Survival Analysis for Breast Cancer

by

Yongcai Liu
B.Eng., Tianjin University, China, 1983
M.Eng., Tianjin University, China, 1988
D.Sc., Technion - Israel Institute of Technion, Israel, 1997

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

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Survival Analysis for Breast Cancer

by

Yongcai Liu
D.Sc., Technion - Israel Institute of Technion, Israel, 1997

Supervisory Committee

Dr. M. Lesperance, (Department of Mathematics and Statistics)

Supervisor

Dr. J. Zhou, (Department of Mathematics and Statistics)

Departmental Member

Supervisory Committee

Dr. M. Lesperance, (Department of Mathematics and Statistics)

Supervisor

Dr. J. Zhou, (Department of Mathematics and Statistics)

Departmental Member

Abstract

This research carries out a survival analysis for patients with breast cancer. The influence of clinical and pathologic features, as well as molecular markers on survival time are investigated. Special attention focuses on whether the molecular markers can provide additional information in helping predict clinical outcome and guide therapies for breast cancer patients. Three outcomes, breast cancer specific survival (BCSS), local relapse survival (LRS) and distant relapse survival (DRS), are examined using two datasets, the large dataset with missing values in markers (n=1575) and the small (complete) dataset consisting of patient records without any missing values (n=910). Results show that some molecular markers, such as YB1, BCL2, should join ER, PR and HER2 to be integrated into cancer clinical practices. Further clinical research work is needed to identify the importance of CK56.

The 10 year survival probability at the mean of all the covariates (clinical variables and markers) for BCSS, LRS, and DRS is 77%, 91%, and 72% respectively.

Due to the presence of a large portion of missing values in the dataset, a sophisticated multiple imputation method is needed to estimate the missing values so that an unbiased and more reliable analysis can be achieved. In this study, three multiple imputation (MI) methods, data augmentation (DA), multivariate imputations by chained equations (MICE) and AREG, are employed and compared. Results shows that AREG is the preferred MI approach. The reliability of MI results are demonstrated

using various techniques. This work will hopefully shed light on the determination of appropriate MI methods for other similar research situations.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I am heartily thankful to my supervisor, Dr. Mary Lesperance, for her encouragement, guidance and support throughout my research and study at University of Victoria. I am grateful for her help, providing me with directed study opportunities so that I did not need to travel much between the UVic campus and BC Ministry of Health.

I began my formal study in statistics in 2005; while studying, I was also engaged in full-time employment. Although I have worked in a number of diverse fields such as chemical engineering, materials science, and physics, I find that I especially enjoy studying statistics. I have been eager and even excited to learn new statistical knowledge and applications in this field.

The generous support from Pacific Leaders Scholarship provided by BC Government is greatly acknowledged. I appreciate the understanding and support from Martha Burd, director of Modeling and Analysis Team of BC Ministry of Health, so that I can make a direct linkage between my study in statistics and work in health care which makes my study more important and attractive.

I wish to thank my family for their love, help, and support. The discussions and arguments with my wife Ruixia were particularly helpful and supportive. I dedicate this thesis to her.

Finally, I would like to offer my regards and thanks to the faculty and staff of the Department of Mathematics and Statistics and to all of those who supported me in any respect during the completion of my study at the UVic, especially Dr. J. Zhou and Dr. B. Reed in the math department, Dr. D. Giles in the economics department, and Dr. S. Dost in mechanical engineering.

Chapter 1

Introduction

1.1 General Background

Breast cancer is the most common malignant disease for females and the second most common type of cancer after lung cancer for both sexes. It primarily affects women older than 50 years. Even though the absolute incidence in women aged 20 - 40 years is low, breast cancer constitutes about 24 percent of new cancers in this age group. Hence treatment of breast cancer, including surgery, drugs (hormone therapy and chemotherapy) and radiation, is a main interest of the public health sector (Wiki Online 2009). In British Columbia (BC), Canada, breast cancer accounts for 29% of all cancer diagnoses for BC women. One in 9 women is expected to develop breast cancer during her lifetime and one in 33 is expected to die of breast cancer. In 2009, approximately 2980 BC women were diagnosed with breast cancer and 615 died of it

(BC Cancer Agency 2010).

Breast cancer is highly curable if diagnosed at an early stage. Traditional prognostic factors include auxiliary lymph node status, tumour size, nuclear grade and histological grade etc. They are important predictors of whether a tumor is localized and therefore amenable to local treatment. Many researches have been studying the relationship between these clinical variables and the survival time of breast cancer patients. Interest in novel prognostic markers is based on the fact that a significant number of patients with early stage breast cancer harbour microscopic metastasis at the time of diagnosis. It is now well established that adjuvant systemic therapy improves survival in patients with early-stage breast cancer (Early BC Group 1998). Treatment options for early stage breast cancer include chemotherapy (e.g. anthracyclines, taxanes) and hormone therapy (e.g. tamoxifen, aromatase inhibitors).

Systemic therapies are potentially toxic, however, and identifying individual patients who are at high risk and likely to benefit from the therapies remains a major challenge. For example, the risk of recurrence for a patient with negative auxiliary lymph nodes and a tumour size 1 to 2 cm is approximately 20 to 30%. Most patients in this group are currently offered adjuvant systemic therapy, although up to 70% of the patients would not need it because they are already cured of their disease. Unfortunately, the histological information is not sufficient to accurately assess individual risk and to possibly avoid adjuvant systemic therapy. A large number of molecular markers have been studied to predict prognosis or response to therapy, or both. Prognostic and

predictive molecular markers commonly used in clinical practice include BCL2, ER, PR, and Her2.

Nevertheless, the validation and appropriate quality control for most of the markers are still big issues to hinder their application. Only a small number of molecular markers have been successfully integrated into cancer clinical practices (McShane *et al.* 2005). Attention to sound statistical practice, in particular the use of statistical approaches that provide clinically relevant information, will help maximize the promise of molecular markers for the care of cancer patients.

It is widely accepted that many factors play an influential role in determining the survival time of breast cancer patients: age, stage diagnosed, screening history, systemic treatment, and genetic factors etc. Many analyses of cancer registry survival data use the Cox proportional hazards (PH) model (Cox 1972), which has had a profound influence on the development within the field of survival analysis. However, it relies on the assumption that the ratio between hazards is constant over time. Because of the long-term follow-up required for breast cancer patients, the PH assumption is often violated, leading to poor model fit.

One of the simplest ways to extend the Cox model is to include the interactions between the covariate effect and time, either as a linear function or as another polynomial function. Stratified models are also very useful, in which a covariate which does not satisfy proportionality can be factorized out (Eide *et al.* 1996). We use the stratification method for a simple treatment of the PH violation. Future work may

employ advanced models to account for complete time varying effects.

The dataset used in this study comes from The Genetic Pathology Evaluation Centre (GPEC) of British Columbia. It contains 2222 records (rows) and 193 fields (columns), which provide us with sufficient information to conduct various investigations. However, there are large portions of missing values in the dataset which need special treatment. Multivariate datasets that contain missing values in one or several of the variables arise frequently in statistical practice. Researchers have become increasingly aware of the problems and biases which can be caused by the missing data. As a result, many different methods of managing missing data have been proposed, of which multiple imputation (MI) has become an important and influential approach in the statistical analysis of incomplete data.

There are several methods to deal with missing data, including ad hoc methods such as case deletion, mean substitution and more principled single imputation methods such as maximum likelihood methods, multiple imputation, or others. The single imputation methods, however, ignore imputation uncertainty and are likely to produce coefficient confidence intervals that are too narrow, and p-values that are too small (Albmer *et al.* 2007). MI is also a statistically principled method. It accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data, which makes it better than single imputation methods. In MI, missing values for any variable are predicted using existing values from other variables. The predicted values are substituted for the missing values,

resulting in a full data set called an imputed data set. This process is performed multiple times, producing multiple imputed data sets. Standard statistical analysis is carried out on each imputed data set to obtain the multiple analysis results. These analysis results are then combined to produce one overall analysis.

This study is intended to get reliable missing value imputations for our breast cancer dataset. Several sophisticated and recently developed multiple imputation methods are employed, validated, and compared.

1.2 Objectives

This research investigates the influence of both standard clinical and pathologic features and molecular markers on the survival time of breast cancer patients. Particularly it seeks independent variable patterns to determine the survival times and identifies the correlations among the variables of interest. It examines whether the molecular markers can provide additional information in helping to predict breast cancer outcomes. The dataset is from a BC institution based on local patient records. The research results are expected to have direct impact on the health care of breast cancer patients for BC residents.

Several sophisticated multiple imputation methods are employed to fit the missing values in the dataset so that an unbiased and more reliable analysis can be achieved. Results are compared with each other, shedding light on the determination of appro-

priate MI methods under various situations.

1.3 Summary of Thesis

This thesis is organized as six main chapters. Chapter 1 gives a brief background and the objectives for this study. Chapter 2 provides a literature review on breast cancer and its prognostic factors including clinical factors and molecular markers. The details of the dataset used in the work are given in Chapter 3. Chapter 4 focuses on multiple imputations for missing values. A brief introduction is given first, and then the basic assumptions and general procedure of MI. The description of MI methods used in the work and imputation results are presented last. Chapter 5 gives details on survival analysis using both the complete dataset after imputation and the small dataset with all missing values removed. Chapter 6 summarizes and discusses the results presented in Chapter 4 and Chapter 5.

Finally, Appendices give R code for the work and additional tables and graphs for references.

Chapter 2

Breast Cancer and Its Prognostic Factors

2.1 Classification of Stages

Breast cancer is classified in four stages based on different risk factors, such as size of the tumor, whether the cancer is invasive or non-invasive, whether lymph nodes are involved, and whether the cancer has spread beyond the breast. The purpose of the staging system is to help organize the different factors and some of the features of the cancer into categories, in order to best understand a patient's prognosis (the most likely outcome of the disease), guide treatment decisions (together with other parts of the pathology report), and provide a common way to describe the extent of breast cancer for doctors and nurses all over the world, so that treatment results can

be compared and understood consistently worldwide (Breastcancer.org Online 2009). Stage I disease is the least advanced stage and the 5 year survival, i.e. the proportion of patients alive after 5 years, is about 90 percent. On the other end of the scale is the most advanced stage, stage IV, where 5 year survival is about 30 percent.

To classify the disease in different stages, a number of prognostic factors are investigated and the overall distribution of these prognostic factors decides which stage of disease is present. The most common system is the TNM staging system.

The TNM system describes the extent of the cancer based on three tumor morphological attributes: the size/extent of the primary tumor (T), regional lymph node involvement (N), and presence or absence of distant metastases (M) (Olivotto *et al.* 1996). The T (size) category describes the primary (original) tumor: T_X means the tumor can not be measured or found. T_0 means there is not any evidence of the primary tumor. T_{is} means the cancer is *in situ* (the tumor has not started growing into the breast tissue). The numbers $T_1 - T_4$ describe the size and/or how much the cancer has grown into the breast tissue. The higher the T number, the larger the tumor and/or the more it may have grown into the breast tissue. The N (node involvement) category describes whether or not the cancer has reached nearby lymph nodes, with the numbers N_0-N_3 describing the size, location, and/or the number of lymph nodes involved. The higher the N number, the more the lymph nodes are involved.

For example, a T_1, N_0, M_0 breast cancer would mean that the primary breast tumor is less than 2 centimeters across (T_1), does not have lymph node involvement (N_0), and

has not spread to distant parts of the body (M_0). This cancer would be grouped as stage I.

2.2 Clinical Factors

A number of factors with great prognostic value for breast cancer have been identified. The investigation of theses factors are important for prognosis and for making treatment decisions. For a cancer with a bad prognosis more aggressive treatment regimes may be chosen and the patient may be willing to accept more severe side effects. Some of the most common clinical factors are given as follows.

Age - In general, breast cancers that occur in women under age 40 tend to be more aggressive than those that occur most often in women over 50. But while age has some influence, it is not a major predictor of how serious any individual case of breast cancer will be.

Tumor size - Tumor size is evaluated by radiologic examination, and is categorized in four groups. Size does not tell the whole story. A small cancer can be very fast-growing. A larger cancer could be a gentle giant.

Lymph node involvement - Lymph nodes are filters along the lymph fluid channels. They try to catch and trap cancer cells before they reach other parts of the body. If lymph nodes have some cancer cells in them, they are called positive, which is associated with an increased risk of the cancer spreading.

Histology - Histology gives the information where the cancer starts from. *Ductal* denotes the cancer begins in the milk duct, and *Lobular* starts inside the milk-making glands (called lobules),

Grade - Grade is used to compare cancer cells to normal breast cells. Grades I to III represent well/fairly differentiated, moderately/partially differentiated, and poorly differentiated from the normal tissues respectively. A higher grade is associated with faster cancer growth, earlier spread of the cancer, and a greater incidence of axillary lymph node invasion.

Lymphatic/vascular invasion (Breastcancer.org Report 2009) - The breast has a network of blood vessels and lymph channels that connect breast tissue to other parts of the body. These are the highways that bring in nourishment and remove waste products. There is an increased risk of cancer coming back when cancer cells are found in the fluid channels of the breast. In these cases, doctors may recommend treatment to the patient's whole body, not just the breast area. Note that lymphatic or vascular invasion is different from lymph node involvement.

Estrogen and Progesterone receptors - A cancer is called ER-positive if it has receptors for the hormone estrogen. It is called PR-positive if it has receptors for the hormone progesterone. Breast cancers that are either ER-positive or PR-positive, or both, tend to respond to hormonal therapy. These cancers can be treated with medicine that reduces the estrogen in your body. They can also be treated with medicine that keeps estrogen away from the receptors. Hormone receptor negative tumors generally

have a more severe prognosis and more aggressive chemotherapy may be warranted. About 60 percent of primary breast cancers contain estrogen receptors and the levels are usually greater in post-menopausal women than in pre-menopausal.

Margin - Margins around a cancer tissue are described in three ways: *Negative* means no cancer cells can be seen at the outer edge. Usually, no more surgery is needed. *Positive* means cancer cells come right out to the edge of the tissue. More surgery may be needed. *Close* means cancer cells are close to the edge of the tissue, but not right at the edge. More surgery may be needed.

2.3 Molecular Markers

Molecular markers are molecules that show up in the blood, urine, or tumor of a cancer patient. These are hormones, proteins, or parts of proteins that are made by the tumor itself, by the surrounding normal tissue in response to the presence of tumor, or by the tissue of metastases (Voorzanger-Rousselot and Garnero 2007).

Although clinical indices such as tumor size and grade and axillary lymph node metastases are useful prognostic factors in breast cancer, there is an urgent need to identify molecular characteristics of breast carcinomas that more accurately predict clinical outcome and guide specific therapies for individual patients (Gradishar 2005). Tumor markers could identify a disease process, a specific tissue or patients characteristics and help establish the severity and extent of the disease. They are usually not

used alone for the diagnosis because most markers can be found in elevated levels in people who have benign conditions, and no tumor marker is yet specific to a particular cancer. Not every tumor will cause an elevation in the tumor marker test, especially in the early stages of cancer. Even though with these limitations, tumor markers may be useful for the four following clinical purposes (Voorzanger-Rousselot and Garnero 2007):

- Screening a healthy population for the presence of cancer or for detecting a group at a higher risk for developing a cancer.
- Making a diagnosis of cancer: a diagnostic tumor marker is a marker that will aid in detection of malignant disease in an individual. Preferably, the marker should be tissue specific and not influenced by benign diseases.
- Determining the prognosis in a patient with cancer. This would provide to the clinician a tool for early prediction of tumor recurrence, progression and development of metastases, following the initial surgical removal of the cancer but without administration of adjuvant therapy.
- Monitoring results of antitumoral treatment: tumor markers may predict how the patient is going to respond to a given therapy which includes surgery, radiation, chemotherapy or more recently targeted treatments.

A clinically useful molecular marker should have minimum requirements for sample preparation, high sensitivity and specificity. When molecular markers are used alone

in cancer diagnosis, in tumor recurrence and in treatment monitoring, they may have lower specificity and/or sensitivity than imaging techniques. Consequently a panel of different markers would be an adequate strategy to monitor cancer patients. This is one of the goals for the present research.

During the past few decades, with the explosion of molecular technology and understanding of the biology of breast cancer, numerous studies have been performed to identify prognostic and predictive factors in breast cancer, although with mixed success (Henry and Hayes 2006). A brief overview of the individual markers used in breast cancer is given in the following. Notice that all the markers discussed below except for Ki67 are in our dataset.

AB (alpha basic-crystalline) - Research showed that AB was commonly expressed in basal-like breast carcinomas, which account for 15 to 20% of breast cancer cases (Moyano *et al.* 2006), and it might contribute to short survival (Moyano *et al.* 2006; Perou *et al.* 2000). Not much research has been done for this marker so far.

BCL2 (Beta-cell lymphoma leukemia 2) - BCL2 is a mitochondrial protein known to inhibit apoptosis triggered by chemotherapy and radiation therapy. Lower levels of apoptosis could lead to malignant cell accumulation and therefore to a more aggressive clinical course for the disease. Although BCL2 can block apoptosis in vitro, several studies have shown that BCL2 overexpression is associated with improved disease-free survival rates (Gasparini *et al.* 1995). This may be in part because of the close association between BCL2 expression and ER expression (Esteva1 and Hortobagyi 2004).

Perhaps more important is the potential association between BCL2 expression and response to chemotherapy. Several studies have shown that patients with BCL2-negative breast cancer were more likely to respond to chemotherapy than patients with BCL2-positive tumors (Bonetti *et al.* 1998; Buchholz *et al.* 2003). However, other studies found no association between BCL2 expression and the response to chemotherapy (Bottini *et al.* 2000; Poelman *et al.* 2000). A recent work shows that BCL2 can be an independent predictor of breast cancer outcome and useful as a prognostic adjunct to the Nottingham Prognostic Index (NPI), particularly in the first 5 years after diagnosis (Callagy *et al.* 2006).

CA9 (carbonic anhydrases 9) – Carbonic anhydrases are a family of zinc metalloenzymes. CA9 is one of the best-known genes associated with tumour cell hypoxia, and is quickly and extensively upregulated under hypoxic conditions (Wykoff *et al.* 2000). CA9 expression was mainly found in high-grade, steroid receptor negative cancer tissues. Research has shown that CA9 levels were not significantly associated with relapse-free survival, and patients with low CA9 levels benefit more from adjuvant treatment than patients with high levels (Span *et al.* 2003; Watson *et al.* 2003).

CK56 (cytokeratin 5/6) - CK56 is a commonly used surrogate immunohistochemical indicator for tumors with the basal-like gene expression profile (Banerjee *et al.* 2006). Research has showed that CK56 can provide prognostic information in the group of breast cancer tumors with certain specific phenotypes to better predict breast cancer survival (Cheang *et al.* 2008; Rakha *et al.* 2007; Fadare *et al.* 2008; Nielsen *et*

al. 2004; Sasa *et al.* 2008).

ER (Estrogen receptor) - ER refers to a group of receptors that are activated by Estrogen. The main function of the estrogen receptor is as a DNA binding transcription factor that regulates gene expression. However, it has additional functions independent of DNA binding (Levin 2005). Estrogen receptors have been used primarily as predictive factors for hormone responsiveness in metastatic breast carcinoma. Indeed, tumors lacking ER respond infrequently to endocrine therapy, whereas response rates of 50 to 60% are observed in ER-positive tumors (Osborne *et al.* 1980). The absence of estrogen receptor is associated with early recurrence and correlates with poor prognosis (Knight *et al.* 1977; Crowe *et al.* 1991). ER is a currently widely used marker in prognostic practice.

GATA3 (GATA binding protein 3)– GATA3 is a key regulator of mammary gland formation and it directs differentiation of a newly identified progenitor population in the adult gland along the luminal-cell lineage (Asselin-Labat *et al.* 2007). It was suggested that patients whose tumors expressed low GATA3 had significantly shorter overall and disease-free survival when compared with those whose tumors had high GATA3 levels (Mehra *et al.* 2005). Another result suggested that GATA3 expression in breast cancer had a strong association with estrogen receptor but lacked independent prognostic value (Voduc *et al.* 2008).

Her1 (epidermal growth factor receptor 1) - Her1 is one of the four receptor members in the epidermal growth factor family of tyrosine kinases (another three are Her2

to Her4). All receptorr are structurally similar and consist of an extracellular region binding ligands, a transmembrane domain and an intracellular tyrosine kinase domain. Unlike thorough investigation for Her2, the possible role of Her1, Her3 and Her4 in breast cancer needs further elucidation. They are interesting because of their ability to heterodimerize with Her2 and theoretically, Her2 being without any known ligand, overexpression of Her2 could be caused by Her1 stimulation (Olsen *et al.* 2009). Research has showed that many basal-like tumors express Her1, which suggests candidate drugs for evaluation in these patients (Nielsen *et al.* 2004).

Her2 (human epidermal growth factor receptor 2) - Her2 is a protein giving higher aggressiveness in breast cancers. It is a gene that helps control how cells grow, divide, and repair themselves. About one out of four breast cancers has too many copies of the Her2 gene, in contrast to two copies of the Her2 gene in a healthy breast cell. The Her2 gene directs the production of special proteins, called Her2 receptors, in cancer cells. Cancers with too many copies of the Her2 gene or too many Her2 receptors tend to grow fast. They are also associated with an increased risk of spread (Breastcancer.org Report 2009).

One of the most promising uses of Her2 is probably to help identify how a patient will respond to different types of treatment like endocrine therapy or chemotherapy (Muss *et al.* 1994; Cornez and Piccart 2000). Her2 levels is useful to identify women who are likely to benefit from trastuzumab (Herceptin) treatment and monitor response to herceptin therapy (Ross *et al.* 2004; Kostler *et al.* 2004). Note that Her2, together

with ER and PR, has been included in a physician's standard pathology report.

IGFB (Insulin-like growth factor binding protein) – IGFB serves as a carrier protein for insulin-like growth factor 1, a polypeptide protein hormone similar in molecular structure to insulin, which plays an important role in childhood growth and continues to have anabolic effects in adults (Hwa *et al.* 1999). It was shown that serum concentrations of IGFB were not related to risk of breast cancer. However, insulin and insulin resistance may play a role in breast pathology in post-menopausal women (Schairer *et al.* 2007; Wolpin *et al.* 2009).

Ki67 - Ki67 is a nuclear antigen found in cells in the proliferative phases of the cell cycle. A strong correlation has been noted between the percentage of cells showing Ki67 staining and the nuclear grade, age, and mitotic rate (Sahin *et al.* 1991; Keshgegian and Cnaan 1995). Patients whose tumors overexpress Ki67 in more than 50% of the cells are at high risk of developing recurrent disease (Veronese *et al.* 1993).

P53 (protein 53) - P53 is a tumor suppressor gene. It is involved in regulating cell proliferation, inducing apoptosis, and promoting chromosomal stability. Disruption of these functions appears to have an important role in carcinogenesis (Wang *et al.* 1993). There is evidence that overexpression of p53 is relevant to breast cancer progression which leads to poor patient survival (Thor *et al.* 1992; Beenken *et al.* 2001; Pharaoh *et al.* 1999). Other research showed that p53 may have little clinical prognostic relevance (Rakha *et al.* 2007).

PR (progesterone receptor) - PR is an intracellular steroid receptor that specifi-

cally binds progesterone. Overexpression of PR serves as a functional assay because it indicates that the ER pathway is intact, even if the tumor is reported as ER-negative (Esteva1 and Hortobagyi 2004). Hence PR has been often in association with ER as a predictive factor for hormone therapy (see, e.g. Esteva1 and Hortobagyi 2004; Voorzanger-Rousselot and Garnero 2007; Ponzone *et al.* 2006), and appears in a standard pathology report. Research has also shown that PR status could define a subset of tumours with distinctive pathological characteristics and may help select those patients who derive the greatest benefit from endocrine adjuvant treatment, particularly within the first few years of follow-up (Ponzone *et al.* 2006).

YB1 (Y-box binding protein-1) – YB1 is a transcription and translation factor that can promote tumor growth and chemotherapy resistance by inducing growth-promoting genes such as Her2 and EGFR (epidermal grwoth factor receptor) (Wu *et al.* 2006). YB1 is found to be a highly predictive biomarker of relapse and poor survival across all breast cancer subtypes. Expression of YB1 universally identifies patients at high risk and in situations where more aggressive treatment may be needed (Habibi *et al.* 2008).

Chapter 3

Dataset Description

3.1 The Data Source

The data was collected and generated by the Genetic Pathology Evaluation Centre (GPEC) of the British Columbia. The research theme for GPEC is the validation of prognostic and predictive cancer biomarkers by immunohistochemical and fluorescence in situ hybridization studies on tissue microarrays of human tumor tissue samples.

The GPEC scientific team make use of extremely well-positioned resources in British Columbia to play a leading role in biomarker validation. There are large archives of cancer samples in the Vancouver General hospital and other BC hospitals, province-wide standardized protocol driven cancer care, and ready access to detailed patient records through the British Columbia Cancer Agency. Moreover, the development of tissue microarray technology enabled GPEC to bring these assets together and create

high throughput systems for the validation of cancer biomarkers.

3.2 The Breast Cancer Dataset

The raw data contains 2222 cases (rows) and 293 variables (columns). We are interested in the 9 clinical and 14 molecular marker variables, among all the available items. Three outcomes, breast cancer specific survival (BCSS), local relapse survival (LRS), and distant relapse survival (DRS), are investigated in the survival analysis. We focus our study on newly referred patients with invasive breast cancer. Some records for very sick patients or with very few cases are removed. Hence the exclusion criteria are: metastatic, tumor stage = 4, tumor size = (5-10) cm, histology = other, or TNM (Tumor, Node, Metastasis) stage = 3. There are 1876 records after the above clinical exclusions. Two markers, CyD1 and EMSY, are dropped due to too many missings (over 70%). Finally we have 9 clinical variables and 12 molecular markers used in the analysis. Furthermore, there are 301 records with unknown clinical information, and these records are excluded from analysis so that there are no missing values in all the clinical variables for the remaining 1575 cases.

The dataset with 21 variables and 1575 observations are summarized in Tables 3.1 and 3.2 [1]. Information for the missing data is given in Table 3.3. 665 observations (42%) have at least one missing value. There are 2082 missing values in total, occupying 6.3% of the total data points (21 times 1575). Table 3.4 presents the top 29 missing

patterns which account for more than half of the total missing counts.

Table 3.1: The dataset with missing values in markers only (n = 1575): molecular markers

Variable	Description	Negative (%)	Positive (%)	Missing (%)
AB	Alpha-basic crystalline	1162 (73.8)	133 (8.4)	280 (17.8)
BCL2	Beta-cell lymphoma leukemia 2	539 (34.2)	799 (50.7)	237 (15.1)
CK56	Cytokeratin 5/6	1235 (78.4)	98 (6.2)	242 (15.4)
Her1	Human epidermal growth factor receptor 1	1200 (76.2)	167 (10.6)	208 (13.2)
ERS	Estrogen receptor	440 (27.9)	1120 (71.1)	15 (1.0)
CA9	Carbon anhydrase 9	1211 (76.9)	223 (14.2)	141 (8.9)
P53	Tumor suppressor gene	1235 (78.4)	308 (19.6)	32 (2.0)
Her2	Human epidermal growth factor receptor 2	1317 (83.6)	195 (12.4)	63 (4.0)
IGFB	Insulin-like growth factor binding protein	836 (53.1)	505 (32.1)	234 (14.8)
PR	Progesterone receptor	664 (42.2)	742 (47.1)	169 (10.7)
YB1	Y-box-1 protein	808 (51.3)	491 (31.2)	276 (17.5)
GATA	A transcription factor	1012 (64.3)	378 (24.0)	185 (11.7)

Tables 3.5 and 3.6 present figures for the dataset without any missing values, that is, the set of 910 patients who have no missing marker values.

Several multiple imputation methods are employed to fit the missing values. Results are compared with each other to determine the best imputation method. Cox proportional hazards model is used for survival analysis using both the full dataset after filling the missing values (n = 1575) and the dataset resulting from the removal of all the missings (n = 910, see Tables 3.5 and 3.6). Detailed analyses and discussions

[1]AGECAT and AGE, GRADECAT and GRADE, PPNODECAT and PPNODE, as well as SIZECAT and SIZE are used interchangeably in this work.

Table 3.2: The dataset with missing values in markers only (n = 1575): clinical variables

Variable names	Description	Category	Number (%)
AGECAT, AGE	Age	1: ≤ 50 2: > 50	477 (30.2) 1098 (69.8)
Histology		1: Ductal 2: Lobular	1461 (92.8) 114 (7.2)
GRADECAT, GRADE	Tumor grade	1: grade 1 2: grade 2 3: grade 3	93 (5.9) 690 (43.8) 792 (50.3)
ERPOSNE	Estrogen receptor status at diagnosis	1: Negative 2: Positive	313 (19.9) 1262 (80.1)
LVNNE	Lymphatic/vascular invasion	1: Negative 2: Positive	913 (58.0) 662 (42.0)
#PosNodes	Number of positive lymph nodes	1: 0 2: 1-3 3: ≥ 4	918 (58.3) 427 (27.1) 230 (14.6)
PPNODECAT, PPNODE	Proportion of positive lymph nodes	1: 0 2: ≤ 0.25 3: > 0.25	918 (58.3) 345 (21.9) 312 (19.8)
SIZECAT, SIZE	Tumor size in cm	1: ≤ 2 2: 2-5	869 (55.2) 706 (44.8)
SYS	initial systemic therapy	1: none 2: hormones only 3: chemotherapy only 4: both	700 (44.4) 500 (25.4) 278 (17.7) 97 (6.2)

are given in Chapters 4 and 5. We use the dataset with n = 1575 if there is no special notation.

Table 3.3: Information about the missing values for markers in the dataset

Name	Missing	Missing %
ERS	15	0.7
P53	32	1.5
Her2	63	3.0
CA9	141	6.8
PR	169	8.1
GATA	185	8.9
Her1	208	10.0
IGFB	234	11.2
BCL2	237	11.4
CK56	242	11.6
YB1	276	13.3
AB	280	13.4
Total	2082	100

Table 3.4: The top 29 missing patterns

	Name	Number of observations	Proportion
01	PR	39	0.101
02	YB1	37	0.095
03	GATA, IGFB, YB1, AB	31	0.080
04	GATA	29	0.075
05	CK56	23	0.059
06	BCL2	21	0.054
07	Her1	21	0.054
08	IGFB	18	0.046
09	AB	17	0.044
10	GATA, IGFB, YB1	17	0.044
11	Her2, PR	13	0.033
12	Her1, BCL2, CK56, AB	10	0.025
13	IGFB, YB1	9	0.023
14	CA9	8	0.020
15	Her2	8	0.020
16	BCL2, CK56	8	0.020
17	GATA, YB1	8	0.020
18	Her1, BCL2, CK56	7	0.018
19	YB1, AB	6	0.015
20	Her1, CK56	6	0.015
21	Her1, BCL2	6	0.015
22	PR, AB	6	0.015
23	BCL2, CK56, AB	6	0.015
24	IGFB, YB1, AB	6	0.015
25	CA9, Her1, BCL2, CK56, AB	6	0.015
26	IGFB, AB	5	0.012
27	GATA, CK56	5	0.012
28	CA9,Her1,BCL2,CK56,YB1,AB	5	0.012
29	CA9,Her1,IGFB,BCL2,CK56,YB1,AB	5	0.012
Total		386	1.000

Table 3.5: The dataset without missing values (n = 910): molecular markers

Variable	Negative (%)	Positive (%)
AB	813 (89.3)	97 (10.7)
BCL2	378 (41.5)	532 (58.5)
CK56	833 (91.5)	77 (8.5)
Her1	789 (86.7)	121 (13.3)
ERS	226 (24.8)	684 (75.2)
CA9	751 (82.5)	159 (17.5)
P53	695 (76.4)	215 (23.6)
Her2	774 (85.1)	136 (14.9)
IGFB	570 (62.6)	340 (37.4)
PR	404 (44.4)	506 (55.6)
YB1	524 (57.6)	386 (42.4)
Gata3	659 (72.4)	251 (27.6)

Table 3.6: The dataset without missing (n = 910): clinical variables

Variable	Category	Number (%)
AGECAT, AGE	≤ 50	283 (31.1)
	> 50	627 (68.9)
Histology	Ductal	869 (95.5)
	Lobular	41 (4.5)
GRADECAT, GRADE	1	45 (4.9)
	2	365 (40.1)
	3	500 (54.9)
ERPOSNE	Negative	192 (21.1)
	Positive	718 (78.9)
LVNNE	Negative	500 (54.9)
	Positive	410 (45.1)
#PosNodes	0	504 (55.4)
	1-3	267 (29.3)
	≥ 4	139 (15.3)
PPNODECAT, PPNODE	0	504 (55.4)
	≤ 0.25	216 (23.7)
	> 0.25	190 (20.9)
SIZECAT, SIZE	≤ 2	478 (52.5)
	2-5	432 (47.5)
SYS	1	388 (42.6)
	2	285 (31.3)
	3	181 (19.9)
	4	56 (6.2)

Chapter 4

Multiple Imputation Methods for Missing Values

4.1 Introduction

The occurrence of missing data is a pervasive problem in data analysis. Data values may be absent from a dataset for numerous reasons, for example, the inability to measure certain attributes and incomplete gathering from data sources. Researchers have become increasingly aware of the problems and biases which can be caused by the missing data. Significant advances have been made in the past two decades regarding methodologies which handle responses to these problems and biases. Among them multiple imputation (MI) has become an important and influential approach in the statistical analysis of incomplete data.

It is important to understand that once data are missing, it is impossible not to treat them: once data are missing, any subsequent procedure applied to that data set represents a response in some form to the missing data problem.

Some of the most popular methods to manage missing data involve ad-hoc deletion or replacement of the missing data. These methods typically edit missing data to produce a complete data set and are attractive because they are easy to implement. However, researchers have been cautioned against using these methods because they have been shown to have serious drawbacks (e.g., Little and Schenker 1995; Graham and Hofer 2000; Graham *et al.* 1997; Schafer and Graham 2002). For example, handling missing data by eliminating cases with missing data (so called listwise deletion) will bias results if the remaining cases are not representative of the entire sample. This method is the default for most statistical software. Another common method available in most statistical packages is mean substitution, which replaces missing data with the average of valid data for the variable in question. Because the same value is being substituted for each missing case, this method artificially reduces the variance of the variable in question, in addition to diminishing relationships with other variables. Graham *et al.* (2003) referred to these traditional methods as unacceptable methods, especially when there is a large portion of missing data in the dataset. Examples of other unacceptable methods include pairwise deletion and regression-based single imputation.

Additionally, there exist more statistically principled methods of handling missing data which have been shown to perform better than ad-hoc methods (e.g., Little and

Rubin 1987; Graham *et al.* 1997; Schafer and Graham 2002). These methods do not concentrate solely on identifying a replacement for a missing value, but on using available information to preserve relationships in the entire data set. Maximum likelihood estimation is one such method. This method requires specification of a statistical model for each analysis and is a sound method for treating missing data, but is often difficult to implement for less-advanced analysts (Schafer 2002). The Expectation Maximization (EM) algorithm is another method which has been applied to missing data (Little and Schenker 1995; Schafer and Olsen 1998), but obtaining standard errors using EM involves auxiliary methods such as bootstrapping. Multiple imputation is a statistically principled method which has been an attractive choice as a solution to missing data problems because it represents a good balance between the quality of results and ease of use.

In multiple imputations, missing values for any variable are predicted using existing values from other variables. The predicted values, called imputes, are substituted for the missing values, resulting in a full data set called an imputed data set. This process is performed multiple times, producing multiple imputed data sets. Standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. These analysis results are then combined to produce one overall analysis. Multiple imputation accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data. Maintaining the original variability of the missing data is done by creating

imputed values which are based on variables correlated with the missing data and causes of missingness. Uncertainty is accounted for by creating different versions of the missing data and observing the variability between imputed data sets. It is important to note that imputed values produced from an imputation model are not intended to be guesses as to what a particular missing value might be; instead, this modeling is intended to create an imputed data set which maintains the overall variability in the population while preserving relationships with other variables. Thus, in performing multiple imputation, a researcher is interested in preserving important characteristics of the data set as a whole (e.g., means, variances, regression parameters).

Like any statistical technique, multiple imputation depends on some assumptions, and responsible use of multiple imputation involves a basic understanding of these assumptions and their implications.

This work is intended to get good missing value imputations for our breast cancer dataset. We compare three different MI methods and two different treatments for the dataset.

