHCSURV,DTHCEVT)~STAGET,par2455ch7,
na.action=na.exclude,type='fleming')
>plot(NAch7.staget,mark.time=FALSE,lty=1:5,lwd=4,fun='cumhaz',
xlab="Days",ylab="Cumulative Hazard")
>legend(500,1.4,c("In situ", "<2 cm", "2-5 cm",">5 cm","Extended"),lty=1:5)
>NAch7.dxer_survfit(Surv(DTHCSURV,DTHCEVT)~DXER,par2455ch7,
na.action=na.exclude,type='fleming')
>plot(NAch7.dxer,mark.time=FALSE,lty=1:2,lwd=4,fun='cumhaz',
xlab="Days",ylab="Cumulative Hazard")
>legend(500,1.3,c("ER -ive", "ER +ive"),lty=1:2)
```

- Extract Hoffer patients in Dataset II

```
>Hoffer
function(a, b) {
n <- length(a)
m <- length(b)
d <- c(rep(0, n))
for(i in 1:m)
d <- d |(a == c(rep(b[i], n)))
d.mat <- matrix(as.numeric(d), nrow = n, byrow = T)
hofpatient <- (1:2560)[1 == d.mat]
(hofpatient)}
>Hoffer(a,b)
[1] 41 87 122 221 300 318 328 389 427 431 451 473 476 484 491
[16] 513 539 567 606 613 631 632 637 652 687 701 734 746 757 776
[31] 786 801 826 831 928 929 1085 1101 1150 1151 1154 1216 1232 1245 1255
[46] 1257 1258 1259 1331 1359 1398 1405 1406 1415 1449 1460 1472 1529 1544 1545
[61] 1566 1587 1609 1610 1622 1697 1819 1847 1868 1889 1892 1907 1911 1943 1951
[76] 1953 1964 2012 2034 2055 2093 2103 2108 2109 2121 2143 2194 2211 2221 2335
```

[91] 2336 2337 2358 2439 2453 2478 2516 2531 2535

- Identify all Bilateral Patients in Dataset II (bilateral patients count twice in the data set)

```
>Bilateralcheck
function(a) {
n <- length(a)
d <- c(rep(0, n))
for(j in 1:n) {
d[j] <- d[j] + as.numeric(a[j] == a[j - 1])
d[j] <- d[j] + as.numeric(a[j] == a[j + 1])
d.mat <- matrix(d, nrow = n)}
bilpatient <- (1:2560)[1 == d.mat]
bilpatient}
>Bilateralcheck(a)
[1] NA 77 78 99 100 177 178 188 189 258 259 303 304 308
[15] 309 356 357 382 383 391 392 409 410 489 490 497 498 500
[29] 501 561 562 564 565 604 605 631 632 668 669 698 699 760
[43] 761 814 815 857 858 860 861 905 906 923 924 932 933 937
[57] 938 954 955 961 962 973 974 1131 1132 1150 1151 1212 1213 1257
[71] 1258 1263 1264 1267 1268 1302 1303 1327 1328 1544 1545 1609 1610 1619
[85] 1620 1640 1641 1712 1713 1781 1782 1840 1841 1850 1851 1875 1876 1903
[99] 1904 1981 1982 2027 2028 2075 2076 2191 2192 2325 2326 2335 2336 2348
[113] 2349 2361 2362 2406 2407 2475 2476 NA
```

A.3 SPSS Codes

- Crosstab for *histcat* vs. *hist1, hist2, hist3*

```
CROSSTABS
/TABLES=hist1 BY histcat
/FORMAT= AVALUE TABLES
/C
```

Vita

Surname:	Forde
Given Names:	Natalie Judith
Place of Birth:	Toronto, Ontario

EDUCATIONAL INSTITUTIONS ATTENDED

University of Victoria, Canada	1998-2000
University of Western Ontario, Canada	1997
York University, Canada	1991-1995

DEGREES AWARDED

B. A. (Mathematics) York University	1995

HONOURS AND AWARDS

York Universtiy In-Course Scholarship	1993
University of Victoria Fellowship	1998-2000
BC Advanced Systems Institute Provincial Scholarship	1998

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:

A Comparison Study of Nonmetastatic Breast Cancer Patients who were Prescribed Mega-doses of Vitamins and Minerals with a Matched Control group and a Control Population

Author: [redacted]

Natalie Judith Forde

November 6, 2000