4.2 Basic Assumptions and General Procedures for Multiple Imputation

4.2.1 Assumption of ignorability

Let M be a set of random indicator variables that partitions the complete data Y_{com} into observed, Y_{obs} and missing, Y_{mis}. In general, M can be regarded as an array of the same size as Y_{com} containing 0 in every position where the corresponding element of Y_{com} is observed and 1 in every position where the element is missing. We will refer to M as the missing indicator(s) or the missingness. Based on the work of Rubin (1987), missing data can be often categorized as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Data are said to be MAR if the conditional distribution of M given the observed data is independent of the unobserved data:

$$P\left(M \mid Y_{obs}, Y_{mis}, \xi\right) = P\left(M \mid Y_{obs}, \xi\right), \tag{4.1}$$

where ξ is an unknown parameter related to the missingness mechanism. Data are said to be MCAR if the probability of being missing is independent of Y_{com}:

$$P\left(M \mid Y_{obs}, Y_{mis}, \xi\right) = P\left(M \mid \xi\right). \tag{4.2}$$

When neither MCAR nor MAR hold the missing data mechanism is said to be MNAR. For example, suppose we need a group of patients for a research program. If the patients were chosen randomly and later some patients drop the program randomly,

then we have MCAR data. If the patients were chosen carefully but some drop randomly, it is MAR. If patients drop the program with some patterns, we get MNAR. In reality we often meet the situations that patients were chosen carefully but drop with unknown reasons. Then we end up with MNAR or MAR. People argue that MAR is plausible under most of these situations.

Much of the MI work has been based on the MAR assumption. Under MAR, the probability distribution of the observed data can be factored into two pieces (Rubin 1987):

$$\begin{aligned} P\left(M, Y_{obs} \mid \theta, \xi\right) &= \int P\left(M, Y_{com} \mid \theta, \xi\right) dY_{mis} \\ &= \int P\left(M \mid Y_{com}, \xi\right) P(Y_{com} \mid \theta) dY_{mis} \\ &= P\left(M \mid Y_{obs}, \xi\right) P\left(Y_{obs} \mid \theta\right). \end{aligned} \tag{4.3}$$

One pertaining to the parameter of interest θ and the other pertaining to the nuisance parameter ξ.

If it is further assumed that the two unknown parameters, θ and ξ are distinct, which means, from a Bayesian perspective, that any joint prior distribution applied to (θ, ξ) must factor into independent marginal priors for θ and ξ, then likelihood-based inferences about θ will be unaffected by ξ or $P\left(M \mid Y_{obs}, \xi\right)$. Maximum-likelihood estimation of θ, likelihood-ratio tests concerning θ and so on, can then be performed without regard for the missing-data mechanism; that is, the missing-data mechanism may be safely ignored. Hence, ignorability requires two conditions, MAR for the miss-

ing data and distinctness for the unknown parameters. The ignorability assumption occupies a very special position in the missing value framework, not only because it is especially plausible in practice, but also because it represents the most general condition under which valid inference can be obtained without reference to the missing data mechanism.

4.2.2 General procedures for MI

Multiple imputation inference involves three distinct phases: 1. The missing data are filled in m times to generate m complete data sets. 2. The m complete data sets are analyzed using standard procedures. 3. The results from the m complete data sets are combined for inference.

Imputation

In MI, we first impute m independent versions of the missing data from the posterior predictive distribution $P(Y_{mis} \mid Y_{obs}, M)$ under a joint model for the complete data $Y_{com} = (Y_{obs}, Y_{mis})$ and M. Under the ignorability assumption, M can be dropped out and we can impute the missing values from $P(Y_{mis} \mid Y_{obs})$ (Schafer 2002).

In practice, MI's are usually created by Bayesian rather than frequentist arguments. That is, they are typically drawn from a posterior predictive distribution for the missing data given the observed data. Let $P(Y_{com} \mid \theta)$ denote a model for the complete data

with unknown parameter θ. The posterior predictive distribution for Y_{mis} is

$$P\left(Y_{mis} \mid Y_{obs}\right) = \int P\left(Y_{mis} \mid Y_{obs}, \theta\right) P\left(\theta \mid Y_{obs}\right) d\theta, \tag{4.4}$$

where

$$P\left(\theta \mid Y_{obs}\right) \propto P(\theta) \int P\left(Y_{mis} \mid Y_{obs}, \theta\right) dY_{mis} \tag{4.5}$$

is the observed-data posterior distribution for θ and $P(\theta)$ is the prior distribution. The right-hand side of Equation (4.4) suggests that MI may be drawn by repeating this two-step process for $i = 1, ..., m$: first, draw $\theta^{(i)}$ from $P(\theta \mid Y_{obs})$, given by Equation (4.5); then draw $Y_{mis}^{(i)}$ from $P(Y_{mis} \mid Y_{obs}, \theta^{(i)})$ (Rubin 1987).

Analysis

Imputing the data results in m complete data sets. If Q is the measure of interest, the estimates ($\hat{Q}^1$, $\hat{Q}^2$, . . . , $\hat{Q}^m$) and their squared standard errors can be computed using common complete-data methods. This analysis will be equivalent to the analysis that would have been done if we had complete data. Only in this case, it will be done m times. Estimates we might consider are regression coefficients, odds ratios, etc.

Combining

Rubin develops the rules for combining the estimates and their standard errors. With m imputations, we can compute m different sets of estimates and their variances for a parameter Q. Let $\hat{Q}^i$ and $\hat{U}^i$ be the point and variance estimates from the ith imputed

data set, $i = 1, 2, ..., m$. Then the point estimate for Q from multiple imputations is the average of the m complete-data estimates:

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}^i. \tag{4.6}$$

Let $\bar{U}$ be the within-imputation variance, which is the average of the m complete-data estimates

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U}^i, \tag{4.7}$$

and B be the between-imputation variance

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}^i - \bar{Q})^2, \tag{4.8}$$

Then the variance estimate associated with $\bar{Q}$ is the total variance:

$$T = \bar{U} + (1 + \frac{1}{m})B. \tag{4.9}$$

The term $(1 + \frac{1}{m})$ adjusts for the fact that we are effectively conditioning on the finite m number of imputations. The statistic $(Q - \bar{Q})/\sqrt{T}$ is approximately distributed as a t-distribution with v_m degrees of freedom (Rubin 1987), where

$$v_m = (m-1) \left[1 + \frac{\bar{U}}{(1 + m^{-1})B} \right]^2, \tag{4.10}$$

with the last term reflecting the relative increase in variances due to multiple imputations.

When the complete-data degrees of freedom v_0 is small and there is only a modest proportion of missing data, the computed degrees of freedom, v_m, can be much larger

than v_0, which is inappropriate. Some adjustment for the degrees of freedom would be needed (Barnard and Rubin 1999) in this case.

Similar to the univariate inferences, multivariate inferences based on Wald's tests can also be derived from the m imputed data sets. See (Rubin 1987) for details.

4.3 MI Methods

There are a number of software packages offering a variety of multiple imputation techniques. Three frequently used and flexible MI methods are employed in this work and described below.

4.3.1 Data augmentation

Drawing from the posterior distribution in the imputation stage of MI is the most complicated task. There are several algorithms that accomplish this function, and one of them, data augmentation, is given in more detail below.

Data augmentation (DA) (Rubin 1987; Tanner and Wong 1987) is closely related to Gibbs sampling, the most popular and well-known form of Markov Chain Monte Carlo (MCMC) methodology. MCMC creates draws from probability distributions (f). Since these distributions are either hard to find or do not have a closed form, MCMC is used to generate a sequence $X^{(1)}, X^{(2)}, ..., X^{(t)}, ...$ such that each X depends in some way on the previous one, and the stationary distribution ($X^{(t)}$ as $t \to \infty$) is a draw from the

target distribution function f.

Gibbs sampling updates a set of variables one by one given others. Suppose we have variables of interest $\mathbf{q} = (\mathbf{q_1}, \mathbf{q_2}, ..., \mathbf{q_n})$. To get $\mathbf{q^{t+1}} = (\mathbf{q_1^{t+1}}, \mathbf{q_2^{t+1}}, ..., \mathbf{q_n^{t+1}})$ using a Gibbs sampling procedure, the first variable is updated given others in t step: $\mathbf{q^{t+1}} = (\mathbf{q_1^{t+1}}, \mathbf{q_2^{t}}, ..., \mathbf{q_n^{t}})$. Then the second one is updated using q_1 in $t+1$ step and others in t step: $\mathbf{q^{t+1}} = (\mathbf{q_1^{t+1}}, \mathbf{q_2^{t+1}}, \mathbf{q_3^{t}}, ..., \mathbf{q_n^{t}})$. This procedure is repeated until all variables are updated.

The name DA arose from applications of this algorithm to Bayesian inference with missing data. In most of the incomplete-data scenarios, the observed-data posterior distribution $P(\theta \mid Y_{obs})$ is intractable. When Y_{obs} is *augmented* by an assumed value of the Y_{mis}, however, the resulting complete-data posterior distribution $P(\theta \mid Y_{obs}, Y_{mis})$ becomes much easier to handle.

The iterative procedure is as follows: for a current guess of the parameter $\theta^{(t)}$, draw values to replace the missing values from the conditional predictive distribution of Y_{mis},

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right). \tag{4.11}$$

Then given $Y_{mis}^{(t+1)}$, we would draw a new value of θ from the complete data posterior distribution,

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right). \tag{4.12}$$

Repeating the iterative procedure until a stationary state is attained will produce draws from $P(\theta, Y_{mis} \mid Y_{obs})$, which implies that at the end of the procedure one can have

estimates of both $P(Y_{mis} \mid Y_{obs})$ and $P(\theta \mid Y_{obs})$.

4.3.2 MICE method

MICE stands for multivariate imputations by chained equations. The MICE procedure (Royston 2004; Royston 2005) assumes that some joint probability distribution exists for a dataset containing mixed sets of variables (continuous, binary, ordinal, or categorical), but it employs only conditional distributions based on the measurement level of the variable being imputed. The underlying joint distribution for all variables is not specified, but it is assumed that such a distribution exists and that the iterative imputation procedure based on conditional distributions will converge to this distribution. Although this is a theoretical weakness of the MICE iterative procedure, simulation studies have shown that it performs well in practice (Raghunathan *et al.* 2001; Van Buuren *et al.* 2006).

The MICE iterative procedure is implemented as follows. Initially, the set of variables with missing data are ordered in terms of the amount of missing data in each, from least to most. Let Y_1 through Y_k be the variables with missing data. For each Y_i $(1 \leq i \leq k)$ random draws from the observed scores on Y_i are used to impute the missing values. Next, the original observed values of Y_1 are regressed on $U = (Y_2, ..., Y_k, \mathbf{X})$ ($\mathbf{X}$ are a set of variables without missing values) using an appropriate (linear, logistic, etc.) regression model, and the initial imputed values are replaced with new imputed values as described below. The observed Y_2 values are then regressed on $U = (Y_1, Y_3, ..., Y_k, \mathbf{X})$,

new imputed values are assigned to replace the previous imputed values, and then Y_3 is regressed on $U = (Y_1, Y_2, Y_4..., Y_k, \mathbf{X})$, and so on through Y_k. After this first round, where now all variables with missing values have been newly imputed, the procedure is started again with observed values of Y_1 regressed on $U = (Y_2, ..., Y_k, \mathbf{X})$, and so on. The entire round of imputations is then repeated for a pre-specified c number of times (default = 10) and the final complete multiply imputed data set is saved. The procedure is started over again for another c rounds and the final complete data set is saved. This is done m times. If the variable being imputed is continuous, binary, ordinal, or categorical (with $k > 2$), then least squares regression, binary or ordinal logistic regression, or multinomial regression is used to obtain estimates of the regression coefficients and the error variance, using only the observed values for the dependent variable. To obtain the imputed values for the missing cases, a small random disturbance is added to the estimates of both the residual error variance and the regression coefficients, and these revised estimates are then used to generate the imputed values. For example, if the variable being imputed, Y_i, is continuous, then a linear regression model is fitted that relates Y_i to the most recently updated U, $Y_i = U\beta + e$, using only cases with observed scores on Y_i. The imputed values are then obtained as follows (Raghunathan *et al.* 2001):

1. Let $B = (U'U)^{-1}U'Y_i$ be the current estimate of β, let $SSE = (Y_i - UB)'(Y_i - UB)$ be the residual sum of squares, and let the residual degrees of freedom be $df =$ number of rows in Y_i (with observed values) - number of columns in U. Select

a chi-square random deviate u with df degrees of freedom and compute ${\sigma_*}^2 = SSE/u$.

2. Factor the asymptotic variance-covariance matrix of B as $(U'U)^{-1} = TT'$ (the Cholesky decomposition). Generate a vector z of independent standard normal variates with the same number of rows as there are coefficients in B. Compute $B_* = B + Tz$. This latter step constitutes taking a draw from the posterior distribution of each of the parameters in β, assuming the usual multivariate normal distribution for a set of regression coefficients.

3. Let U_{miss} denote the U matrix for those cases with missing Y_i values. The imputed values are generated from $Y_{j*} = U_{miss}B_* + \sigma_*\mu$, where μ is a vector of random normal deviates of the same row dimension as U_{miss}.

Note that the imputed values are not simply the predicted scores using the adjusted regression coefficients but also include an adjusted error term to reflect the uncertainty of scores around the regression line. For logistic, Poisson, and multinomial regression, the adjustments to the estimated regression coefficients is essentially the same as in Step 2 above, although the adjustments to the error variances in Step 3 vary for each model. The entire process is cycled through until m complete imputed data sets have been generated. The m imputed complete data sets are then analyzed separately using any standard statistical model, and the results combined using Rubin's rules described above.

4.3.3 AREG method

This MI method uses additive regression, bootstrapping, and Predictive Mean Matching (PMM) in the missing value imputations, and is implemented in the *aregImpute* function of the *Hmisc* package in R (R Documentation 2009). It takes all aspects of uncertainty in the imputations into account by using the bootstrap to approximate the process of drawing predicted values from a full Bayesian predictive distribution. Different bootstrap resamples are used for each of the multiple imputations, i.e., for the ith imputation of a sometimes missing variable, $i = 1, 2, ...m$, a flexible additive model is fitted on a sample with replacement from the original data and this model is used to predict all of the original missing and non-missing values for the target variable.

The sequence of steps used by the aregImpute algorithm is the following.

1. For each variable containing x missing values where $x > 0$, initialize the missing values to values from a random sample (without replacement if a sufficient number of non-missing values exist) of size x from the non-missing values.

2. For (burn-in + m) iterations do the following steps. The first burn-in iterations provide a burn-in, and imputations are saved only from the last m iterations. For each variable containing any missing values, fit a flexible additive model to predict this target variable while finding the optimal transformation of it (unless the identity transformation is forced). Use this fitted flexible model to predict the target variable in all of the original observations. Impute each missing value

of the target variable with the observed value whose predicted transformed value is closest to the predicted transformed value of the missing value, or use a draw from a multinomial distribution with probabilities derived from distance weights, if match = *weighted* (the default).

3. For each variable with missing values, after step 2, use these imputations the next time the current target variable is used as a predictor of other missing variables. When the missingness mechanism for a variable is so systematic that the distribution of observed values is truncated, predictive mean matching does not work. It will only yield imputed values that are near observed values, so intervals in which no values are observed will not be represented by imputed values. For this case, the only hope is to make regression assumptions and use extrapolation. With type = *regression*, aregImpute will use linear extrapolation to obtain a (hopefully) reasonable distribution of imputed values. The *regression* option causes aregImpute to impute missing values by adding a random sample of residuals (with replacement if there are more missing values than measured values) on the transformed scale of the target variable. After random residuals are added, predicted random draws are obtained on the original untransformed scale using reverse linear interpolation on the table of original and transformed target values (linear extrapolation when a random residual is large enough to put the random draw prediction outside the range of observed values).

We note that the AREG method is similar to MICE, as it also takes all aspects of uncertainty into account but using the bootstrap to approximate the drawing of predicted values from a full Bayesian posterior distribution and using different bootstrap samples for each multiple imputation. It also differs from MICE, as it applies predictive mean matching for all variable types (i.e., continuous, binary, and categorical). It does not need iterative maximum probability fitting for binary and categorical variables, and neither does it require computing residuals or for curtailing imputed values to be in the range of actual data (Donders *et al.* 2006; Moons *et al.* 2006).

4.4 Results and Discussion

4.4.1 Sensitivity, selectivity, and agreement analysis

Multiple imputations are applied using the three MI methods described above. These methods are evaluated by calculating several parameters, sensitivity, selectivity, positive predictive value (PPV), negative predictive value (NPV), and total agreement, based on the comparison between the *true* values and imputed ones.

The *true* and imputed data are obtained as follows. In order to avoid possible correlations between markers which lead to unstable models, we use only one marker and all the clinical variables in the model. AB, Her2, or PR was chosen as the representive marker since their behavior on survival may be different based on literature review: Her2 and Pr may have significant effects on survival while AB may not. We

use the dataset with 1575 records so that all the missing values appear in the field of the molecular markers and there are no missing values for clinical variables. The values for all the markers are binary only, either 0 (negative) or 1 (positive).

Firstly, we use AREG to fill in the missing values and treat this generated full dataset as the *true* values. Survival analysis is carried out using Cox proportional hazards model based on the *true* values with breast cancer specific survival years as the dependent variable. Then we randomly assign missing values for the marker according to the given missing proportions 100 times to generate 100 datasets with missing values. Finally we use all the three MI methods to do the imputations for each of the 100 datasets so that the imputed values are obtained.

We obtain the following matrix after the missing value imputation:

	True Positive	True Negative
Imputed Positive	a	b
Imputed Negative	c	d

Sensitivity measures the probability of imputed positive values among all the true positive values, equivalent to $a/(a+c)$; Selectivity measures the probability of imputed negative values among all the true negative values, equivalent to $d/(b+d)$. PPV is the proportion of the true positive values among the total imputed positive values: $a/(a+b)$; NPV is the proportion of the true negative values among the total imputed negative values: $d/(c+d)$. Agreement is given by $(a+d)/(a+b+c+d)$.

Table 4.1 to Table 4.3 give the mean and standard error (SE) values for the 100 replications (where each consists of the average of sensitivity, selectivity, PPV, NPV and agreement over $m = 5$ multiple imputations) using marker = AB, Her2 and PR respectively. Results using all other markers except for AB, Her2, and PR with missing proportion = 5% are given in Appendix A.

Table 4.1: Sensitivity, selectivity, and agreement analysis: marker = AB.

Method	missing %	parameter	sensitivity	selectivity	PPV	NPV	agreement
AREG	2	Mean	0.6152	0.9129	0.5536	0.9278	0.8638
		SE	0.0233	0.0052	0.0205	0.0054	0.0064
	5	Mean	0.5509	0.9228	0.5631	0.9193	0.8651
		SE	0.0127	0.0034	0.0132	0.0034	0.0040
	15	Mean	0.5108	0.9209	0.5056	0.922	0.8643
		SE	0.0076	0.002	0.0085	0.002	0.0024
	35	Mean	0.5257	0.9179	0.5151	0.9202	0.8613
		SE	0.0041	0.0011	0.0064	0.0016	0.0012
MICE	2	Mean	0.5857	0.9147	0.5574	0.9194	0.8581
		SE	0.0233	0.0053	0.0207	0.0058	0.0062
	5	Mean	0.5534	0.9221	0.5633	0.9181	0.8638
		SE	0.0135	0.0035	0.0134	0.0036	0.0042
	15	Mean	0.4938	0.9227	0.5099	0.9175	0.8619
		SE	0.0074	0.0018	0.0083	0.0021	0.0022
	35	Mean	0.5074	0.9171	0.5230	0.9098	0.8529
		SE	0.0055	0.0011	0.0057	0.0026	0.0021
DA	2	Mean	0.3278	0.9275	NA	0.8075	0.7769
		SE	0.0082	0.0043	NA	0.0045	0.0046
	5	Mean	0.3318	0.9203	0.5621	0.8172	0.7809
		SE	0.0053	0.0024	0.0096	0.0028	0.0028
	15	Mean	0.2970	0.9198	0.5428	0.8037	0.7688
		SE	0.0031	0.0016	0.0064	0.0019	0.0017
	35	Mean	0.3507	0.9176	0.6207	0.7859	0.7600
		SE	0.0019	0.0011	0.0046	0.0010	0.0009

The main points from the sensitivity analysis shown in Tables 4.1 to 4.3 are summarized below:

Table 4.2: Sensitivity and agreement analysis: marker = Her2.

Method	missing %	parameter	sensitivity	selectivity	PPV	NPV	agreement
AREG	2	Mean	0.4713	0.8763	0.4631	0.8814	0.7994
		SE	0.0178	0.0067	0.0189	0.0062	0.0073
	5	Mean	0.6248	0.8600	0.6221	0.8601	0.7941
		SE	0.0098	0.0041	0.0094	0.0046	0.0041
	15	Mean	0.5519	0.8695	0.5431	0.8730	0.7992
		SE	0.0052	0.0022	0.0062	0.0024	0.0023
	35	Mean	0.5356	0.8689	0.5331	0.8692	0.7954
		SE	0.0032	0.0011	0.0042	0.0019	0.0012
MICE	2	Mean	0.4204	0.8732	0.4368	0.8648	0.7844
		SE	0.0184	0.0066	0.0191	0.0066	0.0074
	5	Mean	0.6071	0.8640	0.6359	0.8484	0.7899
		SE	0.0102	0.0045	0.0095	0.0050	0.0046
	15	Mean	0.5331	0.8720	0.5551	0.8611	0.7932
		SE	0.0055	0.0022	0.0059	0.0027	0.0026
	35	Mean	0.5308	0.8692	0.5456	0.8611	0.7911
		SE	0.0041	0.0014	0.0043	0.0025	0.0020
DA	2	Mean	0.2837	0.8844	NA	0.7779	0.7263
		SE	0.0070	0.0058	NA	0.0044	0.0050
	5	Mean	0.4543	0.8633	0.6362	0.7517	0.7216
		SE	0.0046	0.0041	0.0089	0.0028	0.0027
	15	Mean	0.3889	0.8656	0.5616	0.7614	0.7187
		SE	0.0026	0.0020	0.0054	0.0020	0.0016
	35	Mean	0.3991	0.8669	0.5927	0.7475	0.7125
		SE	0.0018	0.0012	0.0043	0.0015	0.0009

- Standard Error: SE's are generally two orders smaller than the mean, indicating the consistency of the results over simulated datasets. Although there could be some differences for the SE's using different MI methods, they are too small to influence the values.

- Influence of missing proportion: Under various missing proportions, little change for selectivity, NPV, and agreement can be seen. There are some changes for sensitivity and PPV with different missing proportions, but consistency across

Table 4.3: Sensitivity and agreement analysis: marker = PR.

Method	missing %	parameter	sensitivity	selectivity	PPV	NPV	agreement
AREG	2	Mean	0.6269	0.6245	0.6027	0.6474	0.6194
		SE	0.0114	0.0115	0.0121	0.0109	0.0084
	5	Mean	0.6193	0.6518	0.6250	0.6456	0.6339
		SE	0.0079	0.0073	0.0082	0.0074	0.0058
	15	Mean	0.6233	0.6236	0.6263	0.6201	0.6225
		SE	0.0041	0.0042	0.0048	0.0047	0.0029
	35	Mean	0.6357	0.6119	0.6293	0.6183	0.6238
		SE	0.0028	0.0025	0.0031	0.0029	0.0022
MICE	2	Mean	0.6237	0.6297	0.6137	0.6418	0.6228
		SE	0.0118	0.0130	0.0132	0.0113	0.0097
	5	Mean	0.6090	0.6434	0.6134	0.6394	0.6252
		SE	0.0076	0.0076	0.0075	0.0075	0.0059
	15	Mean	0.6314	0.6318	0.6347	0.6280	0.6307
		SE	0.0039	0.0044	0.0048	0.0042	0.0031
	35	Mean	0.6372	0.6114	0.6316	0.6167	0.6242
		SE	0.0025	0.0030	0.0033	0.0034	0.0021
DA	2	Mean	0.5799	0.5909	0.5636	0.6062	0.5809
		SE	0.0079	0.0065	0.0071	0.0084	0.0045
	5	Mean	0.5681	0.6020	0.5717	0.5984	0.5847
		SE	0.0050	0.0037	0.0042	0.0046	0.0030
	15	Mean	0.5869	0.5760	0.5783	0.5844	0.5809
		SE	0.0026	0.0025	0.0027	0.0029	0.0017
	35	Mean	0.5931	0.5610	0.5815	0.5725	0.5771
		SE	0.0015	0.0014	0.0021	0.0022	0.0009

the MI methods is noticed. For example, the lowest sensitivity when using AB in the model is obtained with missing = 15% for all the three MI methods. Figure 4.1 gives a plot which demonstrates the effect of missing proportions under MI method AREG and marker = Her2. Sensitivity and PPV have a large change over missing proportions, while selectivity, NPV and agreement do not change much. This is generally true under other conditions.

- Markers: For the agreement, NPV and selectivity, AB and Her2 are better and

PR the worst, which roughly follows the proportion of negative to positive for the markers in the dataset (see Table 3.1): the larger the proportion of negatives for the marker, the better the selectivity and agreement. Correspondingly for sensitivity and PPV, PR is the best. See Figure 4.2.

- Comparison of MI methods: Results using all three markers shown in Figures 4.3 to 4.5 suggest that AREG and MICE are generally better than DA. With $marker = AB$ or $Her2$ in the model, DA is slightly better only on PPV; There is no difference for selectivity; and AREG and MICE are significantly better for the other three parameters (sensitivity, NPV and agreement). With $marker = PR$, a consistent result is obtained across all the five parameters that MICE is the best and AREG follows very closely.

4.4.2 Coefficient bias checking

As mentioned previously, we use Cox proportional hazards model to do the survival regression modeling with breast cancer specific survival as the dependent variable. The *true* regression coefficients are obtained using the *true* values (see Section 4.4.1 for the procedure), and the imputed coefficients obtained using the datasets with imputed values. Note that the clinical index HISTOLOGY is excluded from all the survival modeling in this work, mainly due to the small number for lobular cancer. Biases are calculated as relative biases:

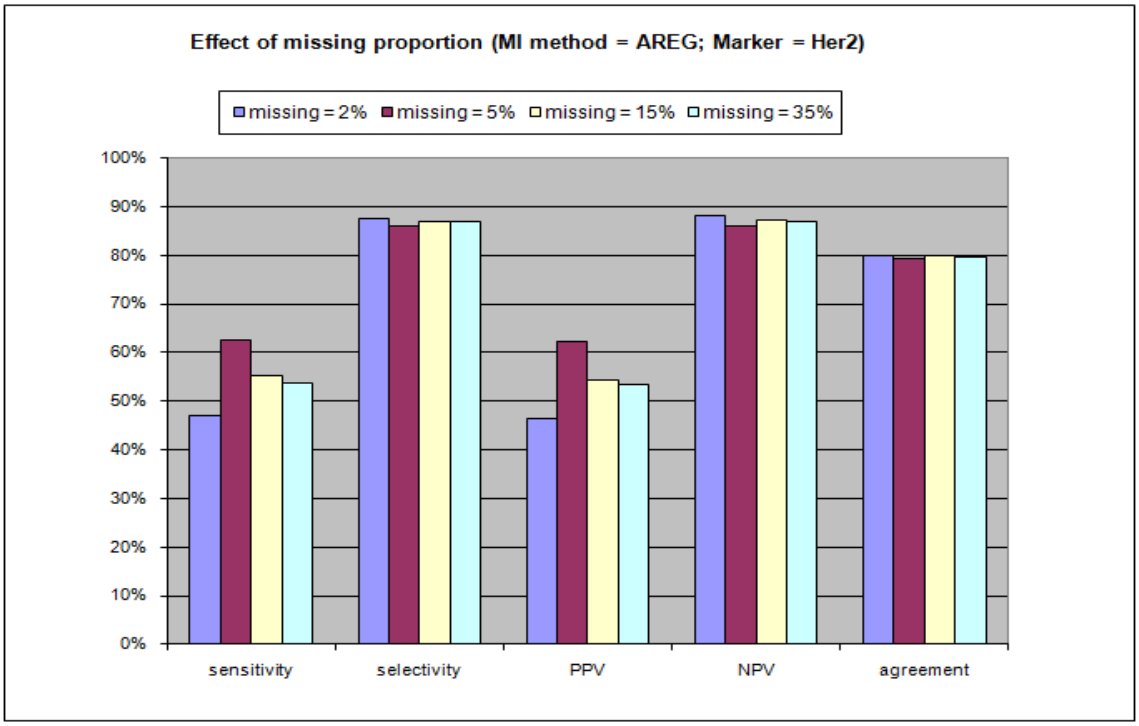


Figure 4.1: Sensitivity analysis: effect of missing proportion (MI method = AREG; marker = Her2).

Bias = 100 % × (imputed coefficient - 'true' coefficient) / 'true' coefficient

The mean and standard error (SE) values for the 100 replications, where each is the average of bias over $m = 5$ multiple imputations, are given in Tables 4.4 to 4.6. Note that the baseline for each variable is the first category of each variable shown in Table 3.2 of Chapter 3. For example, the baseline for AGE is $AGE \leq 50$ and the baseline for SYS is $SYS = 1$. SYS2, SYS3 and SYS4 in the tables denote subsequent categories of the variable SYS.

Some representive results in Tables 4.4 to 4.6 have been plotted for a direct view, see Figures 4.6 to 4.11. Notice the scale change in the figures when comparing biases.

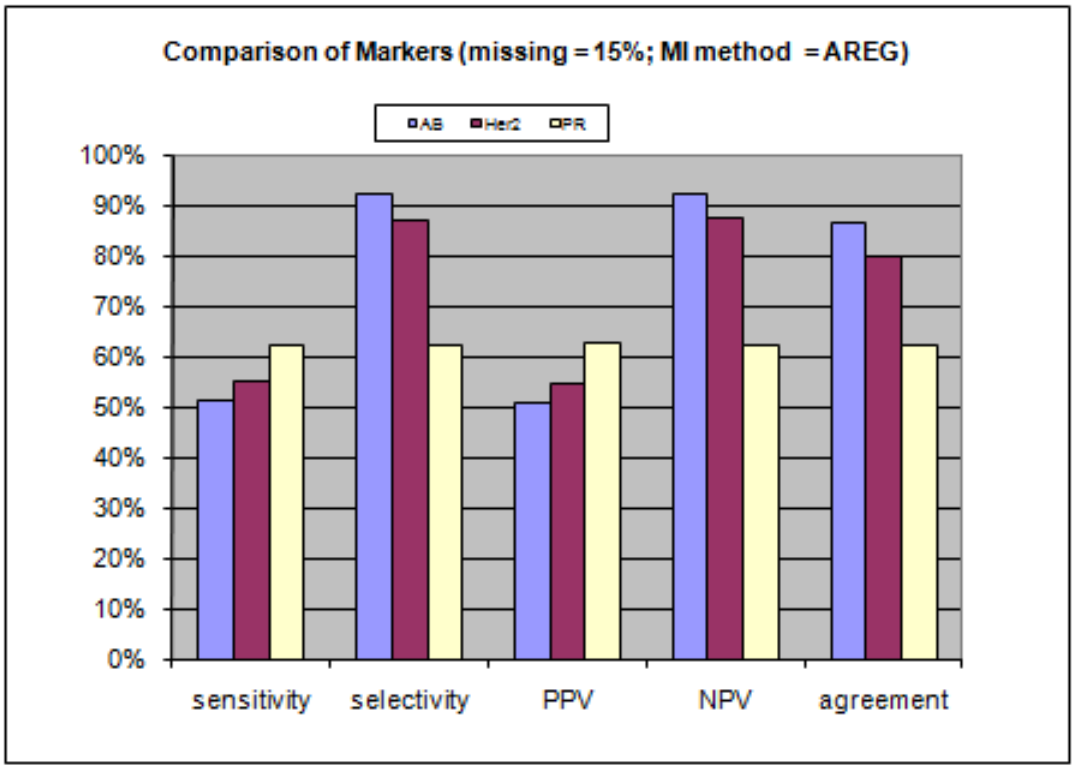


Figure 4.2: Sensitivity analysis: effect of markers (missing proportion = 15%; MI method = AREG).

The following points can be summarized from the comparison of coefficient biases.

- It is not surprising that biases increase significantly as the missing proportion increases. For example, with the MI method = $AREG$ and $marker = AB$, the bias is 4.29% and 31.07% respectively under missing proportions of 5% and 35% for the variable AGE (see Table 4.4). With MI method = DA and $marker = PR$, the bias is 0.09% and 0.54% respectively under missing proportions of 2% and 15% for the variable SIZE (see Table 4.6).

- Biases for AGE and the marker (AB, Her2, or PR) are much larger than those

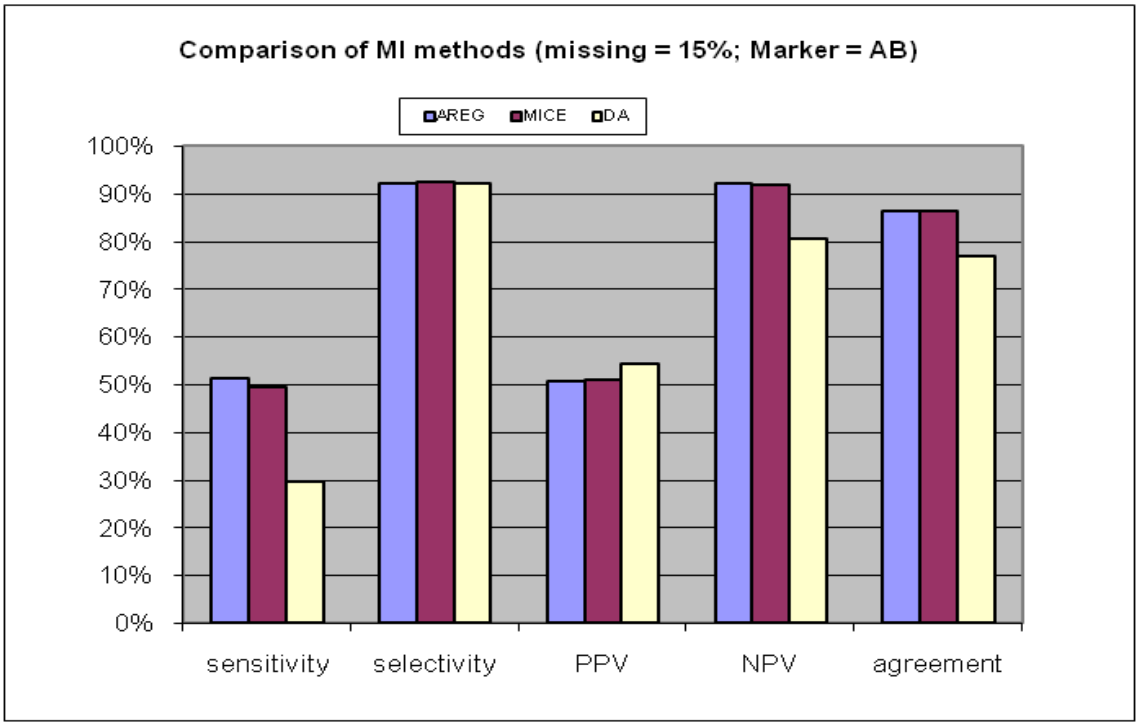


Figure 4.3: Sensitivity analysis: Comparison of MI methods (missing proportion = 15%; marker = AB).

of other variables. Biases for ERPOSNE with some conditions are large also, especially with $marker = PR$. Later in Chapter 5 we show that p-values for AGE and ERPOSNE are large in the full model, suggesting that they are statistically insignificant in the model due to possible correlation with other variables. This could be an important reason for these large biases.

- When using $marker = AB$, no major difference can be found for the three MI methods (see Figures 4.6 and 4.7), although the DA method gives larger biases.

- When using Her2 or PR in the model, AREG and MICE show better results then DA. Figures 4.8 and 4.9 present the results with $marker = Her2$. AREG

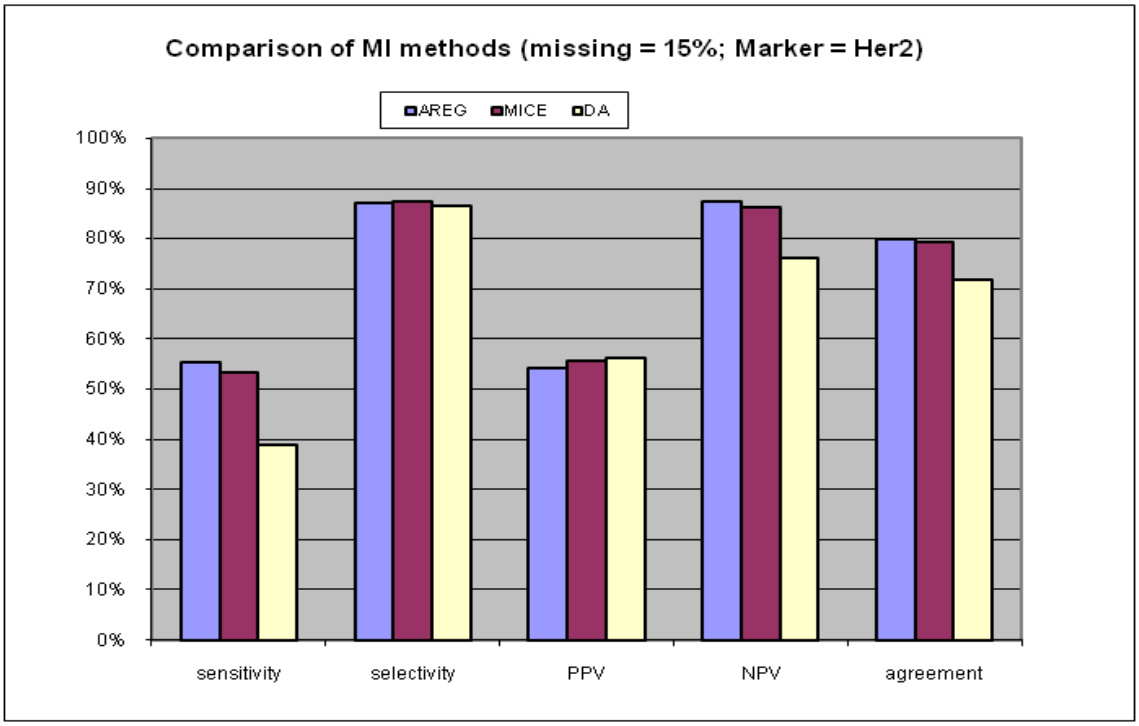


Figure 4.4: Sensitivity analysis: Comparison of MI methods (missing proportion = 15%; marker = Her2).

and MICE methods give lower biases than the DA method for the covariates of SYS4, GRADE2, GRADE3, ERPOSNE and AGE with both $missing = 15\%$ and $missing = 2\%$, and LVNNE with $missing = 2\%$, while biases for other variables are not much different. From Figures 4.10 and 4.11, using PR in the model, AREG and MICE are statistically better for the covariates AGE, GRADE3, and ERPOSNE with $missing = 15\%$ (Figure 4.11), and for the variable ERPONSE only with $missing = 2\%$ (Figure 4.10).

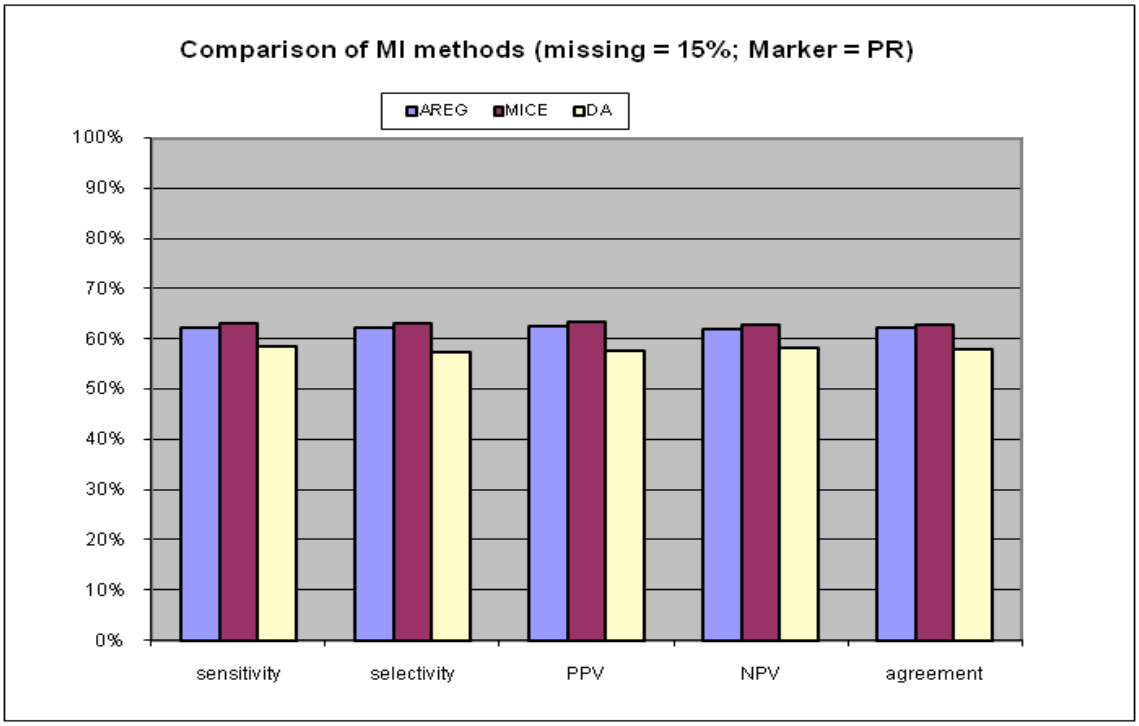
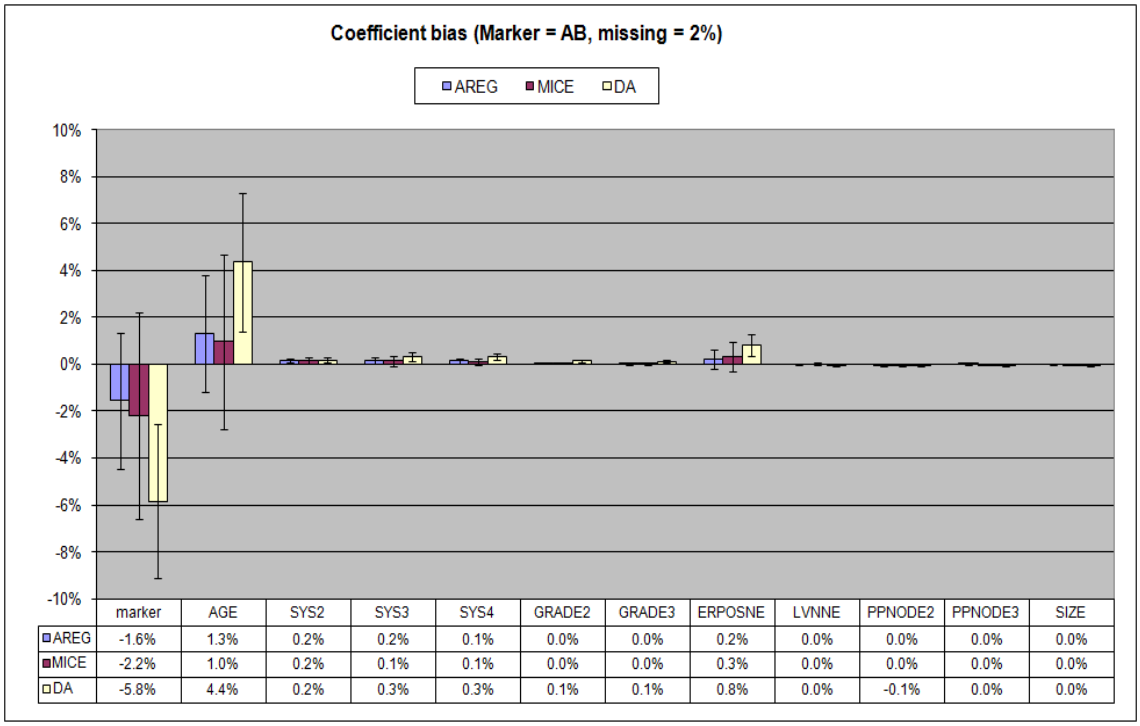


Figure 4.5: Sensitivity analysis: Comparison of MI methods (missing proportion = 15%; marker = PR).

4.4.3 Ten year survival probability

The 10 year breast cancer specific survival probability is calculated at the mean of all the covariates in the Cox model containing all the clinical variables and markers. It is 0.7751 or 77.51% for the dataset with $n = 1575$ after $m = 5$ missing value imputations by the AREG method. One related figure we found from the internet is that the 10 year breast cancer survival proportion is 76% (Well Sphere online, 2010). Table 4.7 gives the results for imputed datasets using different MI methods and markers (AB, Her2, or PR in turn), with missing proportion = 35%. Results show that the 10 year breast cancer specific survival probability is very insensitive to the MI methods and markers.



	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-1.6%	1.3%	0.2%	0.2%	0.1%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%
MICE	-2.2%	1.0%	0.2%	0.1%	0.1%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%
DA	-5.8%	4.4%	0.2%	0.3%	0.3%	0.1%	0.1%	0.8%	0.0%	-0.1%	0.0%	0.0%

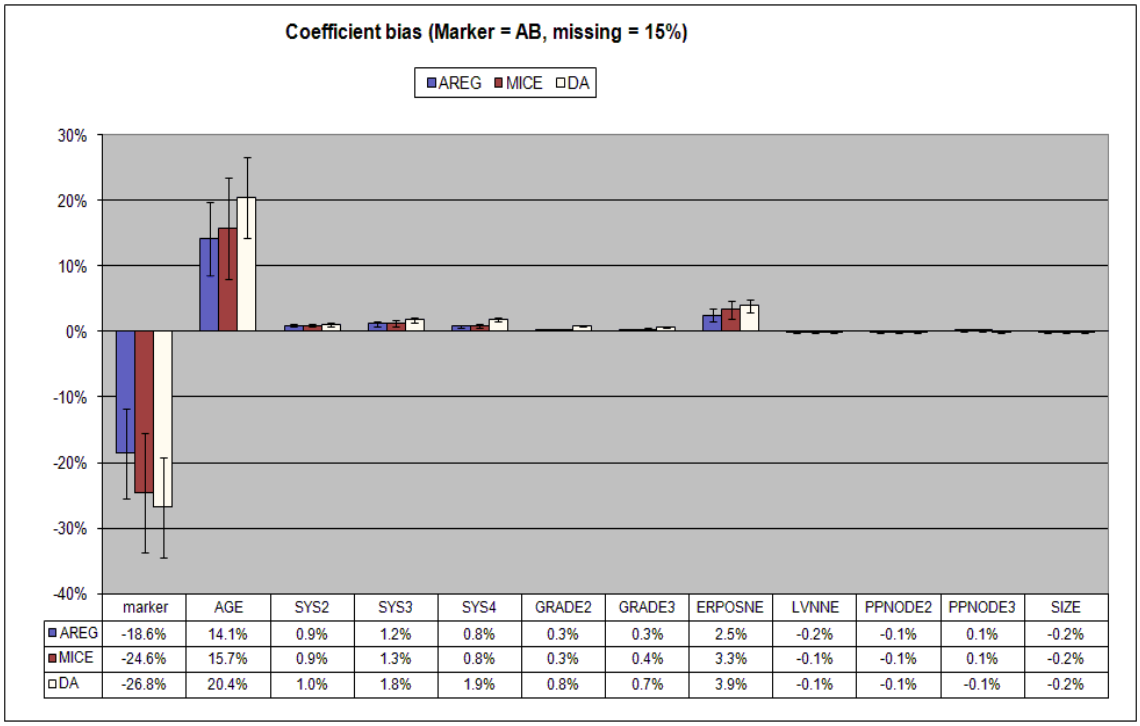
Figure 4.6: Coefficient bias and its 95% confidence interval (missing proportion = 2%; marker = AB).

One of the reasons is that the clinical variables can determine the patients survival time in general so that the values in the marker field do not have much influence.

4.4.4 Results for the dataset without missing values ($n = 910$)

We used the AREG method to generate the *true* dataset in the previous section, which may favor the AREG procedure when imputing the random missing values and correspondingly lead to a better result for the AREG method. Therefore we use the dataset without missing ($n = 910$) to evaluate the MI methods again. This result can also provide us a comparison for the datasets with different sizes.

All the procedures are the same as those using the dataset of $n = 1575$. The only

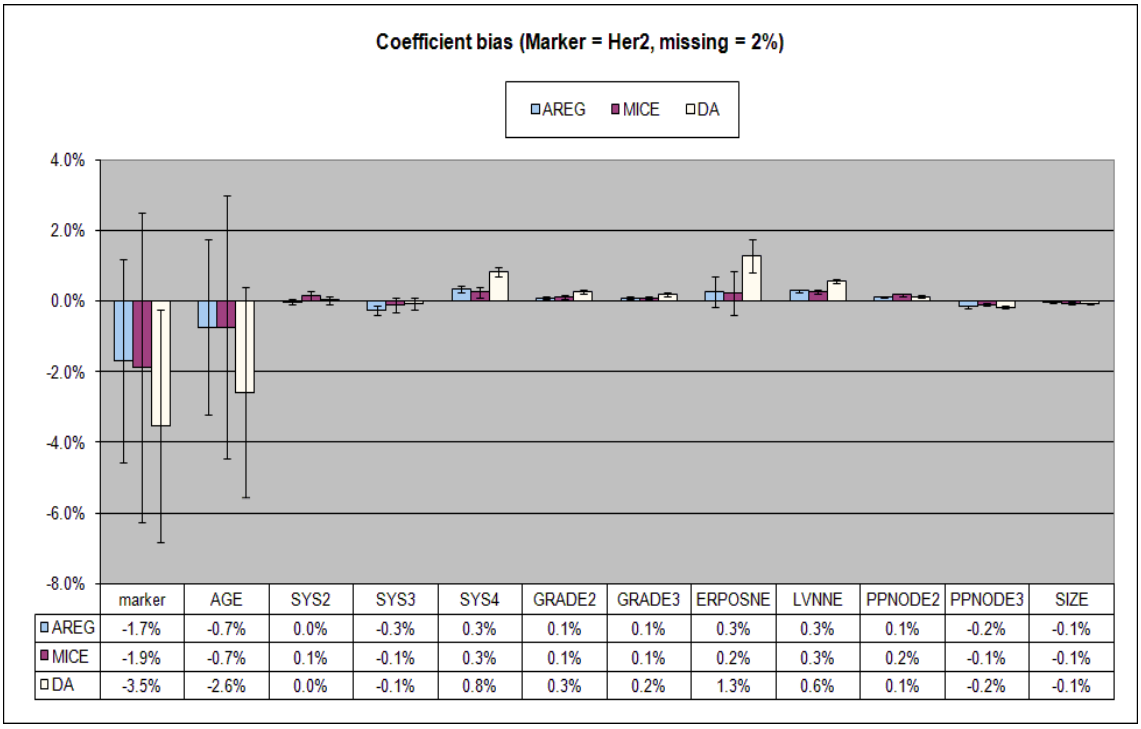


	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-18.6%	14.1%	0.9%	1.2%	0.8%	0.3%	0.3%	2.5%	-0.2%	-0.1%	0.1%	-0.2%
MICE	-24.6%	15.7%	0.9%	1.3%	0.8%	0.3%	0.4%	3.3%	-0.1%	-0.1%	0.1%	-0.2%
DA	-26.8%	20.4%	1.0%	1.8%	1.9%	0.8%	0.7%	3.9%	-0.1%	-0.1%	-0.1%	-0.2%

Figure 4.7: Coefficient bias and its 95% confidence interval (missing proportion = 15%; marker = AB).

difference is that now we have $n = 910$ records in the dataset and imputation is not needed to get the *true* values. Some of the representive results are given here. The complete results can be found in Appendix B.

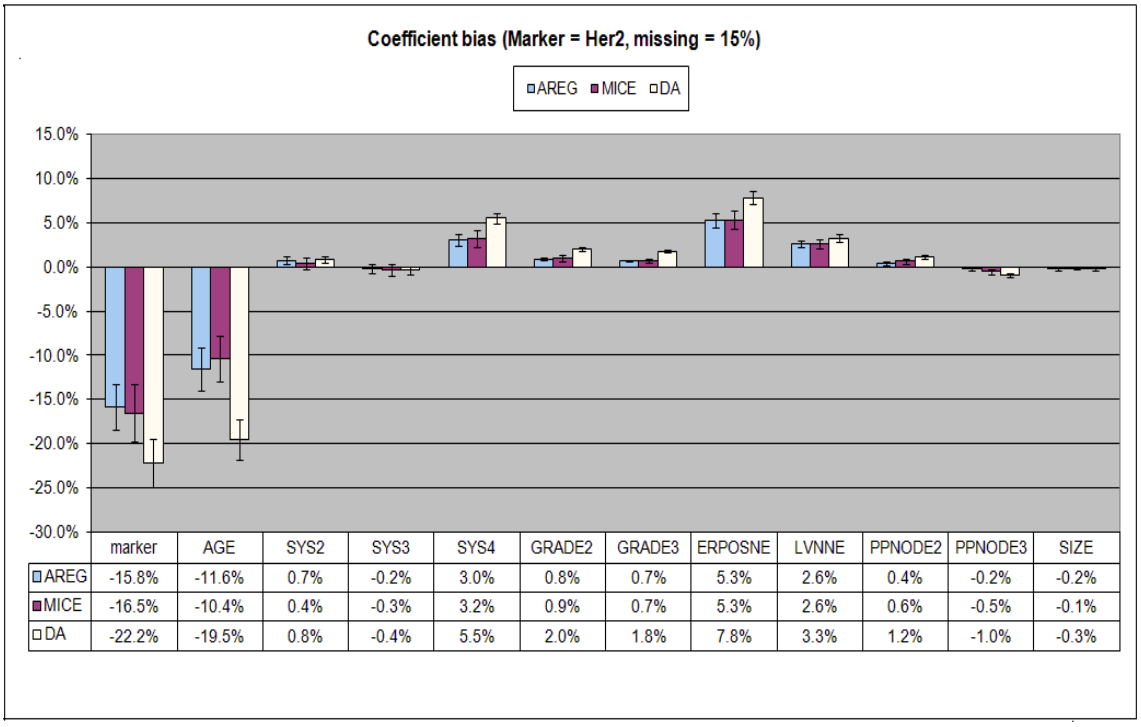
Figures 4.12 to 4.14 show the sensitivity and agreement analysis using the dataset with $n = 910$. The results are similar to those using the larger dataset with $n = 1575$. With marker = Her2 or PR, the values from the two datasets are close to each other (see Figures 4.4, 4.5 and Figure 4.13, 4.14). With marker = AB, the sensitivity and PPV using the small dataset is a bit lower than that using the dataset of $n = 1575$ (see Figure 4.3 and Figure 4.12). Therefore we can draw the same conclusions as those in the section 4.4.1.



	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-1.7%	-0.7%	0.0%	-0.3%	0.3%	0.1%	0.1%	0.3%	0.3%	0.1%	-0.2%	-0.1%
MICE	-1.9%	-0.7%	0.1%	-0.1%	0.3%	0.1%	0.1%	0.2%	0.3%	0.2%	-0.1%	-0.1%
DA	-3.5%	-2.6%	0.0%	-0.1%	0.8%	0.3%	0.2%	1.3%	0.6%	0.1%	-0.2%	-0.1%

Figure 4.8: Coefficient bias and its 95% confidence interval (missing proportion = 2%; marker = Her2).

Figures 4.15 to 4.17 present the results of biases and their 95% confidence intervals using the dataset of $n = 910$. Most of the biases using the $n = 910$ dataset are higher than those using $n = 1575$, with many of them significantly higher, see Figure 4.6 to 4.9 for a comparison with Figures 4.15 to 4.17. With missing proportion = 2%, there is no difference for the biases using different MI methods (see Figures 4.15 and 4.16), which is somewhat different from the conclusion using $n = 1575$ where no difference was found only for marker = AB. AREG and MICE, however, are better than DA with large missing proportions, with lower biases for some variables (see Figure 4.17 and Tables B.4 to B.6 in Appendix B).



	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-15.8%	-11.6%	0.7%	-0.2%	3.0%	0.8%	0.7%	5.3%	2.6%	0.4%	-0.2%	-0.2%
MICE	-16.5%	-10.4%	0.4%	-0.3%	3.2%	0.9%	0.7%	5.3%	2.6%	0.6%	-0.5%	-0.1%
DA	-22.2%	-19.5%	0.8%	-0.4%	5.5%	2.0%	1.8%	7.8%	3.3%	1.2%	-1.0%	-0.3%

Figure 4.9: Coefficient bias and its 95% confidence interval (missing proportion = 15%; marker = Her2).

Table 4.4: Coefficient bias with MI method = AREG.

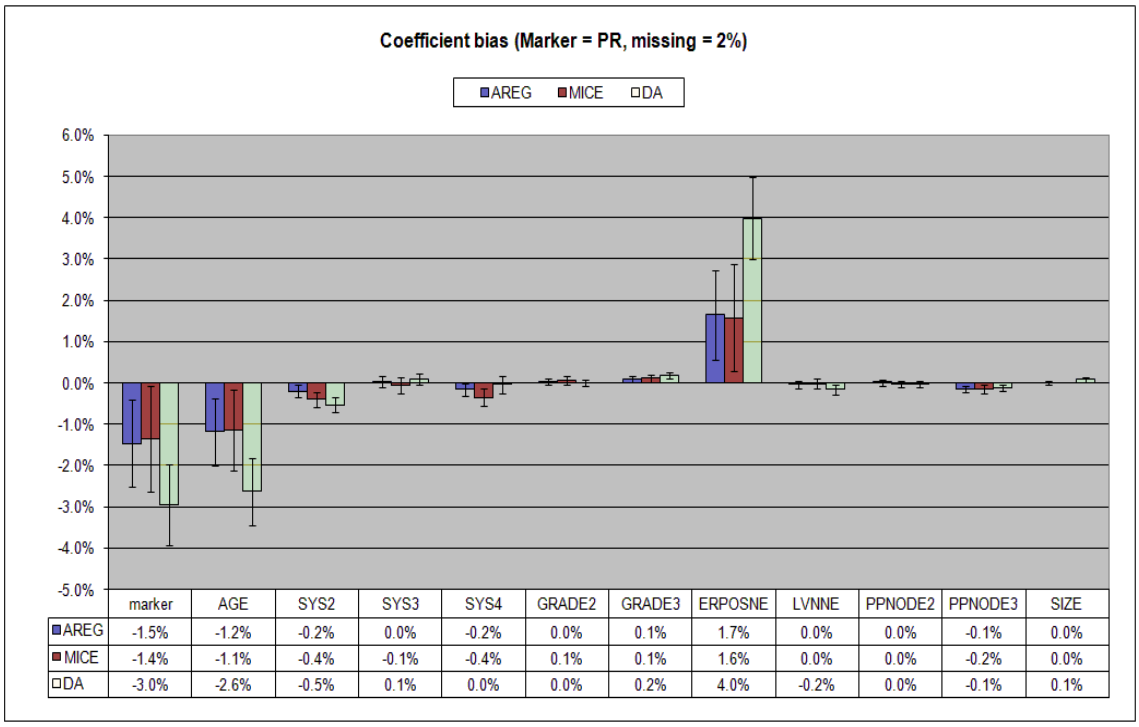
Missing %	Variable	AB		Her2		PR	
		Mean (%)	SE (%)	Mean (%)	SE (%)	Mean (%)	SE (%)
2	marker	-1.56	1.44	-1.69	0.54	-1.46	0.53
	AGE	1.31	1.24	-0.74	0.53	-1.18	0.41
	SYS2	0.15	0.04	-0.02	0.08	-0.20	0.08
	SYS3	0.16	0.07	-0.26	0.10	0.04	0.07
	SYS4	0.13	0.05	0.34	0.12	-0.16	0.08
	GRADE2	0.03	0.01	0.09	0.03	0.03	0.04
	GRADE3	0.02	0.02	0.09	0.03	0.10	0.04
	ERPOSNE	0.22	0.21	0.26	0.19	1.65	0.55
	LVNNE	0.00	0.02	0.29	0.09	-0.03	0.05
	PPNODE2	-0.04	0.012	0.12	0.05	0.01	0.04
	PPNODE3	0.01	0.015	-0.16	0.04	-0.14	0.04
	SIZE	0.00	0.01	-0.05	0.03	0.00	0.03
5	marker	-5.19	2.51	-4.89	0.89	-4.74	0.83
	AGE	4.29	2.15	-4.47	0.83	-3.58	0.63
	SYS2	0.25	0.07	0.41	0.15	-0.72	0.14
	SYS3	0.35	0.13	-0.06	0.18	-0.06	0.11
	SYS4	0.22	0.08	0.87	0.17	-0.50	0.12
	GRADE2	0.08	0.03	0.28	0.05	-0.08	0.07
	GRADE3	0.08	0.03	0.20	0.05	0.21	0.06
	ERPOSNE	0.62	0.36	1.76	0.28	5.21	0.88
	LVNNE	-0.05	0.05	0.78	0.14	-0.18	0.07
	PPNODE2	-0.04	0.02	0.19	0.06	-0.08	0.05
	PPNODE3	0.02	0.02	-0.05	0.05	-0.27	0.06
	SIZE	-0.04	0.02	-0.06	0.05	0.14	0.05
15	marker	-18.61	3.43	-15.83	1.26	-14.00	1.24
	AGE	14.14	2.79	-11.58	1.22	-9.58	0.94
	SYS2	0.91	0.11	0.73	0.24	-2.18	0.22
	SYS3	1.18	0.17	-0.16	0.25	-0.31	0.17
	SYS4	0.76	0.12	3.04	0.31	-1.34	0.21
	GRADE2	0.26	0.04	0.84	0.07	0.04	0.12
	GRADE3	0.29	0.03	0.70	0.06	0.76	0.10
	ERPOSNE	2.47	0.50	5.27	0.41	14.72	1.27
	LVNNE	-0.15	0.06	2.58	0.20	-0.48	0.11
	PPNODE2	-0.12	0.03	0.44	0.12	0.02	0.08
	PPNODE3	0.10	0.04	-0.21	0.11	-0.96	0.09
	SIZE	-0.15	0.03	-0.23	0.07	0.31	0.06
35	marker	-39.07	5.64	-32.57	1.66	-36.41	1.68
	AGE	31.07	4.44	-24.02	1.46	-27.28	1.22
	SYS2	1.63	0.13	1.05	0.26	-4.89	0.24
	SYS3	2.25	0.26	-1.01	0.33	-0.5	0.21
	SYS4	1.46	0.17	5.40	0.36	-3.05	0.27
	GRADE2	0.55	0.04	1.72	0.07	0.10	0.12
	GRADE3	0.57	0.05	1.54	0.08	2.01	0.11
	ERPOSNE	4.82	0.86	9.93	0.56	37.49	1.75
	LVNNE	-0.29	0.11	5.22	0.24	-1.28	0.16
	PPNODE2	-0.28	0.04	1.13	0.13	-0.37	0.11
	PPNODE3	0.10	0.04	-0.73	0.12	-1.83	0.12
	SIZE	-0.27	0.04	-0.76	0.09	0.91	0.08

Table 4.5: Coefficient bias with MI method = MICE.

Missing %	Variable	AB		Her2		PR	
		Mean (%)	SE (%)	Mean (%)	SE (%)	Mean (%)	SE (%)
2	marker	-2.20	2.19	-1.89	0.74	-1.35	0.64
	AGE	0.95	1.86	-0.74	0.63	-1.14	0.49
	SYS2	0.15	0.06	0.14	0.10	-0.40	0.09
	SYS3	0.14	0.11	-0.11	0.13	-0.05	0.10
	SYS4	0.09	0.07	0.25	0.13	-0.35	0.10
	GRADE2	0.04	0.02	0.11	0.04	0.05	0.05
	GRADE3	0.03	0.02	0.09	0.04	0.11	0.05
	ERPOSNE	0.33	0.31	0.23	0.24	1.58	0.65
	LVNNE	0.00	0.03	0.27	0.13	-0.02	0.06
	PPNODE2	-0.04	0.02	0.17	0.06	-0.03	0.04
	PPNODE3	-0.01	0.02	-0.10	0.06	-0.15	0.06
	SIZE	-0.02	0.02	-0.07	0.04	0.00	0.04
5	marker	-6.52	2.94	-4.79	1.10	-5.06	1.04
	AGE	5.22	2.55	-3.94	1.10	-3.70	0.80
	SYS2	0.23	0.11	0.42	0.20	-0.70	0.19
	SYS3	0.40	0.16	-0.11	0.24	-0.37	0.15
	SYS4	0.24	0.12	0.88	0.24	-0.49	0.16
	GRADE2	0.12	0.04	0.19	0.07	-0.01	0.09
	GRADE3	0.10	0.04	0.12	0.06	0.29	0.08
	ERPOSNE	0.79	0.43	1.76	0.33	5.20	1.10
	LVNNE	-0.05	0.06	0.74	0.17	-0.23	0.09
	PPNODE2	-0.07	0.03	0.26	0.08	0.04	0.07
	PPNODE3	0.05	0.03	-0.05	0.09	-0.36	0.09
	SIZE	-0.06	0.03	-0.04	0.07	0.13	0.05
15	marker	-24.56	4.54	-16.52	1.64	-13.24	1.65
	AGE	15.70	3.83	-10.38	1.32	-8.26	1.38
	SYS2	0.94	0.14	0.41	0.33	-1.96	0.28
	SYS3	1.29	0.24	-0.33	0.31	-0.42	0.26
	SYS4	0.77	0.15	3.15	0.48	-1.07	0.29
	GRADE2	0.34	0.06	0.94	0.19	-0.01	0.14
	GRADE3	0.35	0.05	0.70	0.13	0.65	0.13
	ERPOSNE	3.30	0.66	5.30	0.52	14.07	1.75
	LVNNE	-0.09	0.08	2.58	0.27	-0.14	0.18
	PPNODE2	-0.12	0.04	0.64	0.13	-0.01	0.10
	PPNODE3	0.05	0.05	-0.53	0.15	-1.01	0.11
	SIZE	-0.16	0.04	-0.13	0.09	0.14	0.09
35	marker	-39.65	7.74	-33.14	2.61	-36.60	2.28
	AGE	26.83	6.47	-22.62	2.06	-24.58	1.76
	SYS2	1.69	0.17	0.88	0.40	-4.99	0.35
	SYS3	2.27	0.34	-1.47	0.46	-0.94	0.30
	SYS4	1.52	0.21	4.44	0.50	-2.97	0.33
	GRADE2	1.3	0.29	2.26	0.36	-0.06	0.19
	GRADE3	1.00	0.20	1.80	0.25	1.91	0.17
	ERPOSNE	5.12	1.17	9.40	0.86	37.60	2.41
	LVNNE	-0.24	0.15	4.83	0.37	-1.23	0.22
	PPNODE2	-0.25	0.08	1.18	0.17	-0.20	0.13
	PPNODE3	0.23	0.10	-0.74	0.17	-1.98	0.14
	SIZE	-0.33	0.07	-0.81	0.13	0.78	0.11

Table 4.6: Coefficient bias with MI method = DA.

Missing %	Variable	AB		Her2		PR	
		Mean (%)	SE (%)	Mean (%)	SE (%)	Mean (%)	SE (%)
2	marker	-5.84	1.65	-3.54	0.56	-2.95	0.49
	AGE	4.35	1.49	-2.58	0.52	-2.63	0.40
	SYS2	0.16	0.06	0.02	0.10	-0.53	0.09
	SYS3	0.33	0.09	-0.07	0.09	0.09	0.07
	SYS4	0.33	0.06	0.83	0.14	-0.03	0.11
	GRADE2	0.13	0.03	0.27	0.05	0.00	0.04
	GRADE3	0.12	0.02	0.19	0.04	0.18	0.04
	ERPOSNE	0.80	0.24	1.28	0.18	3.99	0.50
	LVNNE	-0.04	0.03	0.56	0.08	-0.16	0.06
	PPNODE2	-0.05	0.02	0.13	0.05	-0.03	0.04
	PPNODE3	-0.03	0.02	-0.18	0.04	-0.12	0.04
	SIZE	-0.04	0.01	-0.07	0.03	0.09	0.03
5	marker	-5.03	2.63	-6.24	0.75	-7.24	0.88
	AGE	4.66	2.22	-5.73	0.81	-6.30	0.67
	SYS2	0.25	0.08	0.16	0.16	-1.34	0.13
	SYS3	0.49	0.14	-0.04	0.15	-0.11	0.12
	SYS4	0.52	0.10	2.29	0.21	-0.43	0.14
	GRADE2	0.28	0.04	0.73	0.11	0.14	0.07
	GRADE3	0.22	0.03	0.62	0.08	0.47	0.06
	ERPOSNE	0.72	0.38	2.32	0.25	9.39	0.88
	LVNNE	0.02	0.04	0.99	0.14	-0.29	0.08
	PPNODE2	-0.03	0.03	0.28	0.08	-0.08	0.05
	PPNODE3	-0.02	0.03	-0.24	0.07	-0.34	0.06
	SIZE	-0.01	0.03	-0.05	0.05	0.16	0.05
15	marker	-26.82	3.81	-22.19	1.36	-19.44	1.48
	AGE	20.40	3.06	-19.48	1.13	-18.51	1.00
	SYS2	1.03	0.14	0.84	0.22	-2.82	0.23
	SYS3	1.82	0.20	-0.39	0.24	-0.37	0.19
	SYS4	1.86	0.12	5.48	0.30	-0.99	0.23
	GRADE2	0.77	0.05	2.03	0.13	0.41	0.11
	GRADE3	0.67	0.04	1.76	0.09	1.38	0.09
	ERPOSNE	3.88	0.53	7.80	0.38	24.68	1.50
	LVNNE	-0.06	0.07	3.26	0.20	-0.84	0.12
	PPNODE2	-0.07	0.04	1.15	0.10	0.03	0.08
	PPNODE3	-0.14	0.04	-0.96	0.10	-1.11	0.10
	SIZE	-0.19	0.03	-0.25	0.07	0.54	0.06
35	marker	-43.38	4.80	-43.47	1.81	-46.55	2.12
	AGE	33.17	3.97	-32.03	1.21	-37.79	1.16
	SYS2	1.60	0.14	1.77	0.31	-6.38	0.24
	SYS3	3.19	0.20	-0.62	0.31	-0.14	0.20
	SYS4	3.32	0.14	10.49	0.31	-1.51	0.23
	GRADE2	1.29	0.09	3.29	0.12	0.50	0.10
	GRADE3	1.18	0.05	3.07	0.08	2.94	0.10
	ERPOSNE	6.42	0.67	14.31	0.52	56.71	1.91
	LVNNE	-0.25	0.11	6.34	0.24	-2.01	0.16
	PPNODE2	-0.34	0.06	1.76	0.11	-0.07	0.09
	PPNODE3	-0.13	0.05	-1.37	0.12	-2.00	0.11
	SIZE	-0.15	0.04	-0.43	0.07	1.22	0.089

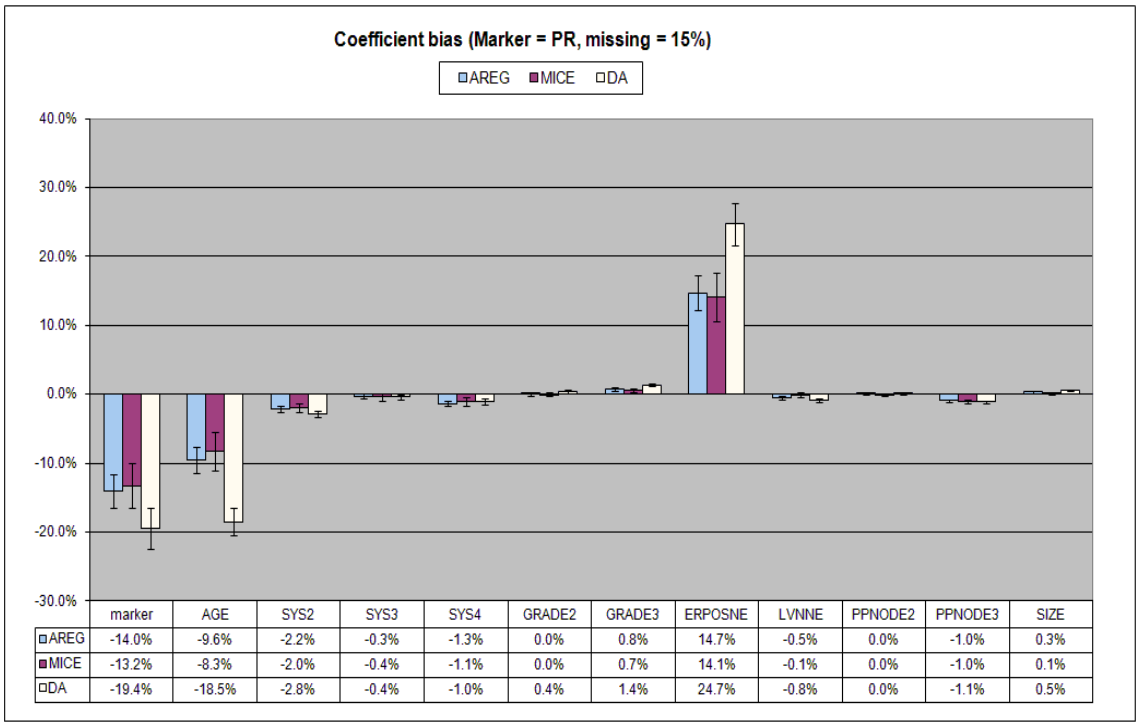


	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-1.5%	-1.2%	-0.2%	0.0%	-0.2%	0.0%	0.1%	1.7%	0.0%	0.0%	-0.1%	0.0%
MICE	-1.4%	-1.1%	-0.4%	-0.1%	-0.4%	0.1%	0.1%	1.6%	0.0%	0.0%	-0.2%	0.0%
DA	-3.0%	-2.6%	-0.5%	0.1%	0.0%	0.0%	0.2%	4.0%	-0.2%	0.0%	-0.1%	0.1%

Figure 4.10: Coefficient bias and its 95% confidence interval (missing proportion = 2%; marker = PR).

Table 4.7: 10 year survival probability at the mean of all the covariates (missing proportion = 35%).

marker	AREG	MICE	DA
AB	0.7744	0.7744	0.7745
Her2	0.7745	0.7745	0.7746
PR	0.7747	0.7747	0.7746



	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-14.0%	-9.6%	-2.2%	-0.3%	-1.3%	0.0%	0.8%	14.7%	-0.5%	0.0%	-1.0%	0.3%
MICE	-13.2%	-8.3%	-2.0%	-0.4%	-1.1%	0.0%	0.7%	14.1%	-0.1%	0.0%	-1.0%	0.1%
DA	-19.4%	-18.5%	-2.8%	-0.4%	-1.0%	0.4%	1.4%	24.7%	-0.8%	0.0%	-1.1%	0.5%

Figure 4.11: Coefficient bias and its 95% confidence interval (missing proportion = 15%; marker = PR).

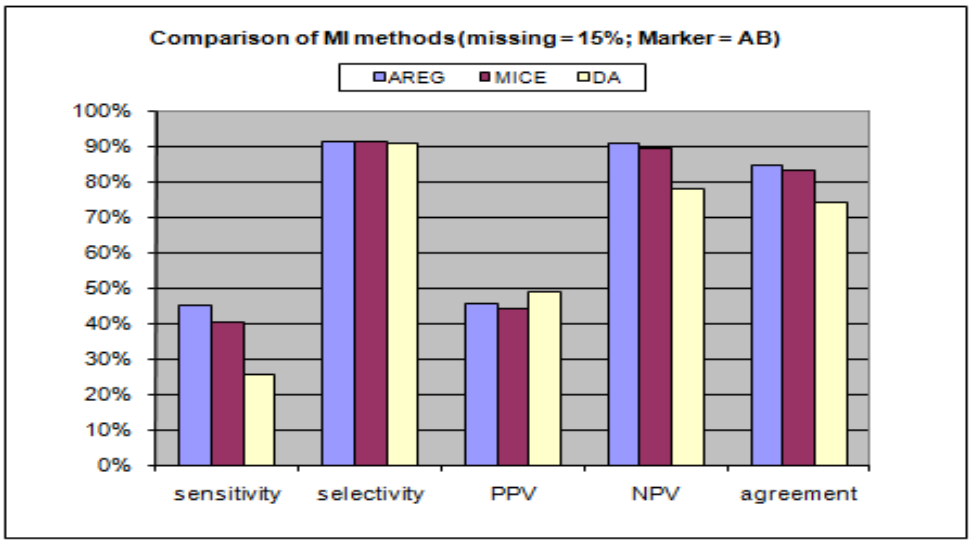


Figure 4.12: Sensitivity analysis: Comparison of MI methods with $n = 910$ (missing proportion = 15%; marker = AB).

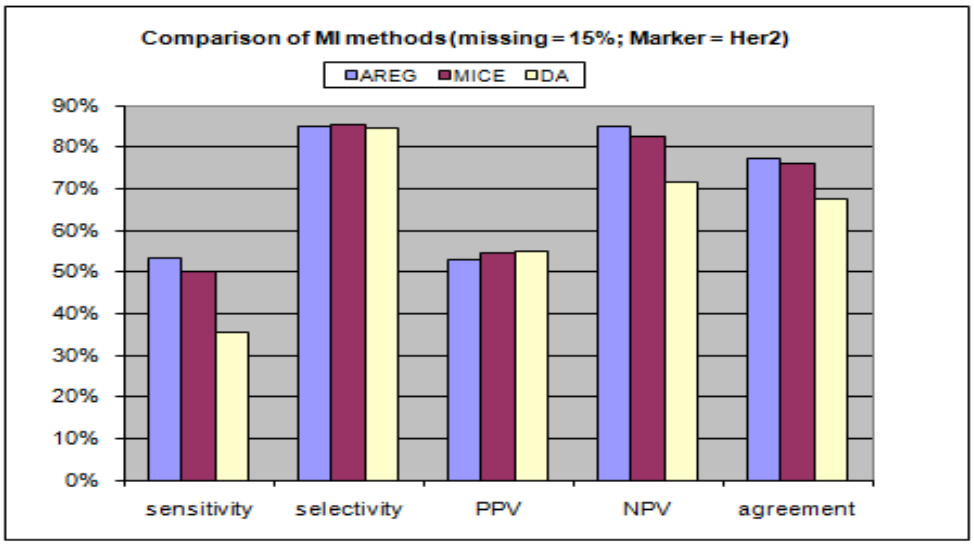


Figure 4.13: Sensitivity analysis: Comparison of MI methods with $n = 910$ *(missing proportion = 15%; marker = Her2).*

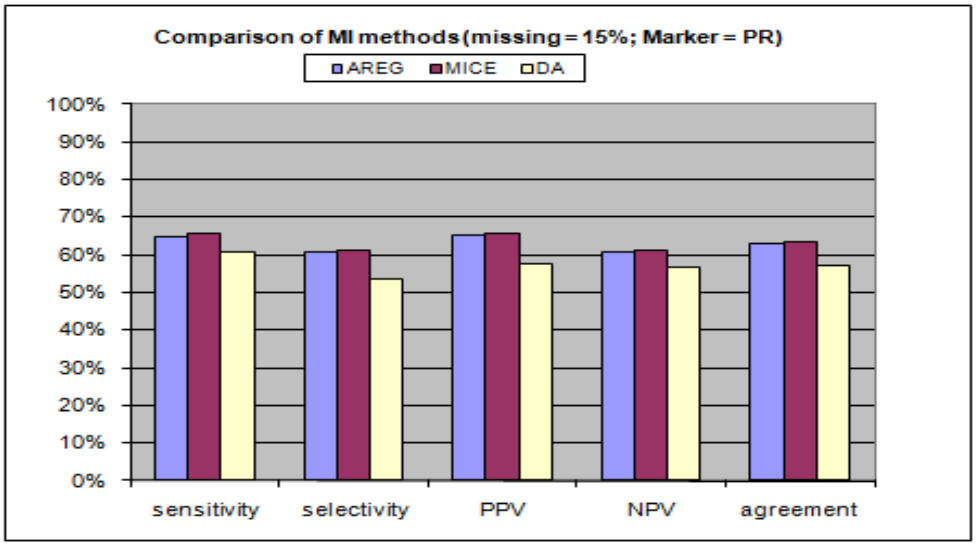
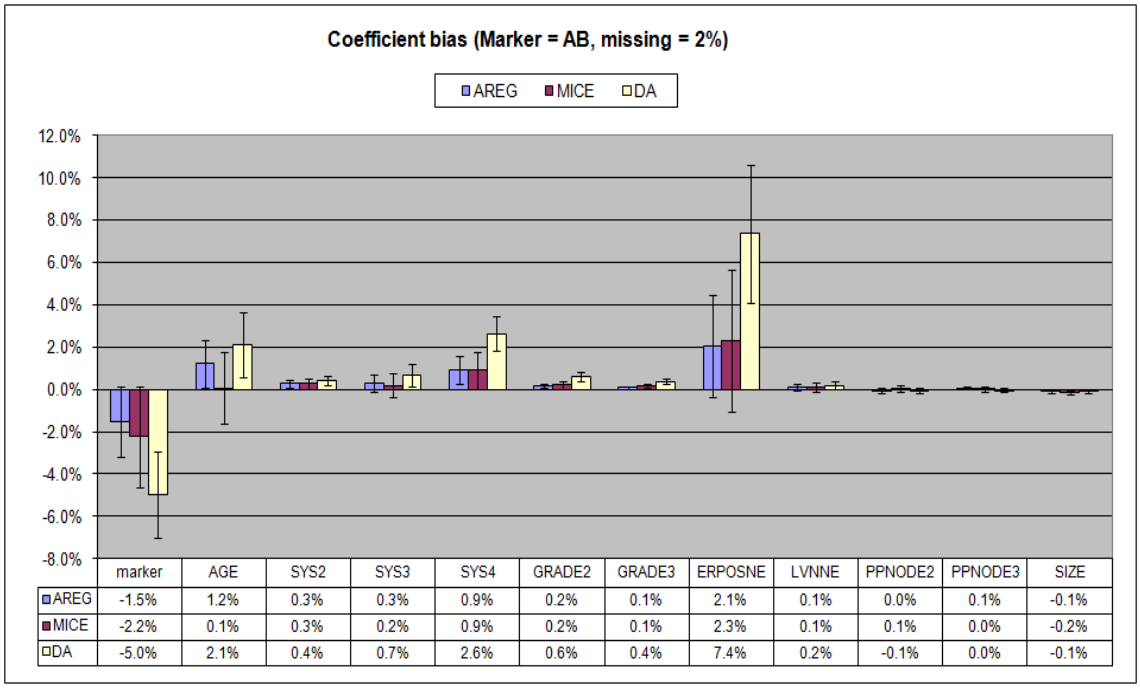
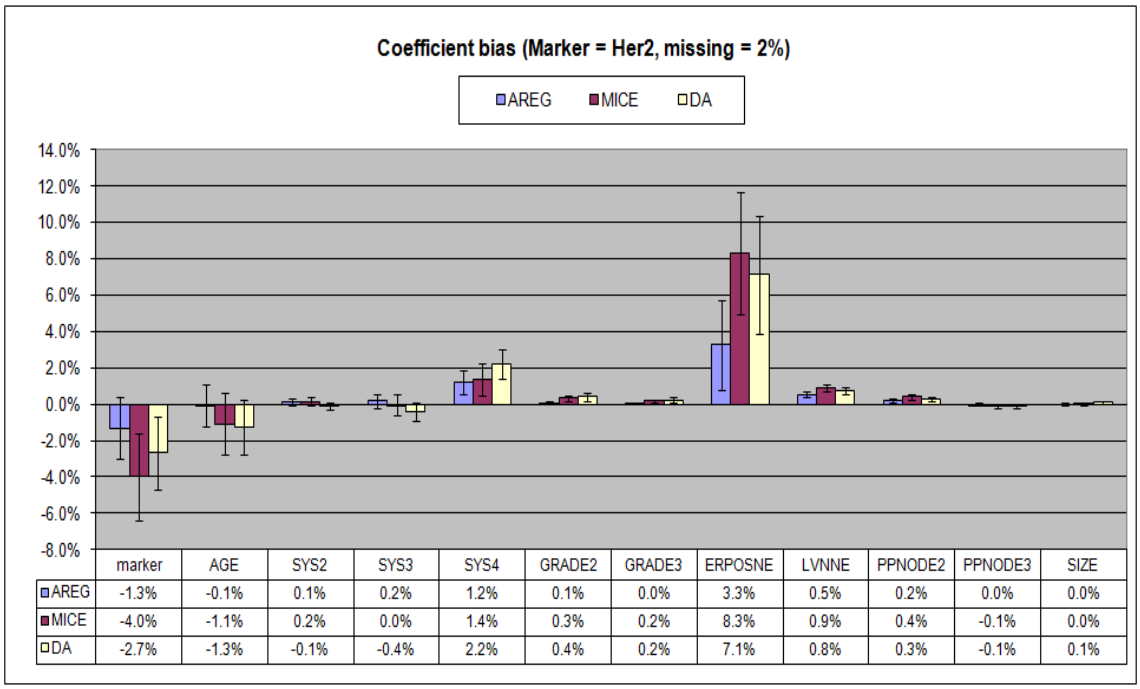


Figure 4.14: Sensitivity analysis: Comparison of MI methods with $n = 910$ *(missing proportion = 15%; marker = PR).*



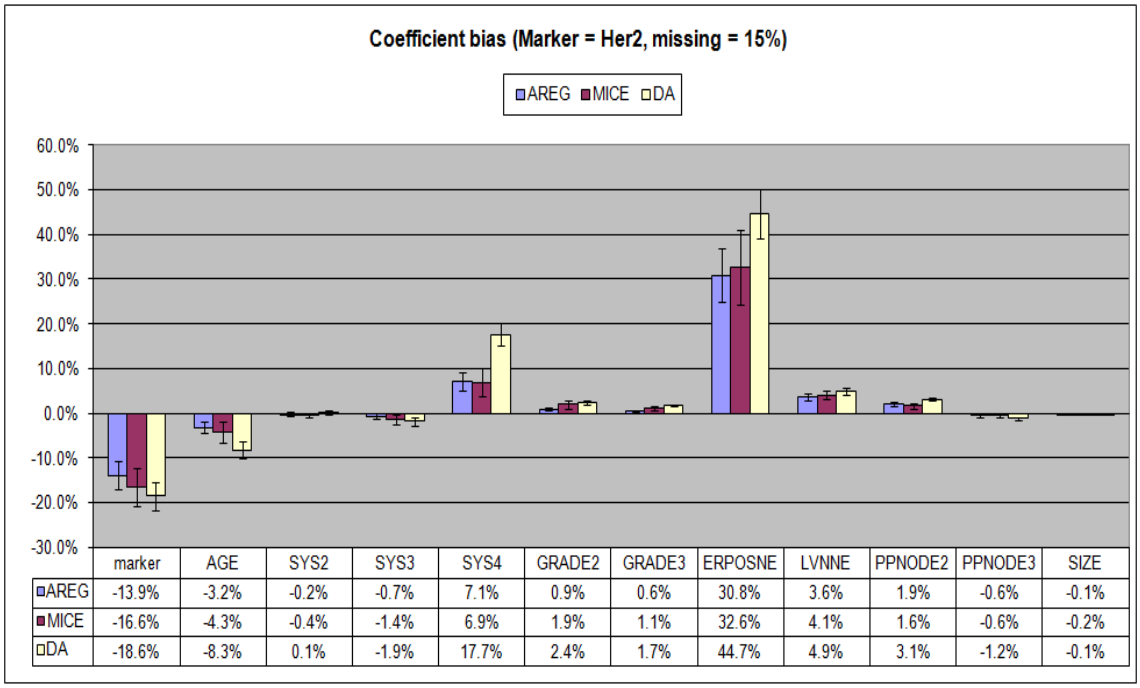
	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-1.5%	1.2%	0.3%	0.3%	0.9%	0.2%	0.1%	2.1%	0.1%	0.0%	0.1%	-0.1%
MICE	-2.2%	0.1%	0.3%	0.2%	0.9%	0.2%	0.1%	2.3%	0.1%	0.1%	0.0%	-0.2%
DA	-5.0%	2.1%	0.4%	0.7%	2.6%	0.6%	0.4%	7.4%	0.2%	-0.1%	0.0%	-0.1%

Figure 4.15: Coefficient bias and its 95% confidence interval with $n = 910$ (missing proportion = 2%; marker = AB).



	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-1.3%	-0.1%	0.1%	0.2%	1.2%	0.1%	0.0%	3.3%	0.5%	0.2%	0.0%	0.0%
MICE	-4.0%	-1.1%	0.2%	0.0%	1.4%	0.3%	0.2%	8.3%	0.9%	0.4%	-0.1%	0.0%
DA	-2.7%	-1.3%	-0.1%	-0.4%	2.2%	0.4%	0.2%	7.1%	0.8%	0.3%	-0.1%	0.1%

Figure 4.16: Coefficient bias and its 95% confidence interval with $n = 910$ (missing proportion = 2%; marker = Her2).



	marker	AGE	SYS2	SYS3	SYS4	GRADE2	GRADE3	ERPOSNE	LVNNE	PPNODE2	PPNODE3	SIZE
AREG	-13.9%	-3.2%	-0.2%	-0.7%	7.1%	0.9%	0.6%	30.8%	3.6%	1.9%	-0.6%	-0.1%
MICE	-16.6%	-4.3%	-0.4%	-1.4%	6.9%	1.9%	1.1%	32.6%	4.1%	1.6%	-0.6%	-0.2%
DA	-18.6%	-8.3%	0.1%	-1.9%	17.7%	2.4%	1.7%	44.7%	4.9%	3.1%	-1.2%	-0.1%

Figure 4.17: Coefficient bias and its 95% confidence interval with $n = 910$ *(missing proportion = 15%; marker = Her2).*

Chapter 5

Survival Analysis

5.1 Introduction

Survival analysis involves the modeling of time to event data. It has been a very active research field for several decades. An important contribution that stimulated the entire field was the counting process formulation given by Aalen (1975). The flexibility of a counting process is that it allows modeling multiple (or recurrent) events. Since then a large number of fine text books have been written on survival analysis and counting processes, with some key references being Andersen *et al.* (1993), Fleming and Harrington (1991), Kalbfleisch and Prentice (2002), and Lawless (2003). Excellent texts aimed at the biostatistical community with biomedical application as the motivating factor include Klein and Moeschberger (1997), Therneau and Grambsch (2000), Tableman and Kim (2003), and Martinussen and Scheike (2006).

Survival analysis generally deals with censored data, that is, when the time to the event is not observed. As an example, when we deal with a lifetime problem, ideally both the birth and death dates of a subject are known, in which case the lifetime is known. If it is known only that the date of death is after some date, this is called right censoring. Right censoring occurs for those subjects whose birth date is known but who are still alive when they are lost to follow-up or when the study ends. If a subject's lifetime is known to be less than a certain duration, the lifetime is said to be left-censored. It may also happen that subjects with a lifetime less than some threshold may not be observed at all: this is called truncation. Note that truncation is different from left censoring, since for a left censored datum, we know the subject exists, but for a truncated datum, we may be completely unaware of the subject. In a so-called delayed entry study, subjects are not observed at all until they have reached a certain age. For example, people may not be observed until they have reached the age to enter school. Any deceased subjects in the pre-school age group would be unknown. Survival analysis represents a collection of statistical procedures which accommodate time to event censored (incomplete) data so that reliable and accurate information can be obtained.

The survival function, $S(t)$, gives the probability that a subject will survive past time t:

$$S(t) = P(T \geq t) = 1 - F(t) = \int f(x)dx, \tag{5.1}$$

where T is a non-negative random variable denoting the time of event occurring, $F(t)$

denotes the cumulative distribution function of T with corresponding probability density function (p.d.f) $f(t)$. Conversely we can express the p.d.f as

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \tag{5.2}$$

The hazard function, $h(t)$, is the instantaneous rate at which events occur, given no previous events.

$$h(t) = \lim_{\delta t \to 0^+} \frac{P\left(t \leq T < t + \delta t \mid T \geq t\right)}{\delta t} = \frac{f(t)}{S(t)}. \tag{5.3}$$

The hazard function is also referred to as risk or mortality rate in the health care field (Tableman and Kim 2003). It must be non-negative, and its integral over t must be infinite, but is not otherwise constrained.

It is easily verified that $h(t)$ specifies the distribution of T, since

$$h(t) = -\frac{dS(t)/dt}{S(t)} = -\frac{dlog\left(S(t)\right)}{dt}. \tag{5.4}$$

The cumulative hazard function, $H(t)$, can be obtained by integrating $h(u)$ over $(0, t)$:

$$H(t) = \int_0^t h(u)du = -log\left(S(t)\right). \tag{5.5}$$

If we assume that every subject follows the same survival function (no covariates or other individual differences), we can easily estimate $S(t)$ and $H(t)$, using nonparametric methods like the Kaplan-Meier estimator (Kaplan and Meier 1958) or parametric methods by making parametric assumptions such as exponential, Weibull, Gamma, or log-normal. Details can be found in many of the text books on survival analysis, such as Tableman and Kim (2003) or Klein and Moeschberger (1997).

If there is no censoring, standard regression procedures could be used. However, these may be inadequate because:

- Time to event is restricted to be positive and has a skewed distribution.
- The probability of surviving past a certain point in time may be of more interest than the expected time of event.
- The hazard function, used for regression in survival analysis, can lend more insight into the failure mechanism than linear regression.

5.2 Cox Proportional Hazard Regression Model

The Cox proportional Hazards model (Cox 1972) is probably the most widely used method for modeling survival data. For data with one explanatory variable, i.e. one covariate, non-parametric methods like plotting of Kaplan-Meier survival probabilities may be adequate if the groups being compared are reasonably similar. Frequently however, the groups being compared differ in many respects. They may have different age distributions, different proportion of men and women, different smoking habits etc. These differences come in addition to the covariate we are really interested in, and the analysis must be adjusted to compensate for these other differences, which may otherwise confound the analysis. The Cox proportional hazards model is a semi-parametric model for fitting survival data.

Let T denote a continuous non-negative random variable representing survival time. The basic Cox model is as follows:

$$h(t \mid \mathbf{Z}) = h_0(t) exp(\beta^{'} \mathbf{Z}), \tag{5.6}$$

where $h_0(t)$ is the baseline hazard which may vary arbitrarily over time, and $\mathbf{Z}$ is the covariate vector. The covariates may be time-dependent but are here assumed to be fixed at the start of study. $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of covariate coefficients. The baseline hazard is treated non-parametrically, but the individual covariate effects β_p are assumed to be constant throughout the study, hence the notation semi-parametric. The model is often called the proportional hazards model because of this constant covariate effect throughout the study. If two individuals are compared that have covariate values $\mathbf{Z}$ and $\mathbf{Z}^*$, the ratio of their hazard rates at any time point simplifies to:

$$\frac{h(t \mid \mathbf{Z})}{h(t \mid \mathbf{Z}^*)} = \frac{h_0(t) exp\left(\sum_{k=1}^{p} \beta_k \mathbf{Z}_k\right)}{h_0(t) exp\left(\sum_{k=1}^{p} \beta_k \mathbf{Z}_k^*\right)} = exp\left(\sum_{k=1}^{p} \beta_k \left(\mathbf{Z}_k - \mathbf{Z}_k{}^*\right)\right), \tag{5.7}$$

This ratio is constant or *proportional* throughout the study, that is, Equation (5.7) does not depend on t. This assumption greatly facilitates the interpretation of covariate effects, as the effect of a given covariate compared to the absence of that covariate is expressed as a single constant. This does not however imply that the absolute difference between the two individuals discussed above is constant; the exponentiated covariates act multiplicatively on a baseline hazard which may vary freely over time.

5.3 Analysis Results

Again two datasets are used for the analysis, the non-missing dataset with $n = 910$ and the $n = 1575$ dataset with unknowns in molecular markers only. The AREG method is employed to impute the missing values and generate the full dataset if needed, with number of imputation sets to be $m = 6$. Formally we should fit the survival model based on each of the 6 imputations (hence 6 full datasets), and average the 6 fitting results to get the final one. We call this procedure a formal approach. Alternatively we could also average the imputation results first and generate one full dataset. Then only one survival model fitting is needed. We call this an alternative approach.

Three outcomes, breast cancer specific survival (BCSS), local relapse survival (LRS), and distant relapse survival (DRS), are investigated in the survival analysis. The number of events with each outcome is given in Table 5.1

Table 5.1: Number of events with the outcome of BCSS, LRS, or DRS.

Outcome	n=1575	n=910
BCSS	476	288
LRS	138	83
DRS	518	320

The numbers of events for all categories of each variable are summarized in Appendix C.

5.3.1 Univariable analysis

Only one independent variable (marker or clinical index) is used in an univariable analysis model. The resulting p-values for three outcomes, breast cancer specific survival (BCSS), local relapse survival (LRS), and distant relapse survival (DRS) are given in Tables 5.2 and 5.3 with the $n = 1575$ and $n = 910$ datasets respectively. These analyses revealed that all clinical variables and most of the molecular markers are of significant importance (using significance level of p-value < 0.05) in predicting survival. Note that results in Table 5.2 are based on the alternative approach for the missing value imputations.

- **BCSS**: IGFB is the only insignificant covariate for breast cancer specific survival when using the $n = 1575$ dataset. With the $n = 910$ dataset, GATA3 and Her1 join IGFB to be insignificant. Indeed, mixed results have been reported for IGFBP-3 (Schairer 2004), with studies showing increased risk (Yu *et al.* 2002), reduced risk (Krajcik *et al.* 2002), or no change in risk (Keinan-Boker 2003; Toniolo 2000) for breast cancer in post-menopausal women. Some research work has shown that Her1 may have a prognostic role in locally advanced breast cancer (Colleoni 2008), but its related research is more uncommon than Her2. It is claimed that GATA3 can be used as a clinical marker to determine response to hormonal therapy and to refine the prognosis of breast cancer patients (Fang *et al.* 2009). We are not clear why GATA3 and Her1 become insignificant to BCSS

with the small dataset, but caution may be needed when using small datasets since some useful information (such as GATA3 and Her1) could be lost.

- **LRS**: ERPOSNE, YB1, IGFB, PR, GRADE and BCL2 are significantly associated with time to local relapse survival with the large dataset, while only ERPOSNE, GRADE and YB1 are significant using the small dataset. Again different results are obtained using these two datasets, with more significant variables for the large dataset. It is amazing that most clinical variables have no significant effects on LRS. On the other hand, some molecular markers seem to be important for LRS.

- **DRS**: With the $n = 1575$ dataset, only IGFB is insignificant (p-value = 0.26) for distant relapse survival, while both IGFB and Her1 are insignificant with the small dataset. Note that the evidence for AGE on the influence of DRS is not very strong compared with other clinical variables, its p-values being 0.021 and 0.03 with $n = 1575$ and $n = 910$ datasets respectively.

- SurvReg, a parametric survival model built in R, was also used for the univariable analysis. Results are very similar with those using the Cox model. The log-logistic distribution was chosen when using SurvReg model.

Table 5.4 give the results using the formal approach for the dataset of $n = 1575$. The numbers in parentheses are the standard errors of the mean p-values based on the 6 imputations. Although there are some small differences for the p-values between

Table 5.2 and Table 5.4, there are no major differences between these values, and most importantly, all the conclusions using the alternative approach are held with the formal approach.

5.3.2 Multivariable analysis

BCSS as the outcome

Fitting of the multivariable Cox proportional hazards model was conducted using all variables (full model) except HISTOLOGY in the model. Insignificant variables were removed from the full model by employing a backward elimination using the function *fastbw* in the package *Design* of R (R Documentation 2009) to produce the final reduced model in which only significant variables were presented.

fastbw is a fast backward variable elimination function. It performs a slightly inefficient but numerically stable version of fast backward elimination on factors, using a method based on Lawless and Singhal (1978). This method uses the fitted complete model and computes approximate Wald statistics by computing conditional (restricted) maximum likelihood estimates assuming multivariate normality of estimates. The cut-off p-value for elimination was p-value$<$ 0.1. The type of statistic on which to base the stopping rule was chosen to be *individual* that uses Wald chi-square statistics of individual factors.

The results based on the alternative approach are shown in Table 5.5. The coeffi-

cients for the remaining variables in the reduced model are very close to the corresponding ones in the full model. In the reduced model, note that three clinical variables, AGECAT, SYS and ERPONSE, are removed. They are likely correlated with other variables and/or their combinations. Four markers remain in the reduced model and all of them except CK56 have been used as prognostic factors to more accurately predict clinical outcome and guide therapies. CK56 was shown to be of prognostic importance in predicting patient outcomes and deciding therapeutic strategy for triple-negative (ER, PR, and Her2 negative) breast cancer (Rakha *et al.* 2007; Sasa *et al.* 2008). The molecular marker PR, which has been shown an important marker in the literature (Esteva1 and Hortobagyi 2004; Voorzanger-Rousselot and Garnero 2007; Ponzone *et al.* 2006), is not included in the model. One of the possible reasons is due to its correlation with other variables (such as ERS).

It is necessary to mention that the final reduced model may not be unique. Other combinations of the factors may be possible. This work, however, is beyond our current research scope.

The full model results using the formal approach are given in Table 5.6. In this table, *coef* and *p(form)* are the average values from $m = 6$ multiple imputations. *SE(between)* and *SE(within)* denote standard errors of the average coefficient between and within imputations respectively. *coef change* denotes the coefficient change compared with the alternative approach, given by *(coef using alternative approach - coef using formal approach) / coef using formal approach.*

From Table 5.6, it is clear that variation within imputations is the dominant source of variation. The difference of coefficients using different approaches is quite large for some factors but small for others, similar to the results in Chapter 4.3 on coefficient biases. However, large coefficient differences are always associated with the large p-value for the factor which leads to instability. The coefficient change is not greater than 15% if p-value$<$ 0.1.

Figure 5.1 gives comparison of the p-values of each variables for the full survival model using both approaches. *p(form)* and *p(alte)* denote the p-value from formal and alternative approach respectively in the figure.

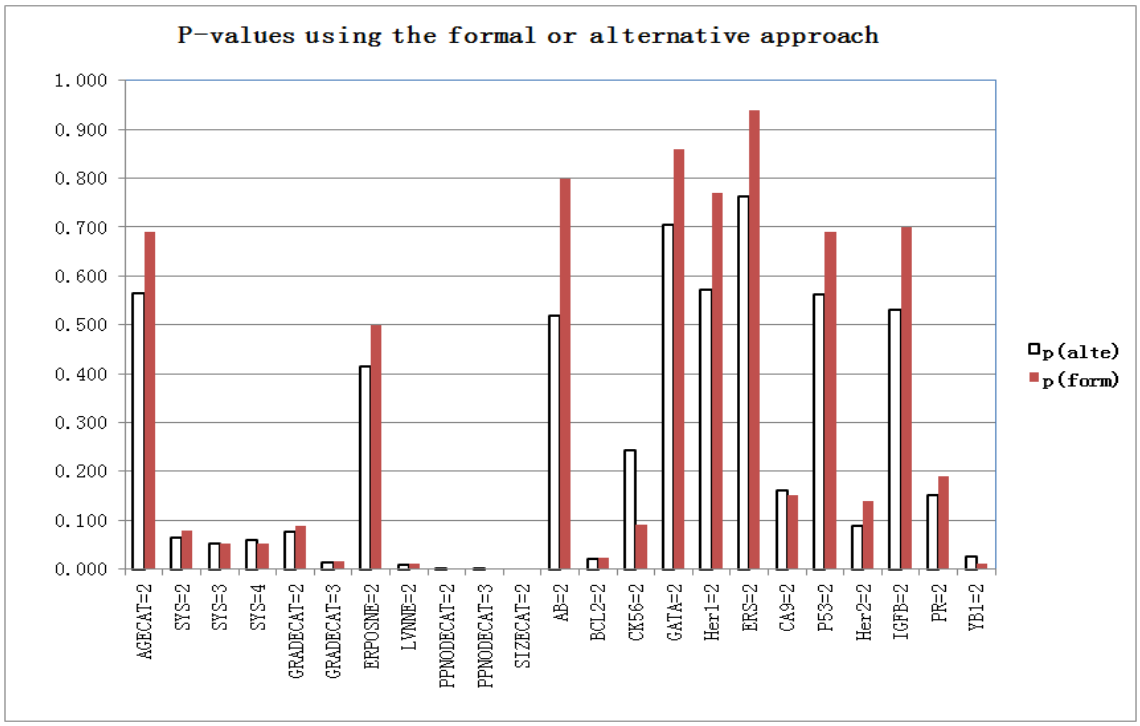


Figure 5.1: Comparison of p-values using formal and alternative approaches.

All the pairs of p-values are close to each other. Only those variables with p-values close to 0.1 may cause differences in the final reduced model, such as Her2 and CK56, because these variables may be included or excluded in the reduced model depending on the approach used. However, it would not influence the survival model fitting much because of their relatively large p-values.

The reason to use the alternative approach in our work is that different imputations may result in slightly different reduced model which brings additional complications for the data summary and analysis, and more importantly, the alternative approach can produce results as good as those using formal approach, based on our current requirements.

Table 5.7 gives the fitting results using the dataset without missing values ($n = 910$). Again, the full model and reduced model using the $n = 910$ dataset do not have much difference for the estimated coefficients. The reduced model with $n = 910$ is almost the same as that with $n = 1575$. LVNNE is replaced by ERPOSNE. The only difference for markers is that the variable AB is included in the reduced model with the $n = 910$ dataset. Note that the p-value for AB in the reduced model of Table 5.7 is quite large (p-value$>$ 0.05). Therefore it is not surprising that AB is included or excluded in the reduced model with the sample size change. This result certainly demonstrates the reliability of using the alternative approach.

LRS as the outcome

The model fitting results using the alternative approach for the large dataset with $n = 1575$ are shown in Table 5.8, while results using the formal approach are given in Table 5.9 for a comparison purpose. Only two clinical variables, GRADECAT and PPNODECAT, and two markers, IGFB and YB1, remain in the reduced model. The coefficients for the remaining variables in the reduced model are again close to the corresponding ones in the full model. Note that YB1 is the only marker to be significant for both BCSS and LRS. The marker IGFB is insignificant for BCSS, but significant for LRS.

Table 5.10 gives the results using the $n = 910$ dataset. The reduced model is the same as that using the $n = 1575$ dataset in Table 5.8. The coefficient for the covariate GRADECAT=2, however, changes from positive to negative. One of the reasons may be due to the large individual p-value for the covariate GRADECAT=2.

Note that the individual p-values for both GRADECAT and PPNODECAT are larger than 0.1, but they remain in the reduced model. The function *fastbw* employs the function *anova.Design* to automatically test combined significance for the categorical variables which have more than 2 degrees of freedom based on Wald statistics. A test result is given in Table 5.11, which shows that both GRADECAT and PPNODECAT are indeed significant.

The comparison of the p-values for the full survival model of LRS using formal or alternative approach is shown in Figure 5.2. Again there is little difference for the pair

of p-values using different approaches.

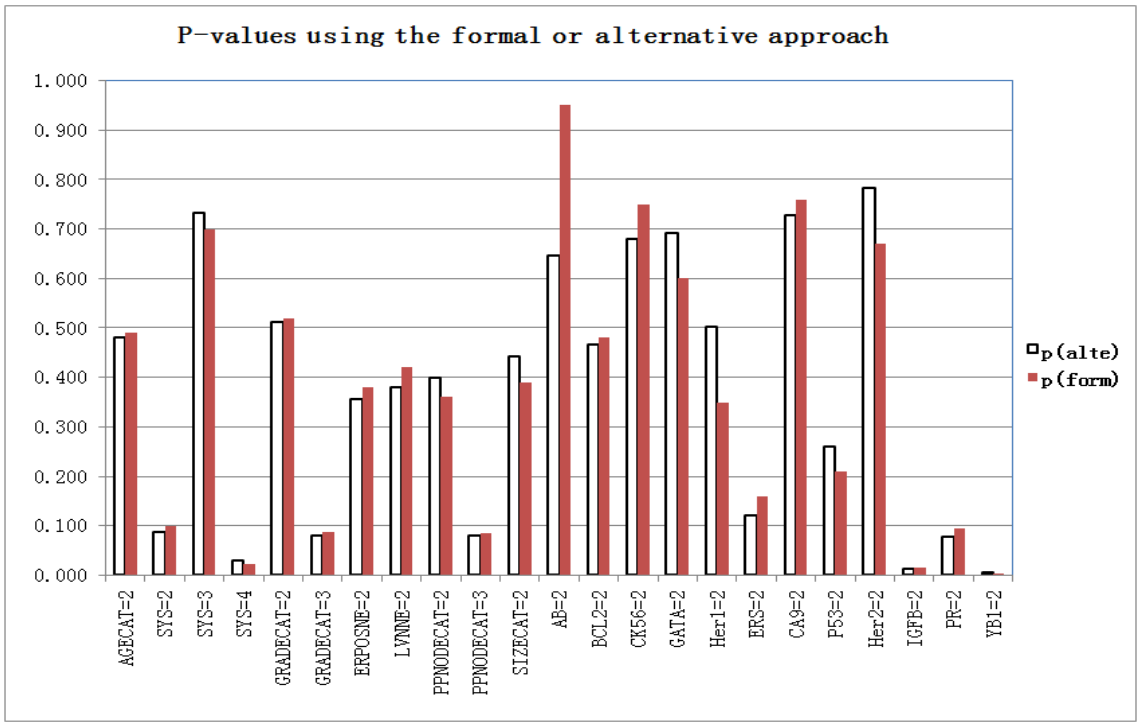


Figure 5.2: Comparison of p-values using formal and alternative approaches (outcome = LRS).

Table 5.2: Individual p-values from univariable analysis ($n = 1575$).

Variable	BCSS	LRS	DRS
AB	0.019	0.17	0.048
BCL2	<0.001	0.017	<0.001
CK56	0.001	0.60	0.012
Her1	<0.001	0.15	0.002
ERS	<0.001	0.12	<0.001
CA9	<0.001	0.22	<0.001
P53	<0.001	0.31	<0.001
Her2	<0.001	0.20	<0.001
IGFB	0.29	0.026	0.26
PR	<0.001	0.005	<0.001
YB1	<0.001	<0.001	<0.001
GATA3	0.045	0.21	0.045
AGE	0.035	0.55	0.021
SYS2	<0.001	0.73	<0.001
SYS3	<0.001	0.055	<0.001
SYS4	<0.001	0.18	<0.001
SYS (overall)	<0.001	0.061	<0.001
GRADE2	0.003	0.37	<0.001
GRADE3	<0.001	0.02	<0.001
GRADE(overall)	<0.001	<0.001	<0.001
ERPOSNE	<0.001	0.002	<0.001
LVNNE	<0.001	0.093	<0.001
PPNODE2	<0.001	0.63	<0.001
PPNODE3	<0.001	0.10	<0.001
PPNODE(overall)	<0.001	0.070	<0.001
SIZECAT	<0.001	0.95	<0.001

Table 5.3: Individual p-values from univariable analysis ($n = 910$).

Variable	BCSS	LRS	DRS
AB	0.007	0.068	0.038
BCL2	<0.001	0.062	<0.001
CK56	0.009	0.94	0.046
Her1	0.054	0.52	0.091
ERS	0.005	0.26	0.003
CA9	<0.001	0.089	<0.001
P53	<0.001	0.15	<0.001
Her2	<0.001	0.047	<0.001
IGFB	0.33	0.14	0.45
PR	<0.001	0.081	<0.001
YB1	<0.001	<0.001	<0.001
GATA3	0.23	0.13	0.025
AGE	0.014	0.84	0.03
SYS2	<0.001	0.92	<0.001
SYS3	<0.001	0.13	<0.001
SYS4	<0.001	0.99	<0.001
SYS (overall)	<0.001	0.473	<0.001
GRADE2	0.044	0.73	0.01
GRADE3	<0.001	0.14	<0.001
GRADE(overall)	<0.001	<0.001	<0.001
ERPOSNE	0.018	0.042	0.005
LVNNE	<0.001	0.059	<0.001
PPNODE2	<0.001	0.59	<0.001
PPNODE3	<0.001	0.14	<0.001
PPNODE(overall)	<0.001	0.102	<0.001
SIZE	<0.001	0.62	<0.001

Table 5.4: Individual p-values (mean and SE) from univariable analysis using formal approach ($n = 1575$).

Variable	BCSS	LRS	DRS
AB	0.032 (0.010)	0.263 (0.064)	0.088 (0.032)
BCL2	<0.001 (0)	0.023 (0.006)	<0.001 (0)
CK56	0.012 (0.004)	0.566 (0.115)	0.050 (0.017)
Her1	<0.001 (0)	0.159 (0.043)	0.001 (0.000)
ERS	<0.001 (0)	0.145 (0.009)	<0.001 (0)
CA9	<0.001 (0)	0.197 (0.045)	<0.001 (0)
P53	<0.001 (0)	0.305 (0.025)	<0.001 (0)
Her2	<0.001 (0)	0.213 (0.025)	<0.001 (0)
IGFB	0.255 (0.040)	0.026 (0.004)	0.222 (0.036)
PR	<0.001 (0)	0.006 (0.001)	<0.001 (0)
YB1	<0.001 (0)	0.001 (0.000)	<0.001 (0)
GATA3	0.049 (0.012)	0.236 (0.036)	0.041 (0.020)

Table 5.5: Cox model with $n = 1575$ and outcome=BCSS

Full model	coef	exp(coef)	se(coef)	z	p
AGECAT=2	-0.05	0.946	0.137	-0.40	0.690
SYS=2	-0.27	0.759	0.156	-1.76	0.078
SYS=3	-0.36	0.694	0.187	-1.94	0.052
SYS=4	-0.42	0.654	0.219	-1.93	0.053
GRADECAT=2	0.586	1.797	0.344	1.700	0.089
GRADECAT=3	0.842	2.322	0.345	2.438	0.015
ERPOSNE=2	0.114	1.121	0.171	0.667	0.500
LVNNE=2	0.284	1.329	0.111	2.555	0.011
PPNODECAT=2	0.625	1.87	0.123	5.053	0.000
PPNODECAT=3	-0.60	0.548	0.151	-3.96	0.000
SIZECAT=2	0.536	1.71	0.098	5.437	0.000
AB=2	0.044	1.045	0.172	0.257	0.800
BCL2=2	-0.26	0.768	0.115	-2.29	0.022
CK56=2	0.311	1.366	0.183	1.697	0.090
GATA=2	0.020	1.021	0.119	0.174	0.861
Her1=2	-0.05	0.950	0.179	-0.280	0.770
ERS=2	-0.01	0.988	0.152	-0.07	0.940
CA9=2	0.189	1.209	0.131	1.439	0.150
P53=2	0.049	1.051	0.123	0.400	0.691
Her2=2	0.202	1.224	0.136	1.475	0.140
IGFB=2	0.038	1.039	0.098	0.390	0.700
PR=2	-0.140	0.862	0.113	-1.30	0.190
YB1=2	0.282	1.327	0.11	2.569	0.011
Reduced model	coef	exp(coef)	se(coef)	z	p
GRADECAT=2	0.584	1.793	0.343	1.7	0.089
GRADECAT=3	0.868	2.382	0.343	2.53	0.011
LVNNE=2	0.218	1.243	0.106	2.05	0.040
PPNODECAT=2	0.619	1.857	0.122	5.06	0.000
PPNODECAT=3	-0.420	0.656	0.125	-3.34	0.000
SIZECAT=2	0.521	1.683	0.096	5.390	0.000
BCL2=2	-0.28	0.751	0.098	-2.90	0.003
CK56=2	0.388	1.474	0.162	2.390	0.016
Her2=2	0.254	1.289	0.127	1.991	0.046
YB1=2	0.283	1.327	0.100	2.810	0.005

Table 5.6: Cox model using formal approach ($n = 1575$ and outcome=BCSS)

Full model	coef	SE(between)	SE(within)	coef change	p(form)
AGECAT=2	-0.05	0.005	0.137	-0.07	0.564
SYS=2	-0.27	0.006	0.156	-0.01	0.063
SYS=3	-0.35	0.005	0.187	0.037	0.052
SYS=4	-0.39	0.006	0.218	0.066	0.058
GRADECAT=2	0.589	0.003	0.344	-0.00	0.075
GRADECAT=3	0.847	0.003	0.345	-0.00	0.012
ERPOSNE=2	0.117	0.013	0.170	-0.02	0.415
LVNNE=2	0.286	0.002	0.111	-0.00	0.008
PPNODECAT=2	0.623	0.003	0.123	0.004	0.000
PPNODECAT=3	-0.59	0.004	0.151	0.003	0.000
SIZECAT=2	0.549	0.002	0.098	-0.02	0.000
AB=2	0.072	0.043	0.164	-0.63	0.517
BCL2=2	-0.25	0.012	0.113	0.022	0.019
CK56=2	0.210	0.029	0.182	0.323	0.243
GATA=2	0.017	0.012	0.118	0.138	0.704
Her1=2	0.020	0.035	0.175	1.396	0.571
ERS=2	0.004	0.008	0.150	1.348	0.763
CA9=2	0.176	0.008	0.129	0.067	0.160
P53=2	0.060	0.010	0.122	-0.21	0.562
Her2=2	0.232	0.016	0.134	-0.15	0.088
IGFB=2	0.045	0.010	0.097	-0.19	0.531
PR=2	-0.16	0.018	0.112	-0.12	0.151
YB1=2	0.264	0.027	0.106	0.065	0.025

Table 5.7: Cox model with $n = 910$ and outcome=BCSS

Full model	coef	exp(coef)	se(coef)	z	p
AGECAT=2	-0.08	0.915	0.179	-0.49	0.62
SYS=2	-0.41	0.66	0.203	-2.04	0.041
SYS=3	-0.32	0.724	0.241	-1.33	0.18
SYS=4	-0.18	0.831	0.274	-0.67	0.50
GRADECAT=2	0.353	1.423	0.464	0.761	0.45
GRADECAT=3	0.682	1.978	0.463	1.473	0.14
ERPOSNE=2	0.345	1.413	0.233	1.487	0.14
LVNNE=2	0.204	1.227	0.15	1.367	0.17
PPNODECAT=2	0.459	1.583	0.158	2.912	0.003
PPNODECAT=3	-0.72	0.484	0.196	-3.70	0.000
SIZECAT=2	0.683	1.98	0.129	5.281	0.000
AB=2	0.350	1.419	0.207	1.69	0.091
BCL2=2	-0.35	0.704	0.151	-2.32	0.02
CK56=2	0.397	1.488	0.217	1.833	0.067
GATA=2	0.150	1.162	0.149	1.008	0.31
Her1=2	-0.23	0.788	0.223	-1.06	0.29
ERS=2	0.098	1.103	0.238	0.413	0.68
CA9=2	0.168	1.184	0.161	1.05	0.29
P53=2	0.197	1.218	0.15	1.313	0.19
Her2=2	0.309	1.363	0.171	1.805	0.071
IGFB=2	0.022	1.023	0.127	0.18	0.86
PR=2	-0.17	0.842	0.151	-1.13	0.25
YB1=2	0.336	1.399	0.137	2.45	0.014
Reduced model	coef	exp(coef)	se(coef)	z	p
GRADECAT=2	0.389	1.475	0.463	0.84	0.401
GRADECAT=3	0.732	2.079	0.462	1.59	0.113
ERPOSNE=2	0.365	1.440	0.181	2.01	0.044
PPNODECAT=2	0.492	1.635	0.155	3.18	0.001
PPNODECAT=3	-0.60	0.548	0.15	-4	0.000
SIZECAT=2	0.664	1.942	0.126	5.27	0.000
AB=2	0.381	1.463	0.197	1.93	0.053
BCL2=2	-0.34	0.708	0.141	-2.44	0.014
CK56=2	0.389	1.475	0.206	1.89	0.058
Her2=2	0.359	1.431	0.163	2.2	0.027
YB1=2	0.404	1.497	0.131	3.08	0.002

Table 5.8: Cox model with $n = 1575$ and outcome=LRS

Full model	coef	exp(coef)	se(coef)	z	p
AGECAT=2	0.179	1.197	0.257	0.698	0.49
SYS=2	-0.45	0.635	0.278	-1.63	0.10
SYS=3	-0.12	0.881	0.334	-0.38	0.70
SYS=4	-1.27	0.28	0.561	-2.27	0.023
GRADECAT=2	0.336	1.40	0.527	0.639	0.52
GRADECAT=3	0.900	2.46	0.526	1.712	0.087
ERPOSNE=2	-0.28	0.754	0.324	-0.87	0.38
LVNNE=2	0.164	1.178	0.205	0.801	0.42
PPNODECAT=2	0.237	1.267	0.257	0.920	0.36
PPNODECAT=3	-0.46	0.63	0.267	-1.72	0.084
SIZECAT=2	-0.15	0.854	0.182	-0.86	0.39
AB=2	-0.02	0.979	0.316	-0.06	0.95
BCL2=2	-0.15	0.859	0.216	-0.70	0.48
CK56=2	-0.11	0.89	0.372	-0.31	0.75
GATA=2	-0.11	0.892	0.22	-0.51	0.60
Her1=2	-0.30	0.735	0.331	-0.93	0.35
ERS=2	0.440	1.553	0.31	1.418	0.16
CA9=2	0.079	1.083	0.261	0.305	0.76
P53=2	-0.29	0.742	0.24	-1.24	0.21
Her2=2	-0.12	0.887	0.278	-0.43	0.67
IGFB=2	-0.48	0.616	0.199	-2.43	0.015
PR=2	-0.35	0.703	0.211	-1.67	0.094
YB1=2	0.641	1.899	0.201	3.197	0.001
Reduced model	coef	exp(coef)	se(coef)	z	p
GRADECAT=2	0.322	1.379	0.525	0.613	0.539
GRADECAT=3	0.921	2.511	0.517	1.78	0.075
PPNODECAT=2	0.162	1.175	0.253	0.641	0.521
PPNODECAT=3	-0.30	0.737	0.206	-1.47	0.139
IGFB=2	-0.47	0.623	0.193	-2.44	0.014
YB1=2	0.527	1.693	0.176	2.996	0.002

Table 5.9: Cox model using formal approach ($n = 1575$ and outcome=LRS)

Full model	coef	SE(between)	SE(within)	coef change	p(form)
AGECAT=2	0.181	0.007	0.249	-0.01	0.480
SYS=2	-0.47	0.006	0.279	-0.04	0.088
SYS=3	-0.11	0.009	0.349	0.107	0.733
SYS=4	-1.20	0.004	0.556	0.050	0.030
GRADECAT=2	0.344	0.003	0.542	-0.02	0.512
GRADECAT=3	0.919	0.003	0.541	-0.02	0.080
ERPOSNE=2	-0.30	0.024	0.344	-0.07	0.355
LVNNE=2	0.180	0.007	0.224	-0.09	0.379
PPNODECAT=2	0.216	0.004	0.260	0.086	0.398
PPNODECAT=3	-0.46	0.003	0.259	-0.00	0.081
SIZECAT=2	-0.13	0.004	0.178	0.114	0.443
AB=2	-0.02	0.068	0.299	-0.08	0.645
BCL2=2	-0.15	0.023	0.215	-0.04	0.465
CK56=2	-0.07	0.077	0.370	0.399	0.678
GATA=2	-0.08	0.025	0.222	0.213	0.691
Her1=2	-0.24	0.073	0.346	0.211	0.502
ERS=2	0.482	0.017	0.319	-0.09	0.121
CA9=2	0.077	0.028	0.253	0.028	0.727
P53=2	-0.27	0.018	0.255	0.086	0.259
Her2=2	-0.07	0.023	0.267	0.357	0.782
IGFB=2	-0.49	0.015	0.172	-0.01	0.013
PR=2	-0.37	0.014	0.230	-0.05	0.078
YB1=2	0.549	0.014	0.182	0.143	0.005

Table 5.10: Cox model with $n = 910$ and outcome=LRS

Full model	coef	exp(coef)	se(coef)	z	p
AGECAT=2	0.466	1.594	0.349	1.334	0.18
SYS=2	-0.36	0.694	0.354	-1.03	0.30
SYS=3	0.013	1.013	0.443	0.029	0.98
SYS=4	-0.57	0.562	0.603	-0.95	0.34
GRADECAT=2	-0.48	0.613	0.631	-0.77	0.44
GRADECAT=3	0.442	1.557	0.614	0.721	0.47
ERPOSNE=2	-0.40	0.665	0.473	-0.86	0.39
LVNNE=2	0.192	1.213	0.274	0.702	0.48
PPNODECAT=2	0.181	1.199	0.323	0.561	0.57
PPNODECAT=3	-0.42	0.654	0.339	-1.25	0.21
SIZECAT=2	0.023	1.023	0.233	0.099	0.92
AB=2	0.498	1.647	0.39	1.279	0.20
BCL2=2	-0.09	0.907	0.287	-0.33	0.73
CK56=2	-0.29	0.745	0.476	-0.61	0.54
GATA=2	-0.16	0.848	0.281	-0.58	0.56
Her1=2	-0.44	0.643	0.428	-1.03	0.30
ERS=2	0.912	2.491	0.523	1.745	0.081
CA9=2	0.339	1.404	0.304	1.116	0.26
P53=2	-0.22	0.796	0.29	-0.78	0.43
Her2=2	0.162	1.176	0.326	0.497	0.62
IGFB=2	-0.52	0.593	0.252	-2.07	0.04
PR=2	-0.18	0.827	0.276	-0.68	0.49
YB1=2	0.899	2.457	0.256	3.505	0.00
Reduced model	coef	exp(coef)	se(coef)	z	p
GRADECAT=2	-0.44	0.638	0.628	-0.71	0.476
GRADECAT=3	0.509	1.663	0.6	0.848	0.396
PPNODECAT=2	0.205	1.227	0.312	0.657	0.511
PPNODECAT=3	-0.41	0.657	0.262	-1.60	0.108
IGFB=2	-0.42	0.652	0.243	-1.75	0.079
YB1=2	0.757	2.131	0.229	3.305	0.000

Table 5.11: Wald test on significance for the reduced model in Table 5.8 ($n = 910$ and outcome=LRS)

Factor	Chi-Square	d.f.	P
GRADECAT	12.59	2	0.0018
PPNODECAT	5.68	2	0.0583
IGFB	3.07	1	0.0797
YB1	10.92	1	0.001

DRS as the outcome

The fitting results using the alternative approach for the large dataset with $n = 1575$ are shown in Table 5.12, and results for the small dataset with $n = 910$ are given in Table 5.13. For the clinical variables, AGECAT and LVNNE are excluded from the reduced models, and SYS is shown to be significant using the dataset of $n = 1575$ but insignificant using the dataset of $n = 910$. Four markers remain in the final reduced models, with (BCL2, Her2, YB1, CK56) using the large dataset and (BCL2, Her2, YB1, P53) using the small one. The differences of corresponding coefficients in the two reduced models are small.

10 year survival probability and survival curve

Table 5.14 gives the summary of 10 year survival probabilities at the mean of the covariates under different conditions. The 10 year survival probability for BCSS, LRS, and DRS is 77%, 91%, and 72% respectively. There is no difference between full models and reduced models. The difference between two datasets($n = 1575$ and $n = 910$) is very small, less than 2% for DRS and less than 1% for BCSS or LRS.

Two representive estimated survival curves at the mean of covariates, which gives the plot of survival proportion verus survival year, are given in Figures 5.3 and 5.4 for BCSS and LRS respectively. The broken lines in the figures represent a 2-standard-error band around the fit. There is a large probability of surviving beyond 1 year. The survival proportion drops quickly between one year and 5 years. And then the

decrease of survival proportion gets mild after 5 years, and there is an approximately linear relationship between the survival proportion and year.

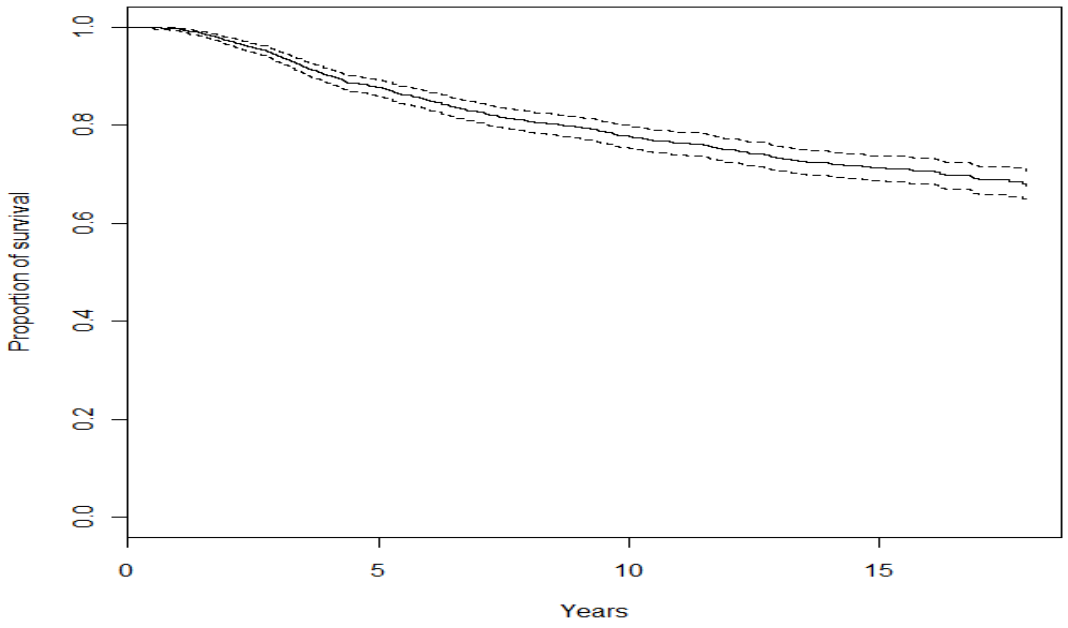


Figure 5.3: Estimated survival curve ($n = 1575$, outcome=BCSS).

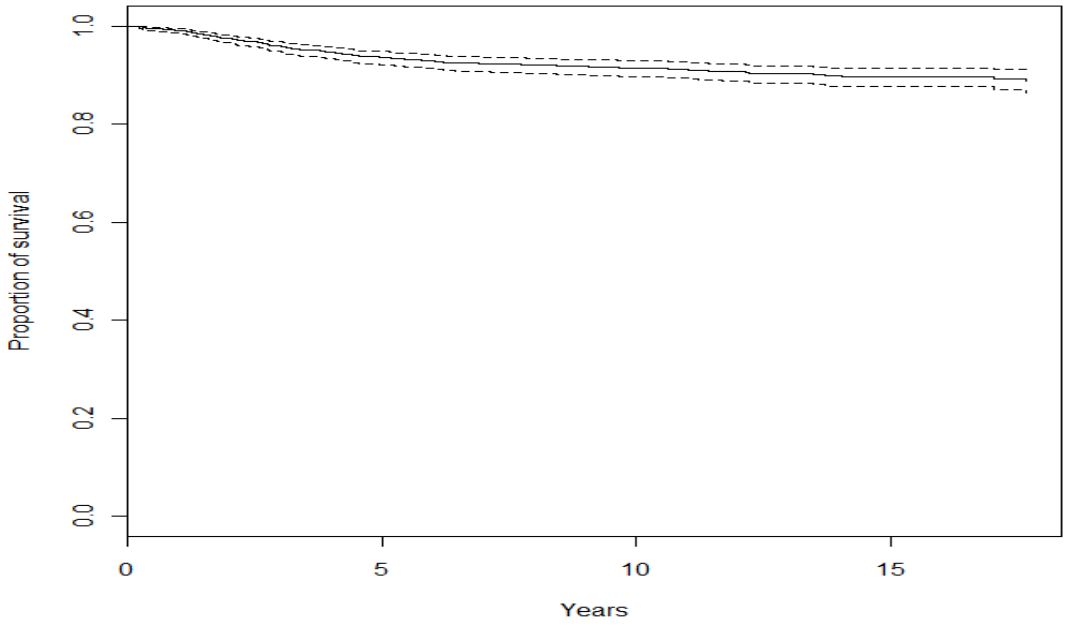


Figure 5.4: Estimated survival curve ($n = 1575$, outcome=LRS).

Table 5.12: Cox model with $n = 1575$ and outcome=DRS

Full model	coef	exp(coef)	se(coef)	z	p
AGECAT=2	-0.07	0.928	0.133	-0.56	0.58
SYS=2	-0.32	0.722	0.149	-2.18	0.029
SYS=3	-0.37	0.689	0.180	-2.06	0.039
SYS=4	-0.47	0.619	0.209	-2.28	0.022
GRADECAT=2	0.827	2.288	0.363	2.278	0.023
GRADECAT=3	1.030	2.802	0.364	2.827	0.004
ERPOSNE=2	0.041	1.043	0.164	0.254	0.80
LVNNE=2	0.309	1.362	0.106	2.90	0.003
PPNODECAT=2	0.531	1.701	0.117	4.535	0.000
PPNODECAT=3	-0.71	0.492	0.143	-4.93	0.000
SIZECAT=2	0.489	1.632	0.094	5.192	0.000
AB=2	0.027	1.028	0.167	0.165	0.87
BCL2=2	-0.26	0.77	0.110	-2.37	0.017
CK56=2	0.236	1.267	0.181	1.304	0.19
GATA=2	0.005	1.005	0.113	0.045	0.96
Her1=2	-0.10	0.902	0.173	-0.59	0.55
ERS=2	0.064	1.067	0.150	0.429	0.67
CA9=2	0.228	1.257	0.126	1.810	0.07
P53=2	0.093	1.098	0.117	0.795	0.43
Her2=2	0.211	1.235	0.131	1.612	0.11
IGFB=2	0.028	1.029	0.094	0.306	0.76
PR=2	-0.15	0.856	0.108	-1.42	0.15
YB1=2	0.235	1.266	0.105	2.235	0.025
Reduced model	coef	exp(coef)	se(coef)	z	p
SYS=2	-0.29	0.746	0.142	-2.06	0.039
SYS=3	-0.31	0.728	0.149	-2.12	0.034
SYS=4	-0.45	0.635	0.203	-2.23	0.025
GRADECAT=2	0.834	2.302	0.362	2.30	0.022
GRADECAT=3	1.067	2.907	0.363	2.94	0.003
LVNNE2	0.295	1.343	0.105	2.79	0.005
PPNODECAT=2	0.539	1.714	0.116	4.63	0.000
PPNODECAT=3	-0.67	0.508	0.141	-4.8	0.000
SIZECAT=2	0.495	1.641	0.093	5.31	0.000
BCL2=2	-0.26	0.771	0.095	-2.72	0.006
CA9=2	0.255	1.29	0.117	2.17	0.03
Her2=2	0.222	1.248	0.123	1.79	0.073
YB1=2	0.276	1.318	0.099	2.78	0.005

Table 5.13: Cox model with $n = 910$ and outcome=DRS

Full model	coef	exp(coef)	se(coef)	z	p
AGECAT=2	-0.08	0.923	0.172	-0.46	0.64
SYS=2	-0.37	0.684	0.19	-1.99	0.046
SYS=3	-0.30	0.74	0.23	-1.30	0.19
SYS=4	-0.26	0.77	0.263	-0.99	0.32
GRADECAT=2	0.757	2.132	0.513	1.476	0.14
GRADECAT=3	1.029	2.799	0.513	2.007	0.045
ERPOSNE=2	0.166	1.181	0.224	0.746	0.46
LVNNE=2	0.236	1.267	0.141	1.675	0.094
PPNODECAT=2	0.406	1.501	0.148	2.751	0.005
PPNODECAT=3	-0.79	0.452	0.184	-4.32	0.000
SIZECAT=2	0.57	1.768	0.121	4.699	0.000
AB=2	0.236	1.267	0.205	1.156	0.25
BCL2=2	-0.31	0.732	0.142	-2.19	0.028
CK56=2	0.299	1.349	0.213	1.409	0.16
GATA=2	0.164	1.179	0.141	1.167	0.24
Her1=2	-0.22	0.802	0.214	-1.03	0.3
ERS=2	0.156	1.17	0.237	0.661	0.51
CA9=2	0.170	1.186	0.152	1.121	0.26
P53=2	0.216	1.241	0.143	1.516	0.13
Her2=2	0.295	1.344	0.162	1.823	0.068
IGFB=2	-0.01	0.987	0.121	-0.10	0.91
PR=2	-0.19	0.823	0.142	-1.36	0.17
YB1=2	0.273	1.314	0.13	2.1	0.036
Reduced model	coef	exp(coef)	se(coef)	z	p
GRADECAT=2	0.803	2.231	0.512	1.57	0.12
GRADECAT=3	1.08	2.944	0.512	2.11	0.035
PPNODECAT=2	0.445	1.56	0.144	3.09	0.002
PPNODECAT=3	-0.68	0.506	0.14	-4.86	0.000
SIZECAT=2	0.564	1.758	0.118	4.8	0.000
BCL2=2	-0.24	0.781	0.126	-1.96	0.05
P53=2	0.227	1.255	0.137	1.66	0.098
Her2=2	0.265	1.303	0.149	1.78	0.075
YB1=2	0.31	1.363	0.124	2.5	0.012

Table 5.14: Summary of 10 year survival probability

dataset	model	BCSS (SE)	LRS (SE)	DRS (SE)
n=1575	full model	0.7774 (0.0116)	0.9174 (0.0008)	0.7233 (0.0129)
	reduced model	0.7769 (0.0116)	0.9148 (0.0008)	0.7227 (0.0129)
n=910	full model	0.7696 (0.0155)	0.9180 (0.0110)	0.7075 (0.0172)
	reduced model	0.7654 (0.0154)	0.9079 (0.0110)	0.7047 (0.0171)

5.3.3 Model diagnostics

It is desirable to determine whether a fitted Cox regression model adequately describes the data. Three kinds of diagnostics are generally carried out for a Cox model: for violation of the assumption of proportional hazards; for influential data; and for non-linearity in the relationship between the log hazard and the covariates. All of these diagnostics use the residuals method for *coxph* or *cph* objects, which calculates several kinds of residuals (along with some quantities that are not normally thought of as residuals). Details can be found in Tableman and Kim (2003). Since all variables are categorical, the nonlinearity check is not needed here.

Check of proportional hazards assumption

Tests and graphical diagnostics for proportional hazards may be based on the scaled Schoenfeld residuals. More conveniently, the *cox.zph* function calculates tests of the proportional-hazards assumption for each covariate, by correlating the corresponding set of scaled Schoenfeld residuals with a suitable transformation of time. The testing results for the reduced models using the $n = 1575$ dataset with the alternative approach are given in Tables 5.15 to 5.17 for the outcomes of BCSS, LRS, and DRS respectively.

With the model of outcome=BCSS, there is strong evidence of non-proportional hazards for BCL2, Her2, YB1, GRADECAT=3 and PPNODECAT=3. Correspondingly the global test is statistically significant. The testing results for the model of outcome=DRS are somewhat similar with those for outcome=BCSS. BCL2, Her2, and

Table 5.15: Tests of proportional-hazards assumption by cox.zph (n = 1575, outcome=BCSS).

outcome=BCSS	rho	Chi-square	p
GRADECAT=2	-0.0633	1.903	0.168
GRADECAT=3	-0.1007	4.839	0.0278
LVNNE=2	0.0717	2.464	0.116
PPNODECAT=2	-0.0138	0.091	0.763
PPNODECAT=3	0.1151	6.299	0.0121
SIZECAT=2	-0.0198	0.192	0.662
BCL2=2	0.2243	21.811	0.0000
CK56=2	-0.0643	1.973	0.1600
Her2=2	-0.1239	6.854	0.0088
YB1=2	-0.1195	6.816	0.0090
GLOBAL	NA	101.641	0.0000

LVNNE show strong evidence of PH violation, while YB1 and GRADECAT=3 show marginal evidence. For the model of outcome=LRS, YB1 is the only covariate that violates the PH assumption, while the global test (on 6 degrees of freedom) shows statistically marginal significant.

The PH testing results for the n = 910 dataset are given in Table 5.18. For the model of LRS and DRS, the results are similar with corresponding ones using the $n = 1575$ dataset. For the outcome of BCSS, however, ERPOSNE is the only covariate showing strong evidence on the PH violation, while BCL2 shows the marginal, which is quite different from those for the large dataset.

Figures 5.5 to 5.7 gives the plots of scaled Schoenfeld residuals against transformed time for each covariate in the reduced models of BCSS, LRS, and DRS respectively using the large dataset, while Figure 5.8 is the plot for BCSS model using the small

Table 5.16: Tests of proportional-hazards assumption by *cox.zph* (n = 1575, outcome=LRS).

outcome=LRS	rho	Chi-square	p
GRADECAT=2	0.10728	1.58	0.2084
GRADECAT=3	0.11952	1.94	0.1632
PPNODECAT=2	-0.09648	1.26	0.2620
PPNODECAT=3	0.01490	0.0303	0.8617
IGFB=2	0.00078	0.00008	0.9927
YB1=2	-0.29212	11.2	0.0008
GLOBAL	NA	13.6	0.0342

dataset. The solid line is a smoothing-spline fit to the plot, with the broken lines representing a 2-standard-error band around the fit. A Schoenfeld residual is the difference between the covariate value at a failure time and its expected value. The rescaled Schoenfeld residuals were used in the analysis due to multiple predictors.

Interpretation of these graphs is greatly facilitated by smoothing, for which purpose *cox.zph* uses a smoothing spline. Systematic departures from a horizontal line are indicative of non-proportional hazards. Results are in agreement with those reported above in Tables 5.15 to 5.18.

One way of accommodating non-proportional hazards is to divide the data into strata based on the value of one or more covariates with non-proportional hazards. Each stratum is permitted to have a different baseline hazard function, while the coefficients of the remaining covariates are assumed to be constant across strata. An advantage of this approach is that we do not have to assume a particular form of interaction between the stratifying covariates and time. A disadvantage is the resulting inability

Table 5.17: Tests of proportional-hazards assumption by *cox.zph* (n = 1575, outcome=DRS).

outcome=DRS	rho	Chi-square	p
SYS=2	0.0340	0.7364	0.391
SYS=3	-0.0165	0.156	0.693
SYS=4	0.01542	0.1386	0.710
GRADECAT=2	-0.04729	1.1666	0.280
GRADECAT=3	-0.08022	3.3883	0.0657
LVNNE=2	0.10487	5.8891	0.0152
PPNODECAT=2	-0.04269	0.9705	0.325
PPNODECAT=3	0.02887	0.4853	0.486
SIZECAT=2	-0.0045	0.0106	0.918
BCL2=2	0.11834	6.9206	0.0085
CA9=2	-0.00711	0.0283	0.866
Her2=2	-0.13322	8.8932	0.0028
YB1=2	-0.08481	4.0067	0.0453
GLOBAL	NA	69.6527	0.0000

to examine the effects of the stratifying covariates. Stratification is most natural when a covariate takes on only a few distinct values, and when the effect of the stratifying variable is not of direct interest.

We compared the six Cox PH models with and without adding the time-dependent covariates and the estimates for all parameters were consistent. Tables 5.19 and 5.20 give the estimated coefficients after the stratification of a covariate with the most non-proportional hazards, as well as the comparison of the coefficients with and without stratification for n = 1575 and n = 910 respectively. The term *coef diff* in these two tables are calculated by *(coef with stratification - coef)/coef*. Notice how close they are for the coefficients with and without stratification. We can conclude that the covariate

effects are robust to model specification.

Check of influential observations

The influence of each observation on the estimated coefficients is examined by calculating the difference, called $dfbeta$, between the estimated coefficients with all observations in the model and the coefficients computed on the sample with the examining observation deleted. Specifying the argument $type = dfbeta$ to residuals produces a matrix of estimated changes in the regression coefficients upon deleting each observation in turn; likewise, $type = dfbetas$ produces the estimated changes in the coefficients divided by their standard errors.

The index plots for BCSS and LRS using the dataset of $n = 1575$ are shown respectively in Figure 5.9 and 5.10. The plots for DRS and/or using the dataset of $n = 910$ are skipped here because they are similar to either Figure 5.9 or Figure 5.10. Results suggest that all the observations are not influential individually since the largest $dfbetas$ value is less than 0.3, which means the difference between estimated coefficients is much smaller than their standard errors.

Table 5.18: Tests of proportional-hazards assumption by *cox.zph* (n = 910, outcome=BCSS).

outcome=BCSS	rho	Chi-square	p
GRADECAT=2	-0.07235	1.5	0.221
GRADECAT=3	-0.10739	3.3	0.0691
ERPOSNE=2	0.22429	14.8	0.0001
PPNODECAT=2	-0.0608	1.1	0.294
PPNODECAT=3	0.06749	1.34	0.246
SIZECAT=2	-0.0847	2.07	0.151
AB=2	-0.00128	0.000439	0.983
BCL2=2	0.11867	3.85	0.0496
CK56=2	0.02562	0.179	0.673
Her2=2	-0.05598	0.876	0.349
YB1=2	-0.07426	1.6	0.206
GLOBAL	NA	80	0.0000
outcome=LRS	rho	Chi-square	p
GRADECAT=2	0.11013	0.99811	0.3177
GRADECAT=3	0.11224	1.01617	0.3134
PPNODECAT=2	-0.08493	0.57646	0.4477
PPNODECAT=3	0.00831	0.00568	0.9399
IGFB=2	0.07089	0.41307	0.5204
YB1=2	-0.32402	8.34287	0.0038
GLOBAL	NA	9.67621	0.1389
outcome=DRS	rho	Chi-square	p
GRADECAT=2	-0.032	0.329	0.566
GRADECAT=3	-0.0611	1.213	0.271
PPNODECAT=2	-0.052	0.889	0.346
PPNODECAT=3	-0.0334	0.365	0.546
SIZECAT=2	-0.0568	1.028	0.311
BCL2=2	0.1772	9.598	0.0019
P53=2	0.0257	0.241	0.624
Her2=2	-0.0832	2.136	0.144
YB1=2	-0.1388	6.665	0.0098
GLOBAL	NA	46.585	0.0000

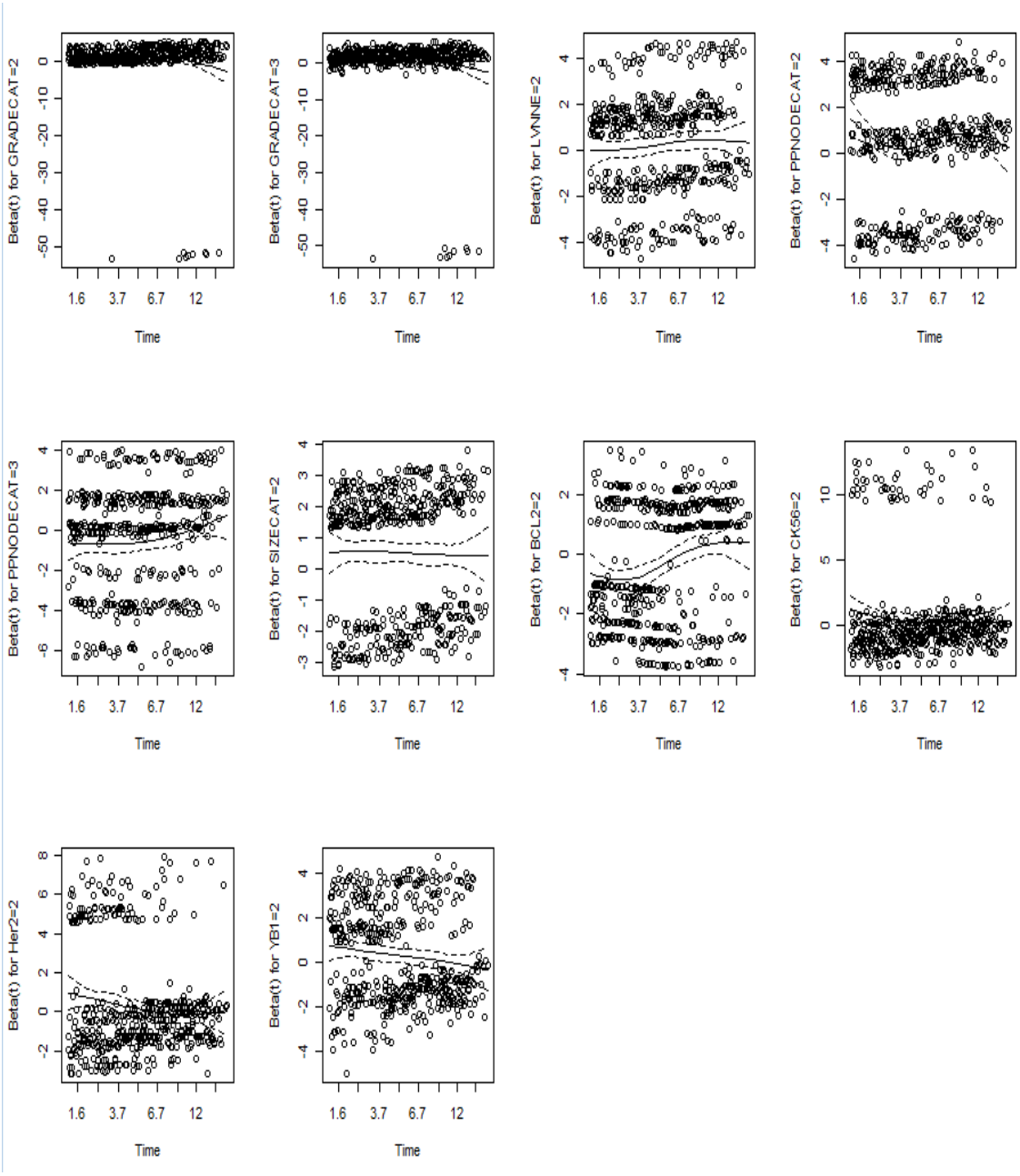


Figure 5.5: Plots of scaled Schoenfeld residuals against transformed time for each covariate ($n = 1575$, outcome=BCSS).

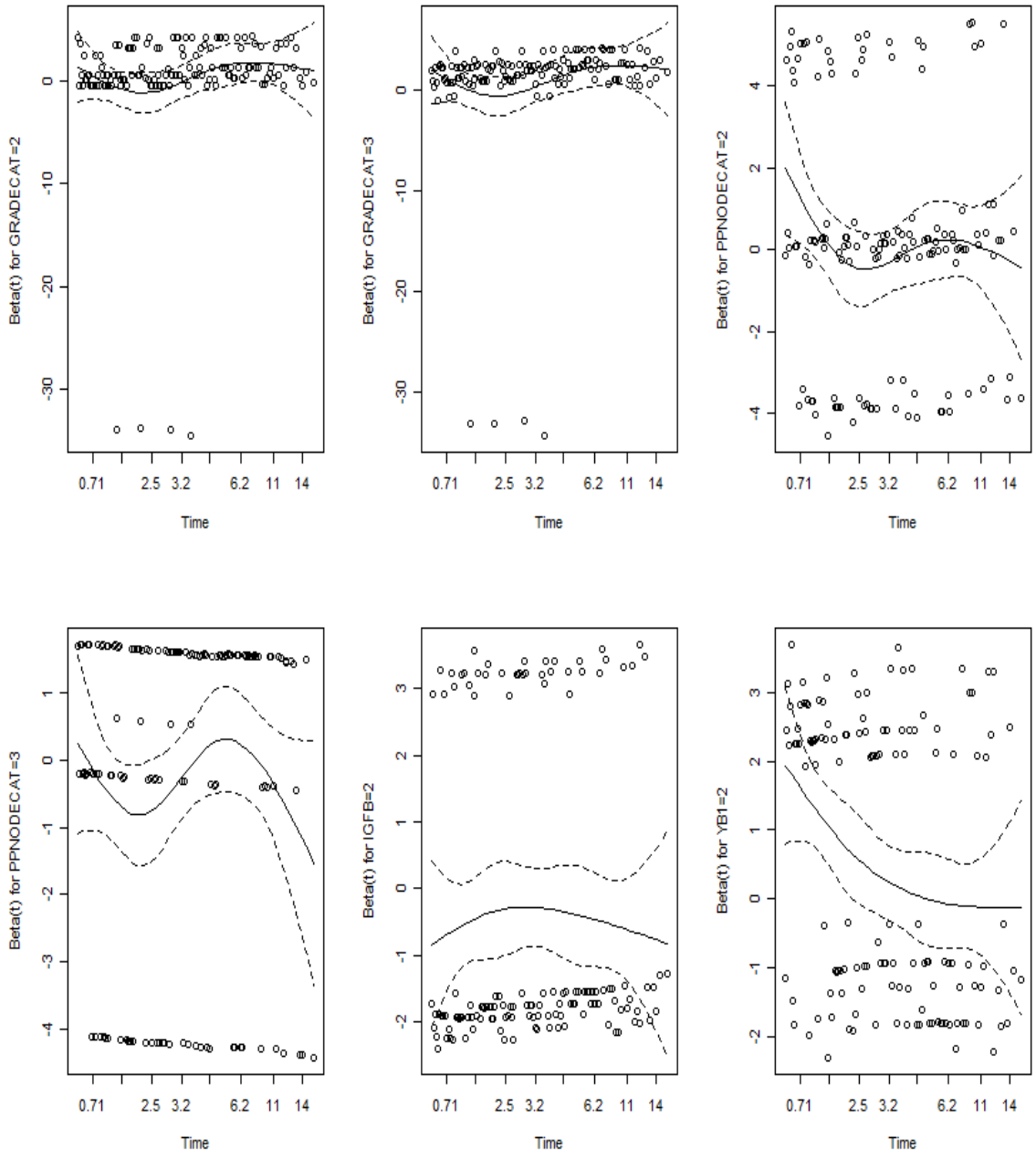


Figure 5.6: Plots of scaled Schoenfeld residuals against transformed time for each covariate ($n = 1575$, outcome=LRS).

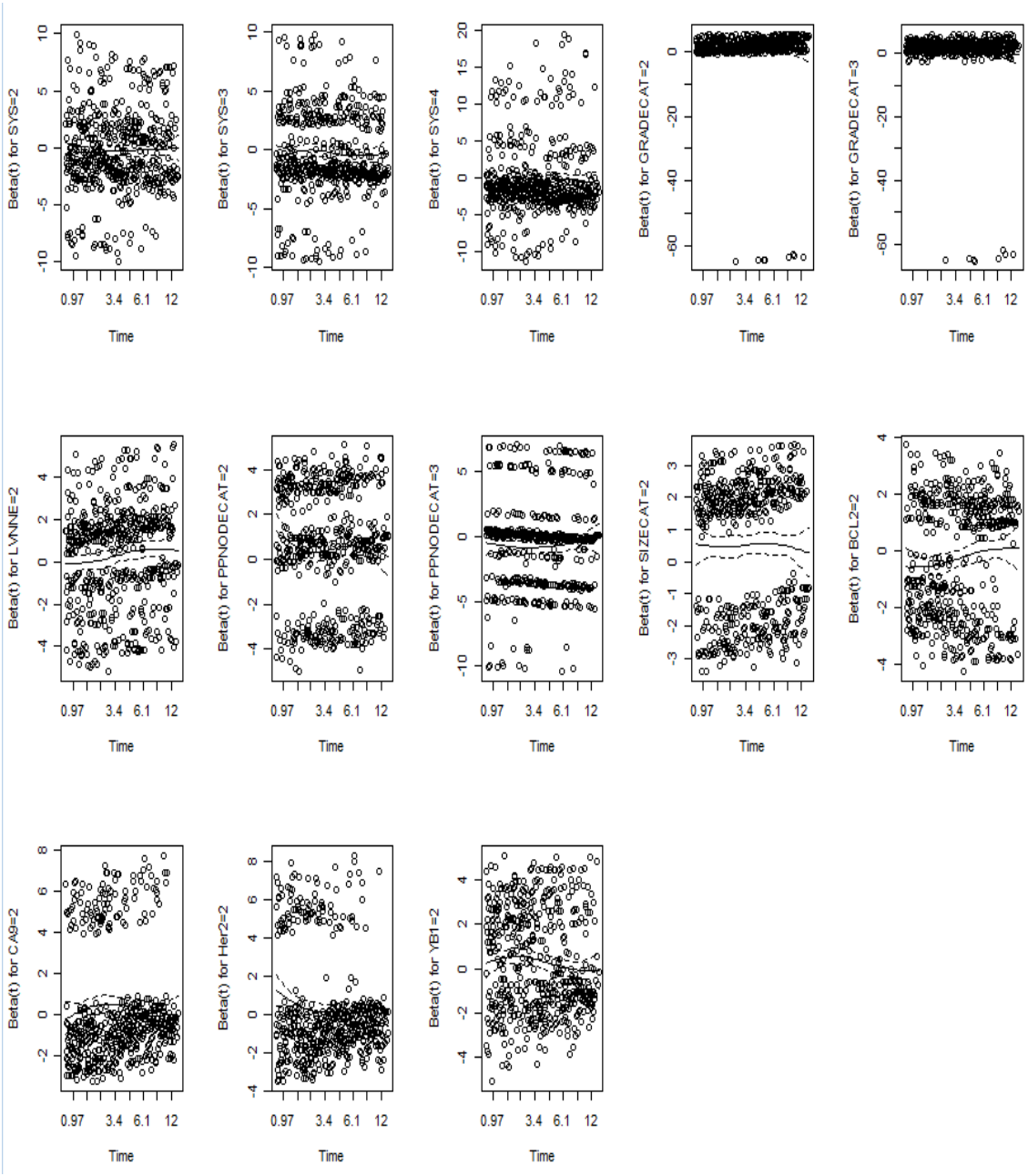


Figure 5.7: Plots of scaled Schoenfeld residuals against transformed time for each covariate ($n = 1575$, outcome=DRS).

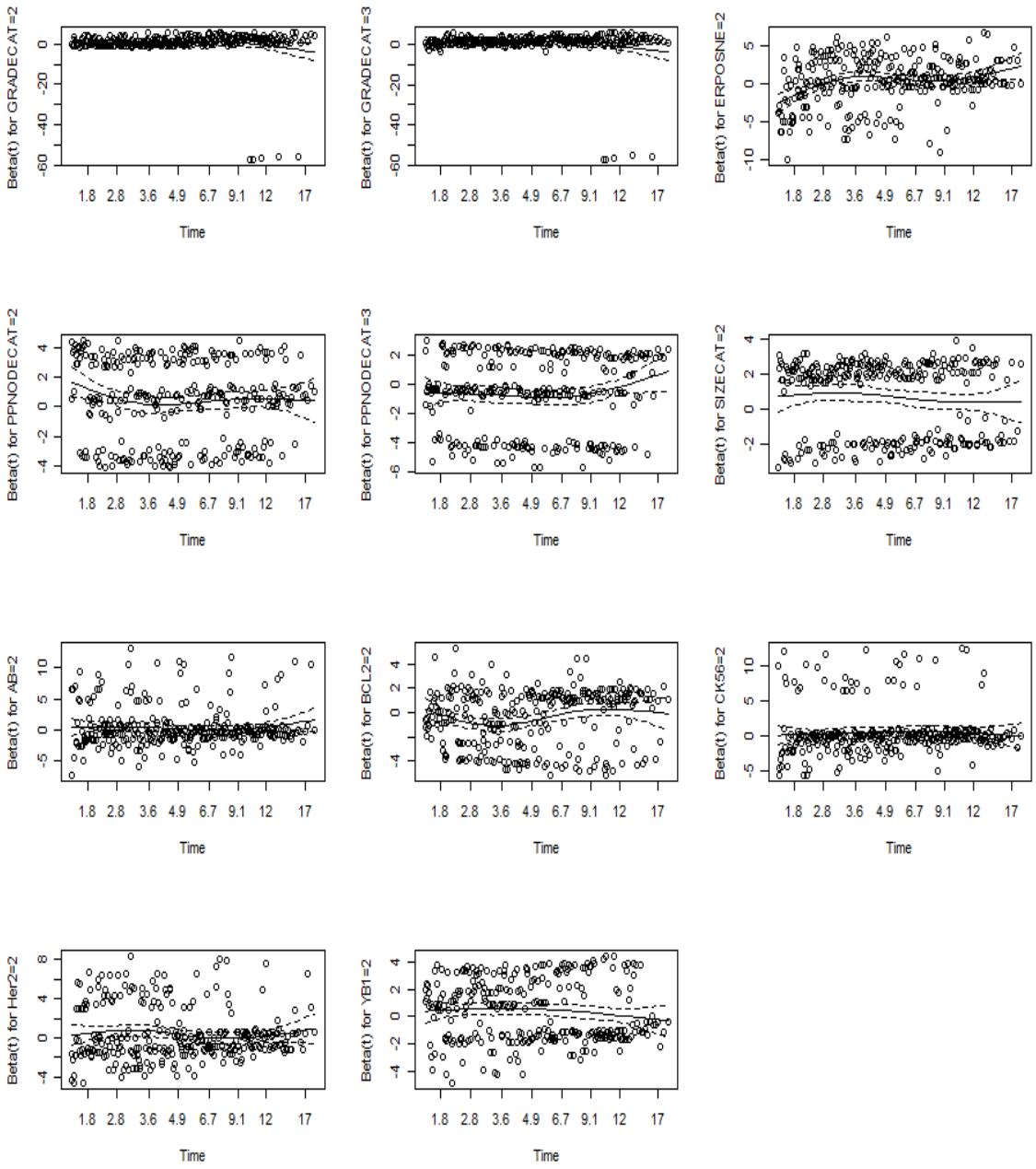


Figure 5.8: Plots of scaled Schoenfeld residuals against transformed time for each covariate ($n = 910$, outcome=BCSS).

Table 5.19: Comparison of estimated coefficients with and without stratification of a covariate with non-proportional hazards ($n = 1575$).

outcome=BCSS	coef	coef (stratifying BCL2)	coef diff
GRADECAT=2	0.584	0.615	0.0530
GRADECAT=3	0.868	0.903	0.0403
LVNNE=2	0.218	0.214	-0.018
PPNODECAT=2	0.619	0.617	-0.003
PPNODECAT=3	-0.421	-0.414	-0.016
SIZECAT=2	0.521	0.518	-0.005
BCL2=2	-0.286		
CK56=2	0.388	0.378	-0.025
Her2=2	0.254	0.214	-0.157
YB1=2	0.283	0.287	0.0141
outcome=LRS	coef	coef (stratifying YB1)	coef diff
GRADECAT=2	0.322	0.327	0.0155
GRADECAT=3	0.921	0.928	0.0076
PPNODECAT=2	0.162	0.139	-0.141
PPNODECAT=3	-0.304	-0.307	0.0098
IGFB=2	-0.472	-0.477	0.0105
YB1=2	0.527		
outcome=DRS	coef	coef (stratifying Her2)	coef diff
SYS=2	-0.293	-0.262	-0.105
SYS=3	-0.318	-0.296	-0.069
SYS=4	-0.454	-0.434	-0.044
GRADECAT=2	0.829	0.855	0.0313
GRADECAT=3	1.062	1.091	0.0273
LVNNE=2	0.295	0.295	0.000
PPNODECAT=2	0.539	0.533	-0.011
PPNODECAT=3	-0.677	-0.657	-0.029
SIZECAT=2	0.495	0.494	-0.002
BCL2=2	-0.26	-0.245	-0.057
CA9=2	0.255	0.250	-0.019
Her2=2	0.222		
YB1=2	0.276	0.275	-0.003

Table 5.20: Comparison of estimated coefficients with and without stratification of a covariate with non-proportional hazards ($n = 910$).

outcome=BCSS	coef	coef (stratifying BCL2)	coef diff
GRADECAT=2	0.389	0.444	0.1413
GRADECAT=3	0.732	0.764	0.0437
ERPOSNE=2	0.365	0.329	-0.098
PPNODECAT=2	0.492	0.499	0.0142
PPNODECAT=3	-0.601	-0.584	-0.028
SIZECAT=2	0.664	0.654	-0.015
AB=2	0.381	0.391	0.0262
BCL2=2	-0.345		
CK56=2	0.389	0.368	-0.053
Her2=2	0.359	0.305	-0.150
YB1=2	0.404	0.405	0.0024
outcome=LRS	coef	coef (stratifying YB1)	coef diff
GRADECAT=2	-0.448	-0.436	-0.026
GRADECAT=3	0.509	0.522	0.0255
PPNODECAT=2	0.205	0.171	-0.165
PPNODECAT=3	-0.419	-0.424	0.0119
IGFB=2	-0.427	-0.433	0.0140
YB1=2	0.757		
outcome=DRS	coef	coef (stratifying BCL2)	coef diff
GRADECAT=2	0.794	0.865	0.0894
GRADECAT=3	1.071	1.131	0.0560
PPNODECAT=2	0.445	0.43	-0.033
PPNODECAT=3	-0.681	-0.669	-0.017
SIZECAT=2	0.564	0.561	-0.005
BCL2=2	-0.247		
P53=2	0.227	0.225	-0.008
Her2=2	0.265	0.224	-0.154
YB1=2	0.310	0.314	0.0129

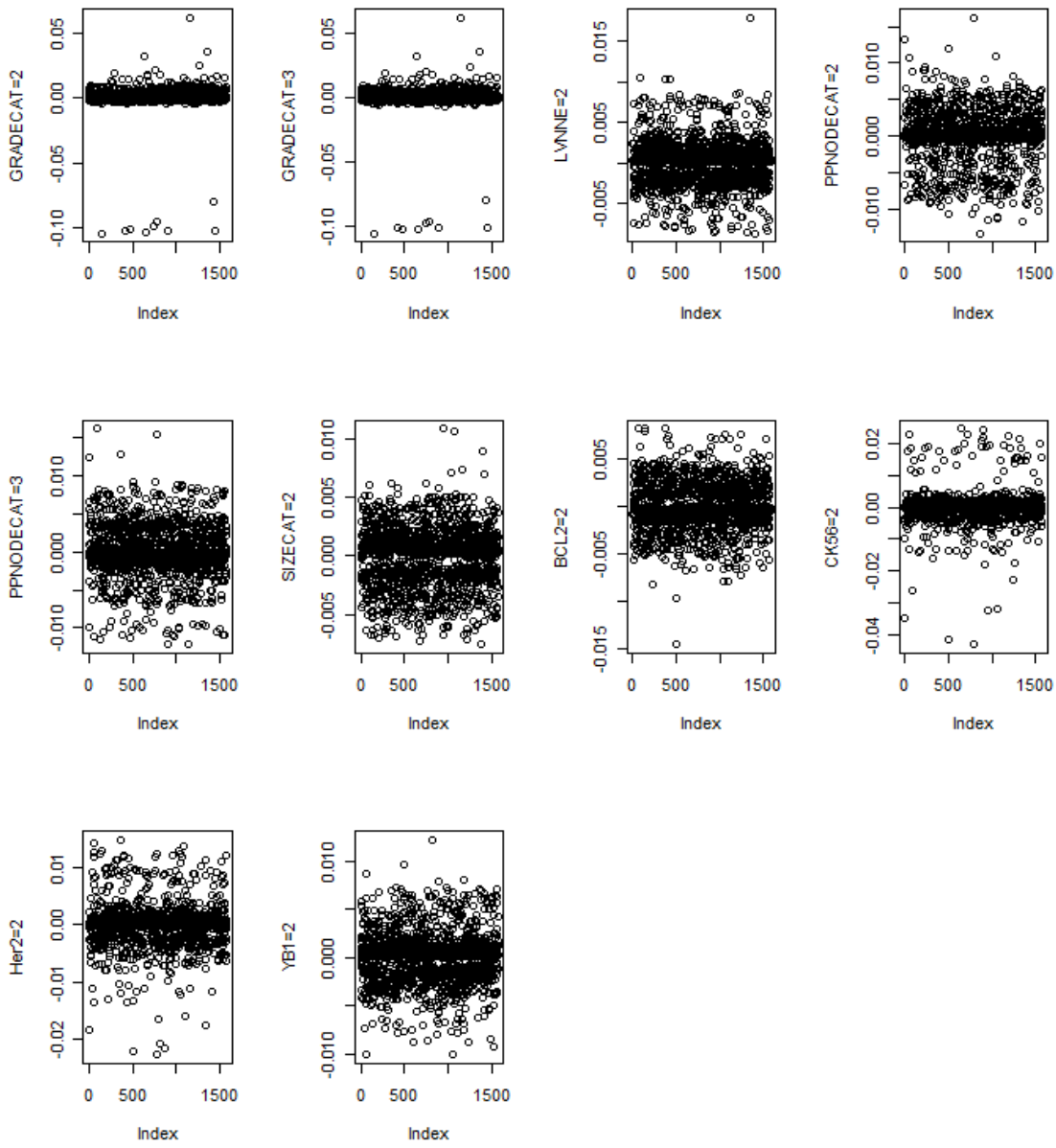


Figure 5.9: Index plots of dfbetas *for the Cox regression of time to each covariate ($n = 1575$, outcome=BCSS).*

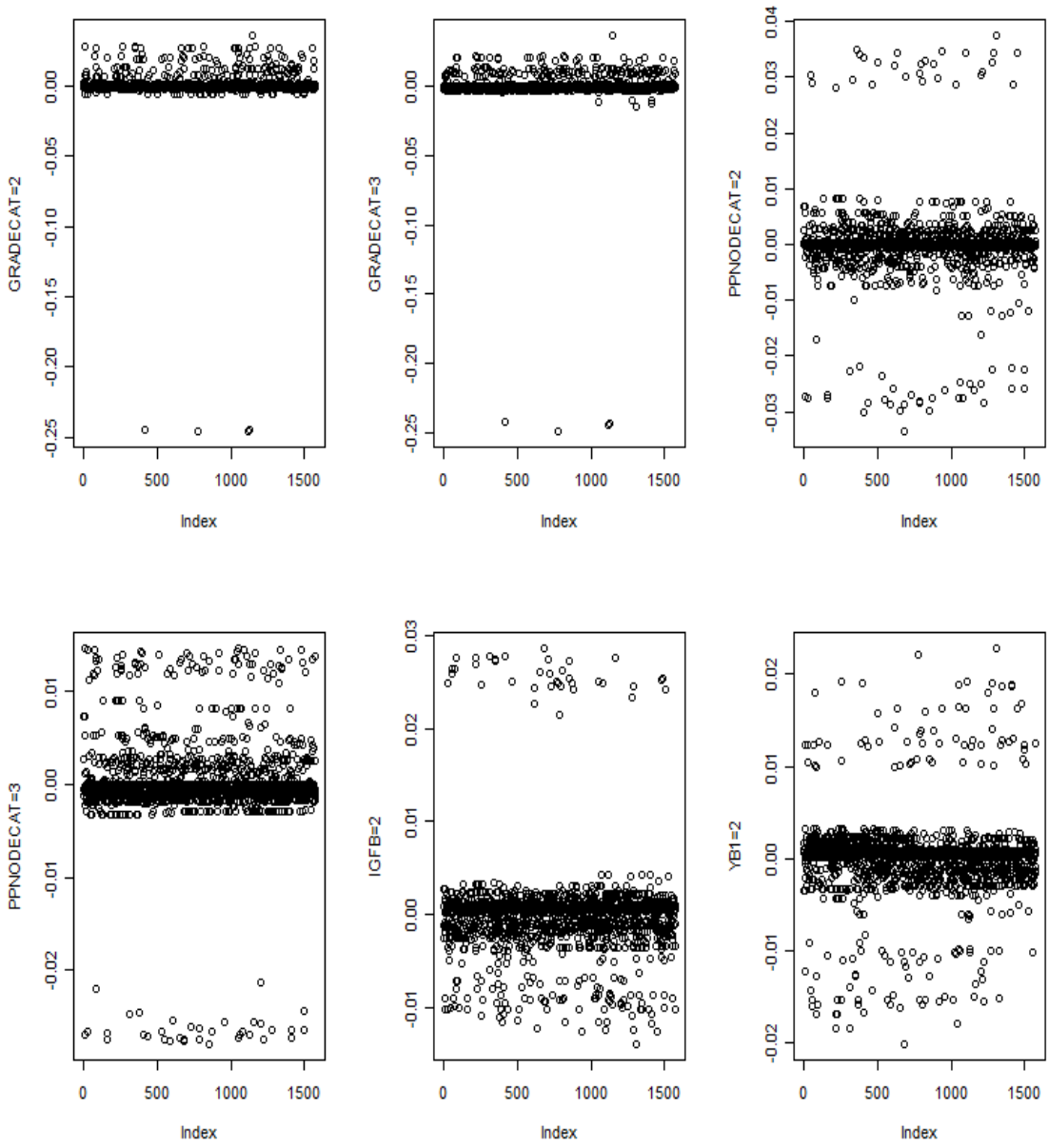


Figure 5.10: Index plots of dfbetas *for the Cox regression of time to each covariate ($n = 1575$, outcome=LRF).*

Chapter 6

Conclusions and Discussions

6.1 Multiple Imputation

Many imputation techniques are based on the idea that any subject in a study sample can be replaced by a new randomly chosen subject from the same source population. Multiple imputation can be applied very generally, to very large datasets with complex patterns of missingness among covariates, and uses only complete data quantities with very simple rules of combination. This makes it especially attractive for observational studies. In this study, three MI methods, DA, MICE and AREG, are employed and compared with each other. Results showed that AREG and MICE performed better than DA, which is consistent with the conclusion obtained by Faris *et al.* (2002) for a cardiac care dataset. In addition, AREG and MICE requires much shorter calculation time on computer than DA. Therefore they may be the preferred MI approaches for

other similar datasets. However, our analysis indicated that the MI performance was affected by the markers used in the model. Caution is needed when choosing an appropriate MI approach for different datasets.

Another issue related to MI is the different approaches, formal or alternative, when doing the survival analysis based on the MI results. The formal approach carries out m times survival modeling based on m imputed datasets from MI, while the alternative approach averages the m imputed datasets to get one final dataset for the survival modeling. It is interesting to notice that the modeling results using both approaches are nearly the same. Although the alternative approach does not make much sense mathematically, it demonstrates the reliability of MI results from another point of view.

6.2 Survival Models

Three outcomes, BCSS, LRS and DRS, are examined using two datasets, the large dataset with missing values in markers (n=1575) and the small (complete) dataset without missing (n=910). It is not surprising to see that results for BCSS and DRS are similar because with the $n = 1575$ dataset 95% BCSS patients have a distant relapse (451 out of 476), which accounts for 87% of total DRS patients (451 out of 518). On the other hand, only 76 BCSS patients have a local relapse which is 55% of the total LRS patients (76 out of 138). The rate for breast cancer with local relapse

in the dataset is relatively low. People with clinical interest may pay more attention on this point.

6.2.1 Univariable models

All clinical variables are significantly associated with BCSS and DRS, although AGE shows only moderate evidence of significance. For local relapse survival, GRADE-CAT shows strong evidence affecting LRS with both datasets, while ERPOSNE shows moderate evidence using the small dataset only. It is not clear why so many clinical variables are insignificant for LRS although research work did show the importance of some clinical variables on LRS (Casalini *et al.* 2008). One reason may be due to the small number of LRS events (see tables in Appendix B).

As for the markers, IGFB is the only insignificant covariate for BCSS and DRS, indicating the importance of markers. For LRS, four markers, YB1, IGFB, PR, and BCL2, are significant using the large dataset, while only YB1 is significant with the small dataset. The univariable analysis results indeed suggest that dropping all the missing values in the dataset may cause biased conclusions in which some important markers may be ignored.

6.2.2 Multivariable models

A very important result this work obtains is that markers can be independent predictors of breast cancer outcomes! The fitting results using the large dataset and the small

dataset are similar, and we think results using the large dataset are more representive. For BCSS and DRS outcomes, four markers (BCL2, CK56, Her2, YB1) remain in the reduced model and all of them except CK56 have been used as prognostic factors to more accurately predict clinical outcome and guide therapies. More clinical research work could be done to identify the importance of CK56. For the outcome of LRS, two markers, IGFB and YB1, remain in the reduced model. YB1 is the only marker to be strongly significant on all the three outcomes. Our results suggest that YB1 could be added to a standard pathology report, although further investigations, both statistically and experimentally, are needed to confirm and clarify it roles.

Another interesting point is that many important markers remained in the final reduced models as independent covariates show strong time-varying effects on outcomes. This result is supported by some available research work. For example, BCL2 can be an independent predictor of breast cancer outcome particularly in the first 5 years after diagnosis (Callagy et al. 2006).

The 10 year survival probability at the mean of all the covariates (clinical variables and markers) for BCSS, LRS, and DRS is 77%, 91%, and 72% respectively. The survival proportion between one year and 5 years shows a relatively quick drop, and then the decrease of survival proportion gets mild after 5 years.

In conclusion, this work provides some statistical indicators which are worthy of further clinical research proof and/or clarification.

6.3 Further Work

We did not fully address the effect of non-proportional hazards since we focused on finding independent patterns for survival. Some covariates were indeed time varying. If the aim is to give a detailed description of covariate effects and to accurately calculate predicted probabilities, more flexible models are needed. Cox-Aalen model, by multiplying the additive and multiplicative hazards models, could fully represent the time varying effect, and has been applied in breast cancer survival (Baldi *et al.* 2006).

The final set of covariate predictors are likely not unique. Further study may focus on specific subsets of patients for specific markers.

Some markers have demonstrated time-varying effects on outcomes. Further and detailed check of these time-varying effects, such as significant change on different time periods, would be of much clinical interest.

Bibliography

[1] Aalen OO. (1975). Statistical inference for a family of counting processes. PhD thesis, Univ. of California, Berkeley.

[2] Ambler G, Omar RZ, Royston P. (2007). A comparison of imputation techniques for handing missing predicator values in a risk model with a binary outcome, *Statistical Methods in Medical Research*, 16, 277-298.

[3] Andersen PK, Borgan O, Gill RD, Keiding N. (1993). Statistical Models Based on Counting Processes. Springer, New York.

[4] Asselin-Labat ML, Sutherland KD, Barker H, et al. (2007). Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *NATURE CELL BIOLOGY*, 9, 201-209.

[5] Baldi I, Ciccone G, Ponti A, Rosso S, Zanetti R, Gregori D. (2006). An Application of the Cox-Aalen Model for Breast Cancer Survival. *AUSTRIAN JOURNAL OF STATISTICS*, 35, 77-88.

[6] Banerjee S, Reis-Filho JS, Ashley S, et al. (2006). Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *J Clin Pathol*, 59, 729-735.

[7] Barnard J and Rubin DB. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. Biometrika, 86, 948-955.

[8] BC Cancer Agency Online. (2010). Cancer Statistics. http://www.bccancer.bc.ca/ABCCA/NewsCentre/stats.htm#breast. Accessed Feb. 2010.

[9] Beenken SW, Grizzle WE, Crowe DR, et al. (2001). Molecular biomarkers for breast cancer prognosis: Coexpression of cerbB2 and p53. *ANNALS OF SURGERY*, 233, 630-637.

[10] Breastcancer.org Online. (2009). Breast Cancer Staging. http://www.breastcancer.org/symptoms/diagnosis/staging.jsp. Accessed July 2009.

[11] Breastcancer.org Report. (2009). Your Guide to the Breast Cancer Pathology Report. download July 2009 from http://www.herceptin.com/pdf/pathology-report.pdf.

[12] Bonetti A, Zaninelli M, Leone R, Cetto GL, Pelosi G, Biolo S, Menghi A, Manfrin E, Bonetti F, Piubello Q. (1998). bcl-2 but not p53 expression is associated with

resistance to chemotherapy in advanced breast cancer. *Clin Cancer Res*, 4, 2331-2336.

[13] Bottini A, Berruti A, Bersiga A, Brizzi MP, Brunelli A, Gorzegno G, DiMarco B, Aguggini S, Bolsi G, Cirillo F, Filippini L, Betri E, Bertoli G, Alquati P, Dogliotti L. (2000). p53 but not bcl-2 immunostaining is predictive of poor clinical complete response to primary chemotherapy in breast cancer patients. *Clin Cancer Res*, 6, 2751-2758.

[14] Buchholz TA, Davis DW, McConkey DJ, Symmans WF, Valero V, Jhingran A, Tucker SL, Pusztai L, Cristofanilli M, Esteva FJ, Hortobagyi GN, Sahin AA. (2003). Chemotherapy-induced apoptosis and Bcl-2 levels correlate with breast cancer response to chemotherapy. *Cancer J*, 9, 33-41.

[15] Callagy G, Pharoah PD, Pinder SE, et al. (2006). Bcl2 is a prognosticMarker in breast cancer independently of the Nottingham Prognostic Index. *Clin Cancer Res*, 12 (8), 2468-2475.

[16] Casalini P, Carcangiu ML, Tammi R, Auvinen P, Kosma V, Valagussa P. (2008). Two Distinct Local Relapse Subtypes in Invasive Breast Cancer: Effect on their Prognostic Impact *Clin Cancer Res*, 25, 25-31.

[17] Cheang MCU, Voduc D, Bajdik C, et al. (2008). Basal like breast cancer defined by five biomarkers has superior prognostic values than triple negative phenotype.

Clin Cancer Res, 14 (5), 1368-1376.

[18] Colleoni M, Viale G, Zahrieh D, et al. (2008). Expression of ER, PgR, HER1, HER2, and response: a study of preoperative chemotherapy. *ANNALS OF ONCOLOGY*, 19, 465-472.

[19] Collins LM, Schafer JL, Kam C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6 (4), 330-351.

[20] Crowe Jr JP, Gordon NH, Hubay CA, et al. (1991). Estrogen receptor determination and long term survival of patients with carcinoma of the breast. *Surg Gynecol Obstet*, 173, 273-278.

[21] Cornez N, Piccart MJ. (2000). Breast cancer and herceptin. *Bull Cancer*, 87, 847-858.

[22] Cox DR. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistics Society*, Series B 34, 187-220.

[23] Donders ART, van der Heijden GJMG, Stijnen T, et al. (2006). Review: A gentle introduction to imputation of missing values, *Journal of Clinical Epidemiology*, 59, 1087-1091.

[24] Early Breast Cancer Trialists Collaborative Group. (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*, 351, 1451-1467.

[25] Eide GE, Omenaas E, Gulsvik A. (1996). The semiparametric proportional hazards model revisited: Practical reparametisations. *Statistics in Medicine*, 15, 1771-1777.

[26] Esteva1 FJ and Hortobagyi GN. (2004). Prognostic molecular markers in early breast cancer. *Breast Cancer Res*, 6, 109-118.

[27] Fadare O, Wang SA, Hileeto D. (2008). The expression of cytokeratin 5/6 in invasive lobular carcinoma of the breast: evidence of a basal-like subset? *HUMAN PATHOLOGY*, 39, 331-336.

[28] Fang SH, Chen YZ, Weigel RJ. (2009). GATA-3 as a Marker of Hormone Response in Breast Cancer. *JOURNAL OF SURGICAL RESEARCH*, 157, 290-295.

[29] Faris PD, Ghali WA, Brant R. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *JOURNAL OF CLINICAL EPIDEMIOLOGY*, 55, 184-191.

[30] Fleming TR and Harrington DP. (1991). Counting Processes and Survival Analysis. Wiley, New York.

[31] Gasparini G, Barbareschi M, Doglioni C, Palma PD, Mauri FA, Boracchi P, Bevilacqua P, Caffo O, Morelli L, Verderio P. (1995). Expression of bcl-2 protein predicts efficacy of adjuvant treatments in operable node-positive breast cancer. *Clin Cancer Res*, 1, 189-198.

[32] Gradishar WJ (2005). The future of breast cancer: the role of prognostic factors. *Breast Cancer Res Treat*, 89, S17-S26.

[33] Graham JW and Donaldson SI. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, 78, 119-128.

[34] Graham JW and Hofer SM. (2000). Multiple imputation in multivariate research. In TD Little, KU Schnabel and J Baumert, (Eds.), Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples. (pp. 201-218). Hillsdale, NJ: Erlbaum.

[35] Graham JW, Cumsille PE, Elek-Fisk E. (2003). Methods for handling missing data. In JA Schinka and WF Velicer (Eds.). Research Methods in Psychology (pp. 87-114). Volume 2 of Handbook of Psychology (I.B. Weiner, Editor-in-Chief). New York: John Wiley and Sons.

[36] Graham JW, Hofer SM, Donaldson SI, MacKinnon DP, Schafer JL. (1997). Analysis with missing data in prevention research. In K Bryant, M Windle, and S West (Eds.), The science of prevention: methodological advances from alcohol and substance abuse research. (pp 325-366). Washington, D.C.: American Psychological Association.

[37] Habibi G, Leung S, Law JH, et al. (2008). Redefining prognostic factors for breast cancer: YB-1 is a stronger predictor of relapse and disease-specific survival than estrogen receptor or HER-2 across all tumor subtypes. *BREAST CANCER RESEARCH*, 10, R86(1-9).

[38] Henry NL and Hayes DF. (2006). Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. *The Oncologist*, 11, 541-552.

[39] Hox JJ. (1999). A review of current software for handling missing data. *Kwantitatieve Methoden*, 20 (62), 123-138.

[40] Hwa V, Oh Y, Rosenfeld RG. (1999). The insulin-like growth factor-binding protein (IGFBP) superfamily. *Endocr Rev*, 20, 761-787.

[41] Kalbfleisch JD and Prentice RL. (2002). The Statistical Analysis of Failure Time Data. Wiley, New York.

[42] Kaplan EL and Meier P. (1958). Nonparametric estimation for incomplete observations. *J Amer Statist Assoc*, 53, 451-481.

[43] Keinan-Boker L, Bas Bueno de Mesquita H, Kaaks R, van Gils CH, van Noord PAH, Rinaldi S, Riboli E, Seidell JC, Grobbee DE, Peeters PHM. (2003). Circulating levels of insulin-like growth factor I, its binding proteins -1, -2, -3, c-peptide and risk of postmenopausal breast cancer. *Int J Cancer*, 106, 90-95.

[44] Keshgegian AA, Cnaan A. (1995). Proliferation markers in breast carcinoma. Mitotic figure count, S-phase fraction, proliferating cell nuclear antigen, Ki-67 and MIB-1. *Am J Clin Pathol*, 104, 42-49.

[45] Klein JP and Moeschberger ML. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag Inc.

[46] Knight WA, Livingston RB, Gregory EJ, McGuire WL (1977). Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. *Cancer Res*, 37, 4669-4671.

[47] Kostler WJ, Steger GG, Soleiman A, et al. (2004). Monitoring of serum Her-2/neu predicts histopathological response to neoadjuvant trastuzumab-based therapy for breast cancer. *Anticancer Res*, 24, 1127-130.

[48] Krajcik RA, Borofsky ND, Massardo S, Orentreich. (2002). Insulin-like growth factor I (IGF-I), IGF-binding proteins, and breast cancer. *Cancer Epidemiol Biomark Prev*, 11, 1566-1573.

[49] Lawless JF. (2003). Statistical Models and Methods for Lifetime Data. New York: John Wiley and Sons.

[50] Lawless JF, Singhal K. (1978). EFFICIENT SCREENING OF NONNORMAL REGRESSION-MODELS. *BIOMETRICS*, 34, 318-327.

[51] Levin ER. (2005). Integration of the extranuclear and nuclear actions of estrogen. *Mol Endocrinol*, 19, 1951-1959.

[52] Little RJA. (1992). Regression with missing Xs: A review. *Journal of the American Statistical Association*, 87 (420) 1227-1237.

[53] Little RJA. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90 (431), 1112-1121.

[54] Little RJA and Rubin DB. (1987). Statistical analysis with missing data. New York: John Wiley and Sons.

[55] Little RJA and Rubin DB. (2002). Statistical analysis with missing data. New York: John Wiley and Sons.

[56] Little RJA and Schenker N. (1995). Missing data. In G Arminger, CC Clogg, DB Sobel (Eds.) Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York, NY: Plenum.

[57] Martinussen T and Scheike TH. (2006). Dynamic Regression Models for Survival Data. Springer, New York.

[58] McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. (2005). Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst*, 97, 1180-1184.

[59] Mehra R, Varambally S, Ding L, et al. (2005). Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *CANCER RESEARCH*, 65, 11259-11264.

[60] Moons KGM, Donders RART, Stijnen T, et al. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59, 1092-1101.

[61] Moyano JV, Evans JR, Chen F, et al. (2006). aB-Crystallin is a novel oncoprotein that predicts poor clinical outcome in breast cancer. *J Clin Invest*, 116, 261-270.

[62] Muss HB, Thor AD, Berry DA, et al. (1994). c-erbB-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *N Engl J Med*, 330, 1260-1266.

[63] Nielsen TO, Hsu FD, Jensen K, Cheang M, et al. (2004). Immunohistochemical and clinical characterization of the basal- like subtype of invasive breast carcinoma. *Clinical Cancer Research*, 10, 5367-5374.

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

[65] Olsen DA, Ostergaard B, Bokmand S, et al. (2009). HER1-4 protein concentrations in normal breast tissue from breast cancer patients are expressed by the same

profile as in the malignant tissue *CLINICAL CHEMISTRY AND LABORATORY MEDICINE*, 47, 977-984.

[66] Osborne CK, Yochmowitz MG, Knight WA, McGuire WL. (1980). The value of estrogen and progesterone receptors in the treatment of breast cancer. *Cancer*, 46(12 Suppl), 2884-2888.

[67] Perou CM, et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406, 747-752.

[68] Pharaoh PDP, Day NE, Caldas C. (1999). Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *Br J Cancer*, 80, 1968-1973..

[69] Poelman SM, Adeyanju MO, Robertson MA, Recant WM, Karrison T, Fleming GF, Olopade OI, Conzen SD. (2000). Human breast cancer susceptibility to paclitaxel therapy is independent of Bcl-2 expression. *Clin Cancer Res*, 6, 4043-4048.

[70] Ponzone R, Montemurro F, Maggiorotto F et al. (2006). Clinical outcome of adjuvant endocrine treatment according to PR and HER-2 status in early breast cancer. *Annals of Oncology*, 17, 1631-1636.

[71] R Documentation, Function "aregImpute" in package *Hmisc*. (2009). http://sekhon.berkeley.edu/library/Hmisc/html/aregImpute.html. Accessed July 2009.

[72] Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85-95.

[73] Rakha EA, El-Sayed ME, Green AR, et al. (2007). Prognostic markers in triple-negative breast cancer. *CANCER*, 109, 25-32.

[74] Ross JS, Fletcher JA, Bloom KJ, et al. (2004). Targeted therapy in breast cancer: the HER-2/neu gene and protein. *Mol Cell Proteomics*, 3, 379-98.

[75] Royston P. (2004). Multiple imputation of missing values. *Stata Journal*, 4, 227-241.

[76] Royston P. (2005). Multiple imputation of missing values: Update of MICE. *Stata Journal*, 5, 527-536.

[77] Rubin DB. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

[78] Rubin DB. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91 (434), 473-489.

[79] Sahin AA, Ro J, Ro JY, Blick MB, el-Naggar AK, Ordonez NG, Fritsche HA, Smith TL, Hortobagyi GN, Ayala AG. (1991). Ki-67 immunostaining in node-negative stage I/II breast carcinoma. Significant correlation with prognosis. *Cancer*, 68, 549-557.

[80] Sasa M, Bando Y, akahashi M, et al. (2008). Screening for Basal Marker Expression Is Necessary for Decision of Therapeutic Strategy for Triple-Negative Breast Cancer. *Journal of Surgical Oncology*, 97, 30-34.

[81] Schafer JL. (2002). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

[82] Schafer JL and Graham JW. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2), 147-177.

[83] Schafer JL and Olsen MK. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, 33 (4), 545-571.

[84] Schairer C, Hill D, Sturgeon SR, et al. (2004). Serum concentrations of IGF-I, IGFBP-3 and C-peptide and risk of hyperplasia and cancer of the breast in postmenopausal women. *INTERNATIONAL JOURNAL OF CANCER*, 108, 773-779.

[85] Sinharay S, Stern HS, Russell D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6 (4), 317-329.

[86] Span PN, Bussink J, Manders P, et al. (2003). Carbonic anhydrase-9 expression levels and prognosis in human breast cancer: association with treatment outcome. *BRITISH JOURNAL OF CANCER*, 89, 271-276.

[87] Tableman M and Kim JS. (2003). Sunrival Analysis Using S. Chapman and Hall/CRC.

[88] Tanner MA and Wong WH. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528- 550.

[89] Therneau T and Grambsch P. (2000). Modeling Survival Data: Extending the Cox Model. Springer-Verlag New York.

[90] Thor AD, Moore DH II, Edgerton SM, et al. (1992). Accumulation of p53 tumor suppressor gene protein: an independent marker of prognosis in breast cancers. *J Natl Cancer Inst*, 84, 845-855.

[91] Toniolo P, Bruning PF, Akhmedkhanov A, Bonfrer JMG, Koenig KL, Lukanova A, Shore RE, Zeleniuch-Jacquotte A. (2000). Serum insulin-like growth factor-I and breast cancer. *Int J Cancer*, 88, 828-832.

[92] Van Buuren S, Brand JP, Gruthuis-Odshoorn CG, Dubin D. (2006). Fully conditional specification in multiple imputation. *Journal of Statistical Computation and Simulation*, 76, 1049-1064.

[93] Veronese SM, Gambacorta M, Gottardi O, Scanzi F, Ferrari M, Lampertico P. (1993). Proliferation index as a prognostic marker in breast cancer. *Cancer*, 71, 3926-3931.

[94] Voduc D, Cheang M, Nielsen T. (2008). GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *CANCER EPIDEMIOLOGY BIOMARKERS and PREVENTION*, 17, 365-373.

[95] Voorzanger-Rousselot N and Garnero P. (2007). Biochemical markers in oncology. Part I: Molecular basis. Part II: Clinical uses. *Cancer Treatment Reviews*, 33, 230-283.

[96] Wang Y, Szekely L, Okan I, et al. (1003). Wild-type p53 triggered apoptosis is inhibited by bcl-2 in a v-myc induced T-cell lymphoma cell line. *Oncogene*, 8, 3427-3431.

[97] Watson PH, Chia SK, Wykoff CC, et al. (2003). Carbonic anhydrase XII is a marker of good prognosis in invasive breast carcinoma. *BRITISH JOURNAL OF CANCER*, 88, 1065-1070.

[98] Wayman JC. (2002a). The utility of educational resilience for studying degree attainment in school dropouts. *Journal of Educational Research*, 95 (3), 167-178.

[99] Wayman JC. (2002b). Practical Considerations in Constructing a Multiple Imputation Model A Data Example. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

[100] Well Sphare Online. (2010). http://www.wellsphere.com/complementary-alternative-medicine-article/the-survival-rate-for-women-with-breast-cancer/597465. Accessed May 2010.

[101] Wiki Online. (2009). http://en.wikipedia.org/wiki/Breast_cancer. Accessed July 2009.

[102] Wolpin BM, Meyerhardt JA, Chan AT, et al. (2009). Insulin, the Insulin-Like Growth Factor Axis, and Mortality in Patients With Nonmetastatic Colorectal Cancer. *JOURNAL OF CLINICAL ONCOLOGY*, 27, 176-185.

[103] Wu J, Lee C, Yokom D, Jiang H, Cheang MCU, Yorida E, Turbin D, Berquin IM, Mertens PR, Iftner T, Gilks B, Dunn SE. (2006). Disruption of the Y-box binding protein-1 (YB-1) results in suppression of the epidermal growth factor receptor and Her-2. *Cancer Res*, 66, 4872-4879.

[104] Wykoff CC, Beasley NJ, Watson PH, Turner KJ, Pastorek J, Sibtain A, Wilson GD, Turley H, Talks KL, Maxwell PH, Pugh CW, Ratcliffe PJ, Harris AL. (2000). Hypoxia-inducible expression of tumor-associated carbonic anhydrases. *Cancer Res*, 60, 7075-7083.

[105] Yu H, Jin F, Shu XO, Li BDL, Dai Q, Cheng JR, Berkel HJ, Zheng W. (2002). Insulin-like growth factors and breast cancer risk in Chinese women. *Cancer Epidemiol Biomark Prev*, 11, 705-712.

Appendix A

Sensitivity analysis for three MI methods using all other markers except for AB, Her2, and PR (n = 1575, missing proportion = 5%)

Table A.1: Sensitivity analysis for three MI methods (n = 1575, missing = 5%).

MI method	Marker		sensitivity	selectivity	PPV	NPV	agreement
AREG	BCL2	Mean	0.6825	0.6423	0.6841	0.6397	0.6622
		SE	0.0071	0.0079	0.0076	0.0077	0.0059
	CK56	Mean	0.6507	0.9391	0.6518	0.9397	0.8960
		SE	0.0118	0.0031	0.0136	0.0028	0.0034
	GATA	Mean	0.5097	0.7271	0.5114	0.7262	0.6477
		SE	0.0081	0.0063	0.0086	0.0060	0.0054
	Her1	Mean	0.6209	0.9134	0.6156	0.9162	0.8596
		SE	0.0111	0.0037	0.0123	0.0031	0.0038
	ERS	Mean	0.8127	0.7160	0.8080	0.7212	0.7718
		SE	0.0061	0.0072	0.0057	0.0085	0.0050
	CA9	Mean	0.6050	0.8454	0.6021	0.8493	0.7781
		SE	0.0083	0.0050	0.0106	0.0039	0.0045
	P53	Mean	0.5049	0.8158	0.5167	0.8069	0.7265
		SE	0.0096	0.0046	0.0087	0.0055	0.0049
	IGFB	Mean	0.4930	0.5918	0.4888	0.5960	0.5468
		SE	0.0079	0.0075	0.0079	0.0073	0.0060
	YB1	Mean	0.5382	0.6847	0.5490	0.6755	0.6221
		SE	0.0074	0.0063	0.0073	0.0062	0.0049
MICE	BCL2	Mean	0.6936	0.6397	0.6877	0.6454	0.6660
		SE	0.0069	0.0082	0.0074	0.0078	0.0054
	CK56	Mean	0.6180	0.9430	0.6630	0.9314	0.8921
		SE	0.0127	0.0029	0.0134	0.0032	0.0036
	GATA	Mean	0.5174	0.7305	0.5175	0.7296	0.6525
		SE	0.0082	0.0055	0.0076	0.0061	0.0049
	Her1	Mean	0.6152	0.9190	0.6366	0.9119	0.8612
		SE	0.0098	0.0031	0.0106	0.0033	0.0036
	ERS	Mean	0.8126	0.7196	0.8106	0.7211	0.7736
		SE	0.0053	0.0071	0.0051	0.0078	0.0045
	CA9	Mean	0.5931	0.8452	0.6009	0.8412	0.7727
		SE	0.0090	0.0051	0.0106	0.0047	0.0046
	P53	Mean	0.5156	0.8141	0.5083	0.8176	0.7312
		SE	0.0100	0.0045	0.0090	0.0050	0.0046
	IGFB	Mean	0.4787	0.5902	0.4754	0.5931	0.5393
		SE	0.0086	0.0070	0.0073	0.0078	0.0059
	YB1	Mean	0.5530	0.6910	0.5609	0.6835	0.6316
		SE	0.0077	0.0064	0.0081	0.0068	0.0053
DA	BCL2	Mean	0.6528	0.5945	0.6435	0.6039	0.6231
		SE	0.0048	0.0042	0.0042	0.0055	0.0026
	CK56	Mean	0.3513	0.9372	0.6482	0.8175	0.7934
		SE	0.0041	0.0026	0.0114	0.0024	0.0023
	GATA	Mean	0.4240	0.7162	0.5002	0.6502	0.5975
		SE	0.0044	0.0051	0.0065	0.0037	0.0030
	Her1	Mean	0.4019	0.9107	0.6167	0.8112	0.7766
		SE	0.0049	0.0027	0.0094	0.0025	0.0025
	ERS	Mean	0.7869	0.6351	0.7435	0.6889	0.7207
		SE	0.0044	0.0040	0.0028	0.0061	0.0028
	CA9	Mean	0.4476	0.8434	0.6056	0.7399	0.7037
		SE	0.0046	0.0042	0.0079	0.0033	0.0030
	P53	Mean	0.4097	0.8191	0.5281	0.7383	0.6829
		SE	0.0051	0.0043	0.0078	0.0034	0.0032
	IGFB	Mean	0.4571	0.6041	0.4922	0.5697	0.5355
		SE	0.0041	0.0048	0.0054	0.0040	0.0027
	YB1	Mean	0.4871	0.6606	0.5221	0.6288	0.5838
		SE	0.0048	0.0056	0.0066	0.0040	0.0034

Appendix B

Evaluation of three MI methods using the dataset without missing values (n = 910)

Table B.1: Sensitivity and agreement analysis: marker = AB.

MI method	missing %	sensitivity	selectivity	PPV	NPV	agreement
AREG	2	0.4671	0.9197	NaN	0.9128	0.8536
	5	0.4825	0.9124	0.4679	0.9145	0.85
	15	0.4478	0.9128	0.4570	0.9086	0.8460
	35	0.5126	0.9078	0.5245	0.9030	0.8417
MICE	2	NaN	0.9120	NaN	0.8894	0.8268
	5	0.4271	0.9141	0.4736	0.9011	0.8404
	15	0.4025	0.9096	0.4391	0.8930	0.8306
	35	0.4796	0.9080	0.5423	0.8817	0.8263
DA	2	0.2017	0.8984	NaN	0.7775	0.7267
	5	0.2262	0.9021	0.4434	0.7785	0.7330
	15	0.2539	0.9068	0.4877	0.7787	0.7390
	35	0.3285	0.9019	0.5997	0.7498	0.7245

Table B.2: Sensitivity and agreement analysis: marker = Her2.

MI method	missing %	sensitivity	selectivity	PPV	NPV	agreement
AREG	2	0.4798	0.8430	0.4502	0.8560	0.76
	5	0.5693	0.8434	0.5481	0.8550	0.7728
	15	0.5321	0.8497	0.5275	0.8512	0.7723
	35	0.5185	0.8461	0.5272	0.8405	0.7639
MICE	2	0.4305	0.8504	NaN	0.8504	0.7578
	5	0.5298	0.8373	0.5271	0.8347	0.7541
	15	0.4993	0.8537	0.5434	0.8265	0.7589
	35	0.5005	0.8461	0.5436	0.8189	0.7511
DA	2	0.2839	0.8588	NaN	0.7461	0.6883
	5	0.3689	0.8369	0.5428	0.7194	0.6757
	15	0.3549	0.8465	0.5489	0.7141	0.6767
	35	0.3828	0.8413	0.5854	0.6989	0.6715

Table B.3: Sensitivity and agreement analysis: marker = PR.

MI method	missing %	sensitivity	selectivity	PPV	NPV	agreement
AREG	2	0.6551	0.6475	0.6689	0.6310	0.6436
	5	0.6249	0.6449	0.6193	0.6493	0.6334
	15	0.6489	0.6095	0.6503	0.6073	0.6292
	35	0.6457	0.6133	0.6467	0.6112	0.6295
MICE	2	0.6852	0.6486	0.6717	0.6615	0.6605
	5	0.6227	0.6561	0.6332	0.6457	0.6373
	15	0.6551	0.6138	0.6561	0.6121	0.6340
	35	0.6446	0.6162	0.6473	0.6126	0.6305
DA	2	0.6008	0.5492	0.5803	0.5704	0.5666
	5	0.5707	0.5872	0.5707	0.5867	0.5765
	15	0.6074	0.5346	0.5775	0.5650	0.5706
	35	0.5928	0.5380	0.5715	0.5593	0.5654

Table B.4: Coefficient bias with MI method = AREG ($n = 910$).

Missing %	Variable	AB		Her2		PR	
		Mean (%)	SE (%)	Mean (%)	SE (%)	Mean (%)	SE (%)
2	marker	-1.52	8.36	-1.31	6.66	-2.52	6.68
	AGE	1.21	5.65	-0.06	3.4	-0.88	2.87
	SYS2	0.28	0.92	0.12	1.07	-0.23	0.69
	SYS3	0.3	1.93	0.19	2.1	0.28	1.21
	SYS4	0.92	3.26	1.19	5.15	-0.91	2.77
	GRADE2	0.17	0.5	0.08	0.62	0.08	0.82
	GRADE3	0.09	0.3	0.04	0.39	0.16	0.5
	ERPOSNE	2.05	12.19	3.25	13.12	-29.81	76.25
	LVNNE	0.12	0.74	0.54	1.64	0.06	0.82
	PPNODE2	-0.03	0.6	0.21	0.99	0.01	0.59
	PPNODE3	0.05	0.36	-0.02	0.65	-0.08	0.42
	SIZE	-0.11	0.35	0	0.38	0.07	0.27
5	marker	-3.66	12.85	-4.99	9.53	-4.32	10.66
	AGE	2.25	8.43	-1.34	4.9	-1.56	4.22
	SYS2	0.58	1.27	-0.17	1.47	-0.43	1.19
	SYS3	0.69	3.26	-0.48	2.74	0.39	2.24
	SYS4	1.81	5.15	2.84	7.39	-1.02	3.97
	GRADE2	0.23	0.83	0.29	0.94	-0.07	1.36
	GRADE3	0.17	0.46	0.22	0.59	0.19	0.78
	ERPOSNE	4.42	21.11	12.5	17.93	-50.77	119.43
	LVNNE	0.12	1.07	1.45	2.24	0.37	1.44
	PPNODE2	-0.16	0.79	0.68	1.55	-0.18	0.93
	PPNODE3	0.14	0.57	-0.29	0.87	-0.18	0.74
	SIZE	-0.21	0.48	-0.02	0.58	0.02	0.48
15	marker	-12.78	23.67	-13.87	15.3	-13.16	20.22
	AGE	8.39	13.85	-3.2	6.67	-4.16	8.05
	SYS2	2.03	1.91	-0.2	1.98	-0.99	1.87
	SYS3	3.19	5.12	-0.69	3.62	1.01	3.35
	SYS4	6.41	8.14	7.09	10.61	-3.09	7.14
	GRADE2	1.08	1.36	0.88	1.28	-0.19	2.29
	GRADE3	0.72	0.84	0.59	0.98	0.42	1.48
	ERPOSNE	15.07	37.8	30.78	29.94	-157.93	230.56
	LVNNE	0.37	1.76	3.64	3.26	0.4	1.93
	PPNODE2	-0.61	1.16	1.94	2.19	0.01	1.38
	PPNODE3	0.37	0.91	-0.58	1.31	-0.46	1.07
	SIZE	-0.42	0.89	-0.07	0.84	0.32	0.86
35	marker	-36.19	28.79	-31.37	20.6	-35.01	27.38
	AGE	18.96	18.53	-8.41	10.75	-10.71	10.79
	SYS2	4.49	2.27	0.03	2.4	-2.4	2.23
	SYS3	7	5.34	-1.7	5.89	3.17	3.92
	SYS4	15.72	8.82	14.26	12.48	-6.81	8.49
	GRADE2	2.52	1.24	1.62	1.78	-0.38	3.02
	GRADE3	1.65	0.71	1.23	1.23	1.24	2.14
	ERPOSNE	48.34	44.64	65.27	40.24	-416.48	315.57
	LVNNE	0.69	1.9	7.46	4.11	1.13	2.47
	PPNODE2	-0.71	1.45	3.7	3.49	0.79	2.13
	PPNODE3	0.75	1.12	-1.1	1.84	-1.28	1.44
	SIZE	-1.06	0.92	-0.06	1.18	0.77	1.09

Table B.5: Coefficient bias with MI method = MICE ($n = 910$).

Missing %	Variable	AB		Her2		PR	
		Mean (%)	SE (%)	Mean (%)	SE (%)	Mean (%)	SE (%)
2	marker	-2.23	11.84	-4	7.86	-1.86	8.99
	AGE	0.07	8.44	-1.1	4	-1.04	3.43
	SYS2	0.26	1.24	0.16	1.14	-0.15	0.96
	SYS3	0.19	2.72	-0.03	2.47	-0.13	1.75
	SYS4	0.89	4.44	1.37	6.73	-0.98	3.28
	GRADE2	0.2	0.81	0.32	0.92	-0.08	1.18
	GRADE3	0.14	0.47	0.18	0.64	0.04	0.74
	ERPOSNE	2.32	16.79	8.28	15.23	-23.22	104.87
	LVNNE	0.11	1	0.89	2.02	-0.13	0.84
	PPNODE2	0.05	0.75	0.41	1.12	-0.04	0.86
	PPNODE3	0.02	0.53	-0.13	0.64	-0.09	0.64
	SIZE	-0.17	0.47	0.03	0.52	0.08	0.39
5	marker	-3.66	17.45	-5.11	12.44	-3.04	13.47
	AGE	0.1	12.56	-0.98	6.22	-1.27	5.37
	SYS2	0.82	1.93	-0.14	1.86	-0.09	1.51
	SYS3	0.42	4.41	-0.3	3.74	0.19	2.44
	SYS4	1.92	6.75	2.33	8.5	-1.67	5.53
	GRADE2	0.31	1.31	0.37	1.13	-0.01	1.74
	GRADE3	0.12	0.76	0.15	0.69	0.12	1.02
	ERPOSNE	5.42	25.14	10.91	23.97	-37.95	153.86
	LVNNE	0.36	1.36	1.7	3.34	0.05	1.57
	PPNODE2	-0.13	1.13	0.92	1.82	0.11	1.24
	PPNODE3	0.22	0.9	-0.27	1.18	-0.2	0.87
	SIZE	-0.2	0.66	0.02	0.83	0.08	0.63
15	marker	-17.17	30.98	-16.59	20.73	-13.86	22.74
	AGE	6.35	23.93	-4.27	11.71	-4.74	8.2
	SYS2	2.34	2.77	-0.43	2.73	-0.84	2.34
	SYS3	3.26	6.86	-1.35	5.7	0.37	4.36
	SYS4	7.69	11.59	6.94	15.03	-4.54	9.34
	GRADE2	3.03	4.82	1.93	4.49	-0.02	2.75
	GRADE3	1.62	2.52	1.1	2.4	0.56	1.67
	ERPOSNE	23.59	46.06	32.59	41.51	-159.89	269.86
	LVNNE	0.59	2.65	4.05	5.22	0.29	2.52
	PPNODE2	-0.48	1.59	1.63	2.75	0.08	1.64
	PPNODE3	0.44	1.15	-0.58	1.64	-0.57	1.31
	SIZE	-0.58	1.11	-0.15	1.25	0.14	1.13
35	marker	-41.71	36.2	-36.04	28.97	-34.21	33.31
	AGE	17.36	31.96	-10.46	12.85	-9.84	13.37
	SYS2	4.97	2.56	0.07	4.34	-2.63	3.36
	SYS3	7.27	8.71	-2.92	7.86	2.47	6.02
	SYS4	17.34	15.09	14.57	20.97	-9.32	11.92
	GRADE2	4.28	5.22	6.36	11.52	-0.69	3.81
	GRADE3	2.54	2.7	3.52	5.71	1.04	2.49
	ERPOSNE	54.93	61.14	70.67	55.44	-412.06	385.79
	LVNNE	1.41	3.34	8.22	7.15	1.03	3.05
	PPNODE2	-1.11	2.47	3.6	4.03	0.05	2.24
	PPNODE3	0.72	1.44	-1.35	2.46	-1.07	1.99
	SIZE	-1.31	1.05	0.06	1.74	0.36	1.79

Table B.6: Coefficient bias with MI method = DA ($n = 910$).

Missing %	Variable	AB		Her2		PR	
		Mean (%)	SE (%)	Mean (%)	SE (%)	Mean (%)	SE (%)
2	marker	-4.95	10.16	-2.69	6.34	-3.73	7.51
	AGE	2.12	7.67	-1.26	3.28	-1.69	2.84
	SYS2	0.43	1.08	-0.1	0.93	-0.28	0.95
	SYS3	0.66	2.56	-0.39	1.64	0.12	1.53
	SYS4	2.64	4.11	2.21	6.28	-0.53	3.12
	GRADE2	0.58	1.07	0.4	1.56	0.07	0.94
	GRADE3	0.37	0.59	0.24	0.85	0.21	0.6
	ERPOSNE	7.35	16.32	7.11	11.74	-49.53	88.48
	LVNNE	0.18	0.96	0.76	1.75	0.1	0.91
	PPNODE2	-0.07	0.63	0.27	1.12	-0.02	0.69
	PPNODE3	-0.03	0.44	-0.13	0.63	-0.11	0.5
	SIZE	-0.09	0.4	0.12	0.43	0.08	0.3
5	marker	-10.8	16.97	-8	10.36	-7.14	10.77
	AGE	6.44	10.26	-3.11	5.17	-2.92	4.85
	SYS2	0.97	1.69	0.05	1.6	-0.79	1.07
	SYS3	1.94	3.94	-0.55	2.38	0.2	1.98
	SYS4	5.76	6.98	7.08	8.58	-0.86	4.44
	GRADE2	1.15	1.56	1.21	2.03	-0.09	1.25
	GRADE3	0.72	0.85	0.81	1.22	0.33	0.85
	ERPOSNE	18.5	25.49	17.58	20.18	-104.37	120.41
	LVNNE	0.07	1.25	2.26	2.72	-0.05	1.27
	PPNODE2	-0.32	0.9	1.04	1.22	0.23	1.01
	PPNODE3	0.17	0.6	-0.34	0.87	-0.23	0.77
	SIZE	-0.21	0.61	-0.1	0.68	0.14	0.46
15	marker	-27.67	20.23	-18.55	15.75	-21.3	16.04
	AGE	10.98	17.06	-8.25	9.13	-9.14	5.86
	SYS2	2.77	1.74	0.05	2.25	-1.74	1.61
	SYS3	4.75	4.46	-1.88	4.76	1.05	3.43
	SYS4	14.86	8.17	17.65	13.15	-2.04	6.71
	GRADE2	2.55	1.65	2.43	2.52	0.28	1.9
	GRADE3	1.61	0.92	1.74	1.39	1.26	1.15
	ERPOSNE	44.64	30.15	44.66	27.99	-305.63	171.75
	LVNNE	0.71	1.94	4.88	4.05	0	1.87
	PPNODE2	-0.46	1.27	3.09	1.86	0.46	1.31
	PPNODE3	0.28	0.92	-1.17	1.3	-0.6	0.93
	SIZE	-0.61	0.84	-0.12	0.88	0.39	0.73
35	marker	-55.5	29.22	-40.63	18.33	-49.96	22.39
	AGE	25.38	16.92	-16.61	8.58	-19.42	7.2
	SYS2	5.12	2.36	-0.04	2.57	-3.81	2.16
	SYS3	9.64	5.53	-2.96	4.89	3.52	3.75
	SYS4	24.81	8.6	29.58	13.53	-6.36	7.52
	GRADE2	4.38	2	4.13	2.85	0.68	2.25
	GRADE3	2.9	1.02	3.15	1.61	2.6	1.45
	ERPOSNE	88.92	36.96	89.39	29.52	-679.04	216.7
	LVNNE	0.85	1.84	10.06	3.97	0.34	2.31
	PPNODE2	-0.81	1.37	5.28	2.76	1.03	1.28
	PPNODE3	0.46	1.23	-1.87	1.45	-1.34	1.16
	SIZE	-1.04	1.04	0.1	1.06	0.98	1.01

Appendix C

Number of outcomes for all categories of each variable

Table C.1: Number of outcomes for all categories of each variable with the n=1575 dataset.

n=1575	BCSS	LRS	DRS
AGECAT=1	171	45	182
AGECAT=2	305	93	336
SYS=1	158	62	170
SYS=2	162	39	180
SYS=3	118	33	127
SYS=4	38	4	41
GRADECAT=1	9	4	8
GRADECAT=2	169	44	191
GRADECAT=3	298	90	319
ERPOSNE=1	119	38	130
ERPOSNE=2	357	100	388
LVNNE=1	209	75	221
LVNNE=2	267	63	297
PPNODECAT=1	117	35	137
PPNODECAT=2	166	29	178
PPNODECAT=3	193	74	203
SIZECAT=1	189	81	210
SIZECAT=2	287	57	308
AB=1	419	121	459
AB=2	57	17	59
BCL2=1	236	64	256
BCL2=2	240	74	262
CK56=1	430	128	472
CK56=2	46	10	46
GATA=1	362	106	393
GATA=2	114	32	125
Her1=1	398	117	437
Her1=2	78	21	81
ERS=1	162	44	172
ERS=2	314	94	346
CA9=1	381	116	415
CA9=2	95	22	103
P53=1	359	109	388
P53=2	117	29	130
Her2=1	390	119	425
Her2=2	86	19	93
IGFB=1	295	101	321
IGFB=2	181	37	197
PR=1	253	76	275
PR=2	223	62	243
YB1=1	263	72	286
YB1=2	214	66	232

Table C.2: Number of outcomes for all categories of each variable with the n=910 dataset.

n=910	BCSS	LRS	DRS
AGECAT=1	112	25	118
AGECAT=2	176	58	202
SYS=1	94	34	102
SYS=2	87	24	103
SYS=3	81	21	88
SYS=4	26	4	27
GRADECAT=1	5	3	4
GRADECAT=2	93	18	108
GRADECAT=3	190	62	208
ERPOSNE=1	71	23	81
ERPOSNE=2	217	60	239
LVNNE=1	122	40	130
LVNNE=2	166	43	190
PPNODECAT=1	77	23	92
PPNODECAT=2	99	19	109
PPNODECAT=3	112	41	119
SIZECAT=1	105	45	123
SIZECAT=2	183	38	197
AB=1	246	70	278
AB=2	42	13	42
BCL2=1	146	40	161
BCL2=2	142	43	159
CK56=1	255	77	287
CK56=2	33	6	33
GATA=1	206	63	228
GATA=2	82	20	92
Her1=1	242	71	271
Her1=2	46	12	49
ERS=1	85	23	94
ERS=2	203	60	226
CA9=1	223	65	249
CA9=2	65	18	71
P53=1	201	60	222
P53=2	87	23	98
Her2=1	226	67	251
Her2=2	62	16	69
IGFB=1	177	59	198
IGFB=2	111	24	122
PR=1	147	42	164
PR=2	141	41	156
YB1=1	137	34	156
YB1=2	151	49	164

Appendix D

Script Files

D.1 R Program of Multiple Imputation

D.1.1 R Program of AREG and MICE

```
#############################################################################
#
# AREG OR MICE METHOD for MI
#
# This code removes all observations with missing clinical observations(n=1575)
#
# Coxph or Survreg method for survival model
#
# if change to use another marker, we need to change 8 in ~markers (i.e. ~AB ~PR),
# 1 in MICE "marker". Also change Imarker correspondingly
#############################################################################
#
# using one marker only (AB, PR, Her2)
# generate random missing and use 1575 obs only (no missing patterns considered)
# use coxph only for MI comparison
# add RMSE ???
#############################################################################

# determine which marker is used in the model
# Imarker =1 for AB, =9 for Her2, =11 for PR
```

```
Imarker = 1 # AB,BCL2,CK56,GATA,Her1,ERS,CA9,P53,Her2,IGFB,PR,YB1



#-----------------------------------------
# Flag of choosing Coxph or Survreg

flg =0  # Coxph
# flg =1 # Survreg

# Flag of choosing MI method MICE or AREG

# mthd =0  # MICE
mthd =1 # Areg
#-----------------------------------------

GPEC<-load("E:\\s-project\\liu\\imputBree")
#GPEC.df, Mark.m.df, Clin.m.df

dim (GPEC.df)

m.vars<-c("AB","BCL2","CK56","GATA","Her1","ERS","CA9","P53","Her2","IGFB","PR","YB1")
# X is missing

c.vars<-c("AGECAT","HISTCA","GRADECAT","ERPOSNE","LVNNE","SYS","PPNODECAT","SIZECAT")
#unknown or category 9 are missing. add SYS in and remove PNODECAT

GPEC.cm.df <- GPEC.df[, c(m.vars,c.vars)]
# choose only clinical variables and markers

summary (GPEC.cm.df)


##remove the cases with missing clinical variables

miss<-(GPEC.df$GRADECAT=="9" | GPEC.df$ERPOSNE=="unknown " | GPEC.df$LVNNE=="9" |
  GPEC.df$SYS=="unknown" | GPEC.df$PPNODECAT=="unknown" | GPEC.df$SIZECAT=="unknown")
sum(miss==T)
GPEC.NCM.df<-GPEC.df[miss==FALSE,]
dim(GPEC.NCM.df)

#Combine M and C
try<-GPEC.NCM.df[,c(m.vars,c.vars)] # 20 columns

dim(try)
summary(try)
```

```
##set the levels of the variables in j
# all levels labeled c(1,2,NA) were originally 0,1,NA

for(i in 1:ncol(try)){
   if( i ==13 | i ==14)
       levels(try[,i])<-c(1, 2) # two levels for AGECAT and HISTCA
   else if (i ==15 | i ==19)
levels(try[,i])<-c(1, 2, 3, 9) # 4 levels for GRADECAT and PPNODECAT
   else
 levels(try[,i])<-c(1, 2, 9) # 3 levels for others
   try[,i]<-as.numeric(try[,i])
}

try_numeric<-as.matrix(try) # keep the original one
summary(try_numeric)

# get categorical data
for(i in 1:ncol(try)){
    try[,i]<-as.factor(try[,i])
    if( i <13 )
 levels(try[,i])<-c(1,2, NA)
    else if (i ==15 |i ==19) # GRADECAT and PPNODECAT
 levels(try[,i])<-c(1,2,3)
    else if (i ==18) # SYS
 levels(try[,i])<-c(1,2,3,4)
    else
       levels(try[,i])<-c(1,2)
}
try_factor<-as.matrix(try) # keep the categorical one
summary(try)  # 20 variables

library (mice)

library(cat)
rngseed(1898) #random generator must be set before using imputation functions


###  aregImputation METHOD to generate complete dataset #####
#####################################################

#Multiple Imputation using Bootstrap and PMM "aregImputation(Hmisc)"

#####################################################

library(Hmisc)
set.seed(200) # Function of this one?
dimnames(try)[[2]] #giving the column names
nnumber = 5 # use 5 unless MICE method uses other values
#Multiple Imputation using Bootstrap and PMM "aregImputation(Hmisc)"y
```

```
#a$imputed[8]

if (Imarker ==1) # AB
a<- aregImpute(~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=try)

if (Imarker ==9) # Her2
a<- aregImpute(~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=try)

if (Imarker ==11) # PR
a<- aregImpute(~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=try)

i=Imarker
    try[,i]<-as.numeric(try[,i])
    try[which(is.na(try[,i])),i]<-round(apply(data.frame(a$imputed[1]),1,mean))
#AB starts from 1st column in "a"

  try[,i]<-as.factor(try[,i])
summary (try)

#m_try for calculating sensitivity etc

if (Imarker ==1) # AB
{
m1.vars<-c("AB")
m_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==9) # Her2
{
m1.vars<-c("Her2")
m_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==11) # PR
{
m1.vars<-c("PR")
m_try <- try[, c(m1.vars,c.vars)]
}


#dim(a$imputed[1])
#Parametric survival analysis "survreg"

library(survival)

full<-cbind(try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(full)<-c(colnames(try),"SURVYRS","BRDEATH")
summary (full)
```

```
if (flg == 1)    # one marker model using survreg
{
# use PPNODECAT and discard PNODECAT because they are highly corrolated
if (Imarker ==1)
model<-survreg(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=full)
if (Imarker ==9)
model<-survreg(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=full)
if (Imarker ==11)
model<-survreg(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=full)

}

if (flg == 0)
{
if (Imarker ==1)
model<-coxph(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=full)
if (Imarker ==9)
model<-coxph(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=full)
if (Imarker ==11)
model<-coxph(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=full)
}
model

# fitted - observed
#exp(model$y)[which(full$BRDEATH==1)]-full$SURVYRS[which(full$BRDEATH==1)]

True_coef_factor <- model$coef  # store true coefficients

# --- add 10 year survival probability ------

TR_pred_p10cox<-rep(0, nrow(full))
TR_pred_p10coxse<-rep(0, nrow(full))

for (i in 1:nrow(full)){
 fit<-survfit(model,newdata=full[i, ])
 jj<-summary(fit,c(10))
 TR_pred_p10cox[i]<-jj$surv
 TR_pred_p10coxse[i]<-jj$std.er
}
summary(TR_pred_p10cox)
```

```
# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er

#--------------------------------------
# Set initial
pmiss <-c(0.02, 0.05, 0.15, 0.35) # missing proportions
Np<-rep(0, 29)
nsampling=100           # number of sampling
nimpute=5               # number of imputation

#coefficients=25
##start a new big big big loop
##sam : number of different sample sizes
##i: number of sampling
##j: number of imputation
##k: number of marks (temp short loops)

for(sam in 1: 4){     # 4 proportions, change if necessary

# DA_rmse_factor<-c(0)
# DA_rmse_numeric<-c(0)

result_sen<-matrix(0, nimpute, nsampling)
result_spe<-matrix(0, nimpute, nsampling)
result_ppv<-matrix(0, nimpute, nsampling)
result_npv<-matrix(0, nimpute, nsampling)
      result_percent<-matrix(0, nimpute, nsampling)

fit_impu<-matrix(0, nimpute, nsampling)


      total_missing<-rep(0, nsampling)

A_coef_factor<-vector("list",nsampling)
A_pred_factor<-vector("list",nsampling)
A_coef<-vector("list",nimpute)
A_pred<-vector("list",nimpute)

 rmse_factor<-c(0)
#  coef_bias <- rep(0, length(model$coef))
coef_bias <-matrix(0, length(model$coef), nsampling)
rownames(coef_bias)<-rownames(data.matrix(model$coef)) #variable names in model
d_print <-matrix(0, length(model$coef), 2)  # print mean and SD
rownames(d_print)<-rownames(data.matrix(model$coef)) #variable names in model
```

```
for( i in 1: nsampling){


#---------------------------------------------------------------
#Randomly select ..% of the data and set it to missing
#change proportion in 'x' to get different percentage missing
#---------------------------------------------------------------

if (Imarker ==1) # AB
{
m1.vars<-c("AB")
sample_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==9) # Her2
{
m1.vars<-c("Her2")
sample_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==11) # PR
{
m1.vars<-c("PR")
sample_try <- try[, c(m1.vars,c.vars)]
}

x<-sample(1:1575,ceiling(pmiss[sam]*1575)) # 1575 rows and missing prop

if (Imarker ==1) # AB
sample_try$AB[x]<-NA
if (Imarker ==9) # Her2
sample_try$Her2[x]<-NA
if (Imarker ==11) # PR
sample_try$PR[x]<-NA
summary (sample_try)

#True_coef_factor <- model0$coef  # store true coefficients

#################################################################
# doing the imputation using MICE


if (Imarker ==1) # AB
mice_sample<-sample_try[,c("AB", c.vars)] #using one marker and all clinical variables
if (Imarker ==9) # Her2
mice_sample<-sample_try[,c("Her2", c.vars)]#using one marker and all clinical variables
if (Imarker ==11) # PR
mice_sample<-sample_try[,c("PR", c.vars)] #using one marker and all clinical variables

imp <- mice(mice_sample) # by default, 5 times imputation
```

```
complete (imp, 1)  # show the first complete data

summary (complete (imp))

#################################################################
# doing the imputation using areg

if (Imarker ==1) # AB
a<- aregImpute(~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=sample_try)

if (Imarker ==9) # Her2
a<- aregImpute(~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=sample_try)

if (Imarker ==11) # PR
a<- aregImpute(~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=sample_try)

#---------------------------------------------------
A_coef_mean<-rep(0, length(model$coef))
A_pred_mean<-rep(0, 1575)

     for( j in 1:nimpute){ #nimpute must equual to 5

temp_try<-sample_try
marker.imputations<-vector("list",nimpute)

if(mthd == 1) # AREG
{

    k=1  # one marker is used,
       temp_try[which(is.na(temp_try[,k])),k]<-a$imputed[[1]][,j]
#the marker starts from 1st column in "a"
}
if(mthd == 0) #MICE
{
temp_try <- complete(imp, nimpute)
}
summary(temp_try)

SamImputed<-cbind(temp_try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(SamImputed)<-c(colnames(temp_try),"SURVYRS","BRDEATH")
summary (SamImputed)
marker.imputations[[j]] <- SamImputed  # for calculating sensitivity etc


if (flg == 1)    # one marker model using survreg
{
```

```
# use PPNODECAT and discard PNODECAT because they are highly corrolated
if (Imarker ==1)
model1<-survreg(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=SamImputed)
if (Imarker ==9)
model1<-survreg(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=SamImputed)
if (Imarker ==11)
model1<-survreg(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=SamImputed)

}

if (flg == 0)
{
if (Imarker ==1)
model1<-coxph(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=SamImputed)
if (Imarker ==9)
model1<-coxph(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=SamImputed)
if (Imarker ==11)
model1<-coxph(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=SamImputed)
}
model1

# fitted(predicted) value for 10 year survival probability
fit_impu[j,i]<-summary(survfit(model1), c(10))$surv

# length(model1$y[,1])
# fitted - observed ???
# exp(model1$y)[,1]-SamImputed$SURVYRS

A_coef[[j]] <-data.frame(model1$coef)

# A_pred[[j]] <-data.frame(predict(model1,type="response"))
# A_pred[[j]]<-exp(model1$y)[, 1]

A_coef_mean<- A_coef_mean +A_coef[[j]]/nimpute
A_pred_mean<- A_pred_mean +A_pred[[j]]/nimpute

     } # end of nimpute

total_missing[i]<-length(which(is.na(sample_try))) # total missing number

for(j in 1:nimpute){
pp<-c(0)
nn<-c(0)
```

```
pn<-c(0)
np<-c(0)
  for (k in 1:1){ ## 1 marker now
                    agreed<-c(which(as.numeric(m_try[,k][where=is.na(sample_try[,k])])-
as.numeric(SamImputed[,k][where=is.na(sample_try[,k])])==0))
                    nn<-c(nn+length(which(SamImputed[agreed,k]==1)))
                    pp<-c(pp+length(which(SamImputed[agreed,k]==2)))
       pn<-c(pn+length(which((as.numeric(m_try[,k][where=is.na(sample_try[,k])])
-as.numeric(SamImputed[,k][where=is.na(sample_try[,k])]))==1)))
                    np<-c(np+length(which((as.numeric(m_try[,k][where=is.na(sample_try[,k])])
-as.numeric(SamImputed[,k][where=is.na(sample_try[,k])]))==-1)))
             } ## end of markers

# Calculate sensitivity, specificity, positive predictive
# value, and negative predictive value
# kappa is the cohen's kappa, calculated by (#real-#expected)/(total - #expected)


  result_percent[j, i]<-c((pp+nn)/total_missing[i])
             nreal=pp+nn
ntotal=pp+nn+pn+np
nexpected=(pp+pn)*(pp+np)/ntotal +(nn+pn)*(nn+pn)/ntotal
# result1_ka[j,i]=(nreal-nexpected)/(ntotal-nexpected)
result_sen[j,i]=pp/(pp+np)
result_spe[j,i]=nn/(nn+pn)
result_ppv[j,i]=pp/(pp+pn)
result_npv[j,i]=nn/(nn+np)
            } ##end of nimpute

# A_coef_factor[[i]]<-A_coef_mean
# A_pred_factor[[i]]<-A_pred_mean

# RMSE calculation: rmse= (1/nsampling)*sum{sqrt[sum (predict - true)^2/nsize]}
# Bias for coefficients: (1/nsampling)*sum(A_coef-True_coef)/abs(True_coef)

A_coef_factor[[i]] <- (A_coef_mean-True_coef_factor)/True_coef_factor
coef_bias [, i] <- data.matrix (A_coef_factor[[i]])

    } ## end of sampling

# Writing out to the files
# Important: change the file name if you want to save the results
# separately. Otherwise the results will be APPENDED to the present
# file because we use the option "append=TRUE")

d_print [, 1] <-data.matrix(apply(coef_bias, 1, mean))
d_print [, 2] <-data.matrix(apply(coef_bias, 1, sd))

if (mthd ==0) # MICE
```

```
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB", "MICE Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_mice_one_AB", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_mice_one_AB",append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "Marker = Her2", "MICE Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_mice_one_Her2", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_mice_one_Her2",append=TRUE)
}
if (Imarker ==11) # PR
{
cat( "\n","\n", "Marker = PR", "MICE Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_mice_one_PR", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_mice_one_PR",append=TRUE)
}

}

if (mthd ==1) # AREG
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB", "AREG Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_areg_one_AB", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_areg_one_AB",append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "Marker = Her2", "AREG Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_areg_one_Her2", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_areg_one_Her2",append=TRUE)
}
if (Imarker ==11) # PR
```

```
{
cat( "\n","\n", "Marker = PR", "AREG Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_areg_one_PR", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_areg_one_PR",append=TRUE)
}

}

# Print out sensitivity etc

if (mthd ==0) # MICE
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB","MICE Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Mean  ", mean(apply(result_sen, 2, mean)),
mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe,2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{

cat( "\n","\n", "Marker = Her2","MICE Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Mean  ", mean(apply(result_sen, 2, mean)),
mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe,2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_Her2", append=TRUE)
}
```

```
if (Imarker ==11) # PR
{

cat( "\n","\n", "Marker = PR","MICE Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Mean  ", mean(apply(result_sen, 2, mean)),
mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe,2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_PR", append=TRUE)
}
}

if (mthd ==1) # AREG
{
if (Imarker ==1) # AB
{

cat( "\n","\n", "Marker = AB","AREG Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Mean  ", mean(apply(result_sen, 2, mean)),
mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe,2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_areg_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{

cat( "\n","\n", "Marker = Her2","AREG Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n", "Mean  ", mean(apply(result_sen, 2, mean)),
 mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe,2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
```

```
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_areg_one_Her2", append=TRUE)
}

if (Imarker ==11) # PR
{

cat( "\n","\n", "Marker = PR","AREG Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Mean  ", mean(apply(result_sen, 2, mean)),
mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe,2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_areg_one_PR", append=TRUE)
}
}

# print out 10 year survival probability

if (mthd ==0) # MICE
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB","MICE Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survival:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "Marker = Her2","MICE Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survival:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_Her2", append=TRUE)
}
```

```
if (Imarker ==11) # PR
{
cat( "\n","\n", "Marker = PR","MICE Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survival:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_PR", append=TRUE)
}
}

if (mthd ==1) # AREG
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB","AREG Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survival:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_AREG_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "Marker = Her2","AREG Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survival:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_AREG_one_Her2", append=TRUE)
}

if (Imarker ==11) # PR
{
cat( "\n","\n", "Marker = PR","AREG Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survival:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_AREG_one_PR", append=TRUE)
}
}


} ## end of missing proportion size

if (mthd ==0) # MICE
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "True Results","\n",
```

```
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_Her2", append=TRUE)
}

if (Imarker ==11) # PR
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_mice_one_PR", append=TRUE)
}
}

if (mthd ==1) # AREG
{
if (Imarker ==1) # AB
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_AREG_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_AREG_one_Her2", append=TRUE)
}

if (Imarker ==11) # PR
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_AREG_one_PR", append=TRUE)
}
}
```

D.1.2 R Program of DA

```
#############################################################################
#
#          DA METHOD
#

# This code removes all observations with missing clinical observations(n=1575)
# and runs a Cox model when you remove the missing observations
# for each specific marker and then when you remove all of the observations
# with any missing at all. The output from this code is used in all of the
# tables.
#
# using one marker only (AB, PR, Her2)
# generate random missing and use 1575 obs only (no missing patterns considered)
# use coxph only for MI comparison
#
#############################################################################

# determine which marker is used in the model
# Imarker =1 for AB, =9 for Her2, =11 for PR


Imarker = 1 # AB,BCL2,CK56,GATA,Her1,ERS,CA9,P53,Her2,IGFB,PR,YB1

#-----------------------------------------
# Flag of choosing Coxph or Survreg

flg =0  # Coxph
# flg =1 # Survreg
#-----------------------------------------

GPEC<-load("E:\\s-project\\liu\\imputBree")
#GPEC.df, Mark.m.df, Clin.m.df

dim (GPEC.df)

m.vars<-c("AB","BCL2","CK56","GATA","Her1","ERS","CA9","P53","Her2","IGFB","PR","YB1")
# X is missing

c.vars<-c("AGECAT","HISTCA","GRADECAT","ERPOSNE","LVNNE","SYS","PPNODECAT","SIZECAT")
#unknown or category 9 are missing. add SYS in and remove PNODECAT

GPEC.cm.df <- GPEC.df[, c(m.vars,c.vars)]
# choose only clinical variables and markers

summary (GPEC.cm.df)
```

```
##remove the cases with missing clinical variables

miss<-(GPEC.df$GRADECAT=="9" | GPEC.df$ERPOSNE=="unknown " | GPEC.df$LVNNE=="9" |
  GPEC.df$SYS=="unknown" | GPEC.df$PPNODECAT=="unknown" | GPEC.df$SIZECAT=="unknown")
sum(miss==T)
GPEC.NCM.df<-GPEC.df[miss==FALSE,]
dim(GPEC.NCM.df)

# remove the column 'PNODECAT'(#6) because it is highly correlated with 'PPNODECAT'
try<-GPEC.NCM.df[,c(m.vars,c.vars)]  # 20 columns

dim(try)
summary(try)

##set the levels of the variables in j
# all levels labeled c(1,2,NA) were originally 0,1,NA

for(i in 1:ncol(try)){
   if( i ==13 | i ==14)
       levels(try[,i])<-c(1, 2) # two levels for AGECAT and HISTCA
   else if (i ==15 | i ==19)
levels(try[,i])<-c(1, 2, 3, 9) # 4 levels for GRADECAT and PPNODECAT
   else
 levels(try[,i])<-c(1, 2, 9) # 3 levels for others
   try[,i]<-as.numeric(try[,i])
}

try_numeric<-as.matrix(try) # keep the original one
summary(try_numeric)

# get categorical data
for(i in 1:ncol(try)){
    try[,i]<-as.factor(try[,i])
    if( i <13 )
 levels(try[,i])<-c(1,2, NA)
    else if (i ==15 |i ==19) # GRADECAT and PPNODECAT
 levels(try[,i])<-c(1,2,3)
    else if (i ==18) # SYS
 levels(try[,i])<-c(1,2,3,4)
    else
       levels(try[,i])<-c(1,2)
}
try_factor<-as.matrix(try) # keep the categorical one
summary(try)  # 20 variables


library(cat)
rngseed(1898) #random generator must be set before using imputation functions
```

```
###  aregImputation METHOD to generate complete dataset #####
#####################################################

#Multiple Imputation using Bootstrap and PMM "aregImputation(Hmisc)"

#####################################################

library(Hmisc)

set.seed(200) # Function of this one?
dimnames(try)[[2]]
nnumber = 5   # imputation
#Multiple Imputation using Bootstrap and PMM "aregImputation(Hmisc)"y
#a$imputed[8]

if (Imarker ==1) # AB
a<- aregImpute(~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=try)

if (Imarker ==9) # Her2
a<- aregImpute(~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=try)

if (Imarker ==11) # PR
a<- aregImpute(~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
,n.impute=nnumber, nk=3, data=try)

i=Imarker
    try[,i]<-as.numeric(try[,i])
    try[which(is.na(try[,i])),i]<-round(apply(data.frame(a$imputed[1]),1,mean))
#marker starts from 1st column in "a"

  try[,i]<-as.factor(try[,i])

summary (try)

if (Imarker ==1) # AB
{
m1.vars<-c("AB")
m_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==9) # Her2
{
m1.vars<-c("Her2")
m_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==11) # PR
{
```

```
m1.vars<-c("PR")
m_try <- try[, c(m1.vars,c.vars)]
}

#survival analysis

library(survival)

full<-cbind(try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(full)<-c(colnames(try),"SURVYRS","BRDEATH")
summary (full)

if (flg == 1)    # one marker model using survreg
{
# use PPNODECAT and discard PNODECAT because they are highly corrolated
if (Imarker ==1)
model<-survreg(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=full)
if (Imarker ==9)
model<-survreg(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=full)
if (Imarker ==11)
model<-survreg(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=full)

}

if (flg == 0) #coxph
{
if (Imarker ==1) #AB
model<-coxph(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=full)
if (Imarker ==9)
model<-coxph(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=full)
if (Imarker ==11)
model<-coxph(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=full)
}
model


# --- add 10 year survival probability ------

TR_pred_p10cox<-rep(0, nrow(full))
TR_pred_p10coxse<-rep(0, nrow(full))

for (i in 1:nrow(full)){
 fit<-survfit(model,newdata=full[i, ])
```

```
 jj<-summary(fit,c(10))
 TR_pred_p10cox[i]<-jj$surv
 TR_pred_p10coxse[i]<-jj$std.er
}
summary(TR_pred_p10cox)

fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er

# fitted - observed
#exp(model$y)[which(full$BRDEATH==1)]-full$SURVYRS[which(full$BRDEATH==1)]

True_coef_factor <- model$coef  # store true coefficients using factors
#True_predict_factor<-predict(model,type="response")

# Set initial
pmiss <-c(0.02, 0.05, 0.15, 0.35) # missing proportions
Np<-rep(0, 29)
nsampling=100           # number of sampling
nimpute=5                # number of imputation

#-------------------------------------------------------------------------
# m sets the margins.
# Seems that m does not have influence on the imputation results
#m<-c(1,0,2,0,3,0,4,0,5,0,6,0,7,0,8,0,9,0,10)
m<-c(1,2,3,4,5)   # missing only for the marker

for(sam in 1: 4){     # 4 proportions, change if necessary


# DA_rmse_factor<-c(0)
# DA_rmse_numeric<-c(0)

result_sen<-matrix(0, nimpute, nsampling)
result_spe<-matrix(0, nimpute, nsampling)
result_ppv<-matrix(0, nimpute, nsampling)
result_npv<-matrix(0, nimpute, nsampling)
      result_ka<-matrix(0, nimpute, nsampling)
      result_percent<-matrix(0, nimpute, nsampling)

fit_impu<-matrix(0, nimpute, nsampling)

      total_missing<-rep(0, nsampling)

DA_coef_factor<-vector("list",nsampling)
DA_pred_factor<-vector("list",nsampling)
DA_coef_f<-vector("list",nimpute)
DA_pred_f<-vector("list",nimpute)
```

```
 rmse_factor<-c(0)
#  coef_bias <- rep(0, length(model$coef))
coef_bias <-matrix(0, length(model$coef), nsampling)
rownames(coef_bias)<-rownames(data.matrix(model$coef))
#variable names in the model
d_print <-matrix(0, length(model$coef), 2) # print mean and SD
rownames(d_print)<-rownames(data.matrix(model$coef))
#variable names in the model

   for( i in 1: nsampling){
#True_coef_factor <- model0$coef  # store true coefficients

#-----------------------------------------------------------------
#Randomly select ..% of the data and set it to missing
#change proportion in 'x' to get different percentage missing
#-----------------------------------------------------------------

if (Imarker ==1) # AB
{
m1.vars<-c("AB")
sample_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==9) # Her2
{
m1.vars<-c("Her2")
sample_try <- try[, c(m1.vars,c.vars)]
}
if (Imarker ==11) # PR
{
m1.vars<-c("PR")
sample_try <- try[, c(m1.vars,c.vars)]
}

x<-sample(1:1575,ceiling(pmiss[sam]*1575)) # 1575 rows and missing prop

if (Imarker ==1) # AB
sample_try$AB[x]<-NA
if (Imarker ==9) # Her2
sample_try$Her2[x]<-NA
if (Imarker ==11) # PR
sample_try$PR[x]<-NA
summary (sample_try)

#########################################################
## Doing the imputation using DA for all markers
#
marker.miss<-data.matrix(sample_try)
s.marker<-prelim.cat(marker.miss)
```

```
s.marker$nmis

#try the imputations under a desired prior

theta.marker<-ecm.cat(s.marker,margins=m,prior=1.1)

##Now (nimpute) multiple imputations of the missing variable are generated
##by running a chain and taking every 50th observation.

marker.imputations<-vector("list",nimpute)


rngseed(1575) #random generator must be set before imputation functions


DA_coef_mean<-rep(0, length(model$coef))
DA_pred_mean<-rep(0, 1575)


for(j in 1:nimpute){

#cat("Doing imputation",i,"\n")
    theta.h3<-da.cat(s.marker,theta.marker,steps=50, prior=1)
    marker.imputations[[j]]<-imp.cat(s.marker,theta.h3)
    dimnames(marker.imputations[[j]])<-dimnames(marker.miss)

#Bring in the SURVYRS and BRDEATH from the dataset
#Change back to dataframe for survdiff function
#SamImputed<-vector("list",nimpute)

# change marker.imputations to factor
marker.ftor=matrix(as.factor(marker.imputations[[j]]),
nrow(marker.imputations[[j]]), ncol(marker.imputations[[j]]))
            SamImputed<-cbind(data.frame(marker.ftor),
GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
            colnames(SamImputed)<-c(colnames(marker.imputations[[j]]),
"SURVYRS","BRDEATH")  #
     summary (as.data.frame(SamImputed))


if (flg == 1)    # one marker model using survreg
{
# use PPNODECAT and discard PNODECAT because they are highly corrolated
if (Imarker ==1)
model1<-survreg(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=SamImputed)
if (Imarker ==9)
model1<-survreg(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=SamImputed)
```

```
if (Imarker ==11)
model1<-survreg(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT
           ,y=TRUE,dist="loglogistic",data=SamImputed)

}

if (flg == 0)
{
if (Imarker ==1)
model1<-coxph(Surv(SURVYRS,BRDEATH==1)~AB+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=SamImputed)
if (Imarker ==9)
model1<-coxph(Surv(SURVYRS,BRDEATH==1)~Her2+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=SamImputed)
if (Imarker ==11)
model1<-coxph(Surv(SURVYRS,BRDEATH==1)~PR+AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT
           +SIZECAT,y=TRUE,model=TRUE,data=SamImputed)
}
model1

# fitted(predicted) value for 10 year survival probability
fit_impu[j,i]<-summary(survfit(model1), c(10))$surv

DA_coef_f[[j]] <-data.frame(model1$coef)

DA_coef_mean<- DA_coef_mean +DA_coef_f[[j]]/nimpute
# DA_pred_mean<- DA_pred_mean_f +DA_pred_f[[j]]/nimpute

} ## end nimpute


total_missing[i]<-length(which(is.na(sample_try))) #total missing num

for(j in 1:nimpute){
pp<-c(0)
nn<-c(0)
pn<-c(0)
np<-c(0)
  for (k in 1:1){ ##1 marker
        agreed<-c(which(as.numeric(m_try[,k][where=is.na(sample_try[,k])])-
as.numeric(marker.imputations[[j]][,k]
[where=is.na(sample_try[,k])])==0))
        nn<-c(nn+length(which(marker.imputations[[j]][agreed,k]==1)))
        pp<-c(pp+length(which(marker.imputations[[j]][agreed,k]==2)))
     pn<-c(pn+length(which((as.numeric(m_try[,k][where=is.na(sample_try[,k])])
-as.numeric(marker.imputations[[j]][,k]
[where=is.na(sample_try[,k])]))==1)))
        np<-c(np+length(which((as.numeric(m_try[,k][where=is.na(sample_try[,k])])
-as.numeric(marker.imputations[[j]][,k]
```

```
[where=is.na(sample_try[,k])]))==-1)))
             } ## end of markers

# Calculate sensitivity, specificity, positive predictive
# value, and negative predictive value
# kappa is the cohen's kappa, calculated by (#real-#expected)/(total - #expected)

  result_percent[j, i]<-c((pp+nn)/total_missing[i])
      nreal=pp+nn
ntotal=pp+nn+pn+np
nexpected=(pp+pn)*(pp+np)/ntotal +(nn+pn)*(nn+pn)/ntotal
result_ka[j,i]=(nreal-nexpected)/(ntotal-nexpected)
result_sen[j,i]<- c(pp/(pp+np))
result_spe[j,i]<- c(nn/(nn+pn))
result_ppv[j,i]<- c(pp/(pp+pn))
result_npv[j,i]<- c(nn/(nn+np))
            } ##end of nimpute

# DA_coef_factor[[i]]<-DA_coef_mean_f
# DA_pred_factor[[i]]<-DA_pred_mean_f

# RMSE calculation: rmse= (1/nsampling)*sum{sqrt[sum (predict - true)^2/nsize]}
# Bias for coefficients: (1/nsampling)*sum(A_coef-True_coef)/abs(True_coef)

DA_coef_factor[[i]] <- (DA_coef_mean-True_coef_factor)/True_coef_factor
coef_bias [, i] <- data.matrix (DA_coef_factor[[i]])

    } ## end of sampling

# Writing out to the files
# Important: change the file name if you want to save the results
# separately. Otherwise the results will be APPENDED to the present
# file because we use the option "append=TRUE")

d_print [, 1] <-data.matrix(apply(coef_bias, 1, mean))
d_print [, 2] <-data.matrix(apply(coef_bias, 1, sd))

# print out coefficient bias

if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB", "DA Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_da_one_AB", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_da_one_AB",append=TRUE)
}
```

```
if (Imarker ==9) # Her2
{
cat( "\n","\n", "Marker = Her2", "DA Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_da_one_Her2", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_da_one_Her2",append=TRUE)
}
if (Imarker ==11) # PR
{
cat( "\n","\n", "Marker = PR", "DA Method ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Variable  ", "Mean   ", "SD   ", "\n",
file="E:\\s-project\\liu\\result\\SA_da_one_PR", append=TRUE)
write.table(round(d_print, digits=4), sep="   ",
file="E:\\s-project\\liu\\result\\SA_da_one_PR",append=TRUE)
}



# Print out sensitivity etc

if (Imarker ==1) # AB
{

cat( "\n","\n", "Marker = AB","DA Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Mean  ", mean(apply(result_sen, 2, mean)),
mean(apply(result_spe, 2, mean)), mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)),mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen, 2, mean)),
median(apply(result_spe, 2, mean)), median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)),median(apply(result_percent,2,mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{

cat( "\n","\n", "Marker = Her2","DA Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Mean  ", mean(apply(result_sen, 2, mean)),mean(apply(result_spe, 2, mean)),
mean(apply(result_ppv, 2, mean)),
```

```
            mean(apply(result_npv, 2, mean)), mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen,2,mean)), median(apply(result_spe, 2, mean)),
median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent, 2, mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_Her2", append=TRUE)
}

if (Imarker ==11) # PR
{

cat( "\n","\n", "Marker = PR","DA Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"Mean  ", mean(apply(result_sen, 2, mean)), mean(apply(result_spe, 2, mean)),
mean(apply(result_ppv, 2, mean)),
            mean(apply(result_npv, 2, mean)),mean(apply(result_percent, 2, mean)),"\n",
"Median", median(apply(result_sen, 2, mean)),
median(apply(result_spe, 2, mean)), median(apply(result_ppv, 2, mean)),
            median(apply(result_npv,2,mean)), median(apply(result_percent,2,mean)),"\n",
"SE    ", sd(apply(result_sen, 2, mean))/sqrt(nsampling),
sd(apply(result_spe, 2, mean))/sqrt(nsampling),
sd(apply(result_ppv, 2, mean))/sqrt(nsampling),
            sd(apply(result_npv, 2, mean))/sqrt(nsampling),
sd(apply(result_percent, 2, mean))/sqrt(nsampling),"\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_PR", append=TRUE)
}
# print out 10 year survival probability

if (Imarker ==1) # AB
{
cat( "\n","\n", "Marker = AB","DA Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survuval:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "Marker = Her2","DA Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survuval:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_Her2", append=TRUE)
}
```

```
if (Imarker ==11) # PR
{
cat( "\n","\n", "Marker = PR","DA Method  ", "Missing proportion = ",
pmiss[sam], sep="    ","\n",
"10 year survuval:     Mean  ", mean(apply(fit_impu, 2, mean)),
"SD  ", sd(apply(fit_impu, 2, mean)), "\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_PR", append=TRUE)
}


} ## end of missing proportion size

if (Imarker ==1) # AB
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_AB", append=TRUE)
}

if (Imarker ==9) # Her2
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_Her2", append=TRUE)
}

if (Imarker ==11) # PR
{
cat( "\n","\n", "True Results","\n",
"10 year survival:     Mean  ", fit_true$surv,
"SD  ", fit_true$std.er, "\n",
            file="E:\\s-project\\liu\\result\\SA_da_one_PR", append=TRUE)
}
```

D.2 R Program of Cox models

D.2.1 Using the dataset with missing (n=1575)

```
#############################################################################
#
# AREG METHOD for MI, Coxph for survival model
```

```
#
# This code removes all observations with missing clinical observations(n=1575)
#
# examples written in SAcourse_project_new.doc and survdiff for missing values.doc
#
#############################################################################

# difference with MAY10: add fastbw, survival calculation within inputat....

GPEC<-load("G:\\s-project\\liu\\imputBree")
#GPEC.df, Mark.m.df, Clin.m.df

dim (GPEC.df)

m.vars<-c("AB","BCL2","CK56","GATA","Her1","ERS","CA9","P53","Her2","IGFB","PR","YB1")
# X is missing

c.vars<-c("AGECAT","HISTCA","GRADECAT","ERPOSNE","LVNNE","SYS","PPNODECAT","SIZECAT")
#unknown or category 9 are missing. add SYS in

GPEC.cm.df <- GPEC.df[, c(m.vars,c.vars)]
# choose only clinical variables and markers

summary (GPEC.cm.df)

##remove the cases with missing clinical variables
miss<-(GPEC.df$GRADECAT=="9" | GPEC.df$ERPOSNE=="unknown " | GPEC.df$LVNNE=="9" |
   GPEC.df$SYS=="unknown" | GPEC.df$PPNODECAT=="unknown" | GPEC.df$SIZECAT=="unknown")
sum(miss==T)
GPEC.NCM.df<-GPEC.df[miss==FALSE,]
dim(GPEC.NCM.df)

#####Check the influence of missing values

try<-GPEC.NCM.df[,c(m.vars,c.vars)]

dim(try)
summary(try)

##set the levels of the variables in j
# all levels labeled c(1,2,NA) were originally 0,1,NA

for(i in 1:ncol(try)){
   if( i ==13 | i ==14)
       levels(try[,i])<-c(1, 2) # two levels for AGECAT and HISTCA
   else if (i ==15 | i ==19)
levels(try[,i])<-c(1, 2, 3, 9) # 4 levels for GRADECAT and PPNODECAT
   else
 levels(try[,i])<-c(1, 2, 9) # 3 levels for others
```

```
   try[,i]<-as.numeric(try[,i])
}
#summary(try)
try_numeric<-as.matrix(try) # keep the original one

# get categorical data
for(i in 1:ncol(try)){
    try[,i]<-as.factor(try[,i])
    if( i <13 )
 levels(try[,i])<-c(1,1, 2)   # remove NA's and keep Non-NA=1,NA =2
    else if (i ==15 |i ==19) # GRADECAT and PPNODECAT
 levels(try[,i])<-c(1,2,3)
    else if (i ==18) # SYS
 levels(try[,i])<-c(1,2,3,4)
    else
       levels(try[,i])<-c(1,2)
}

library(survival)

###############
# compare difference for NA and NON-NA
###############

full1<-cbind(try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(full1)<-c(colnames(try),"SURVYRS","BRDEATH")
summary (full1)

# number of BCSS records
sum(full1$BRDEATH==1)
# return 476

diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~AB,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~BCL2,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~CK56,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~GATA,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~Her1,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~ERS,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~CA9,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~P53,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~Her2,data=data.frame(full1))
```

```
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~IGFB,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~PR,data=data.frame(full1))
diff
diff<-survdiff(Surv(SURVYRS,BRDEATH==1)~YB1,data=data.frame(full1))
diff

full1<-cbind(try,GPEC.NCM.df$LOCSURV,GPEC.NCM.df$LOCSTAT)
colnames(full1)<-c(colnames(try),"LOCSURV","LOCSTAT")
summary (full1)

# number of LRS records
sum(full1$LOCSTAT==1)
# return 138

diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~AB,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~BCL2,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~CK56,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~GATA,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~Her1,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~ERS,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~CA9,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~P53,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~Her2,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~IGFB,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~PR,data=data.frame(full1))
diff
diff<-survdiff(Surv(LOCSURV,LOCSTAT==1)~YB1,data=data.frame(full1))
diff

full1<-cbind(try,GPEC.NCM.df$DISTSURV,GPEC.NCM.df$DISTSTAT)
colnames(full1)<-c(colnames(try),"DISTSURV","DISTSTAT")
summary (full1)

# number of DRS records
sum(full1$DISTSTAT==1)
# return 518
```

```
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~AB,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~BCL2,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~CK56,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~GATA,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~Her1,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~ERS,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~CA9,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~P53,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~Her2,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~IGFB,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~PR,data=data.frame(full1))
diff
diff<-survdiff(Surv(DISTSURV,DISTSTAT==1)~YB1,data=data.frame(full1))
diff

######## Prepare the data try for modeling and imputation
#Combine M and C
try<-GPEC.NCM.df[,c(m.vars,c.vars)]

dim(try)
summary(try)

##set the levels of the variables in j
# all levels labeled c(1,2,NA) were originally 0,1,NA

for(i in 1:ncol(try)){
   if( i ==13 | i ==14)
       levels(try[,i])<-c(1, 2) # two levels for AGECAT and HISTCA
   else if (i ==15 | i ==19)
levels(try[,i])<-c(1, 2, 3, 9) # 4 levels for GRADECAT and PPNODECAT
   else
 levels(try[,i])<-c(1, 2, 9) # 3 levels for others
   try[,i]<-as.numeric(try[,i])
}
#summary(try)
try_numeric<-as.matrix(try) # keep the original one

# get categorical data
for(i in 1:ncol(try)){
```

```
    try[,i]<-as.factor(try[,i])
    if( i <13 )
 levels(try[,i])<-c(1,2, NA)
    else if (i ==15 |i ==19) # GRADECAT and PPNODECAT
 levels(try[,i])<-c(1,2,3)
    else if (i ==18) # SYS
 levels(try[,i])<-c(1,2,3,4)
    else
       levels(try[,i])<-c(1,2)
}
try_factor<-as.matrix(try) # keep the categorical one

summary(try)  # 20 variables

library(cat)
rngseed(1898) #random generator must be set before imputation functions

##################################################

#Multiple Imputation using Bootstrap and PMM "aregImputation(Hmisc)"

##################################################

library(Hmisc)
set.seed(200) # Function of this one?
dimnames(try)[[2]]
nnumber = 6
#Multiple Imputation using Bootstrap and PMM "aregImputation(Hmisc)"y
#a$imputed[8]

a<- aregImpute(~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT
+AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1
,n.impute=nnumber, nk=3, data=try)

# A_coef_mean_reg<-rep(0, length(model0.reg$coef))
# A_coef_mean_cox<-rep(0, length(model0.reg$coef))
# A_pred_mean<-rep(0, nsize[sam])

#not necessary to run the j loop below; can give j=1 to j=nnumber
# and copy results to EXCEL or word
# then calculate coef mean and overall se(coef) according to formula in PDF

# Also bestbw may get different results for the imputations


j=1 #set j=1 to j=nnumber separately
#j=nnumber

#for( j in 1:nnumber) # start each imputation
```

```
{

t_try<-try

for(i in 1:20)  # 20 total variables
{
    t_try[,i]<-as.numeric(try[,i])
  if(i <13)   # AB starts from 1st column in "try" and 12 markers
    t_try[which(is.na(t_try[,i])),i]<-a$imputed[[i+7]][,j]
#AB starts from 8th column in "a"
t_try[,i]<-as.factor(t_try[,i])
}

summary (t_try)

library(survival)

# Full is called BCSS model
full<-cbind(try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(full)<-c(colnames(try),"SURVYRS","BRDEATH")
summary (full)
sum(full$BRDEATH==1) # how many with BCSS


# single variable check; clinical variable should get same results each imputat

model<-coxph(Surv(SURVYRS,BRDEATH==1)~AGECAT,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~SYS,y=TRUE,model=TRUE,data=full)
model
anova(model) # check significance
model<-coxph(Surv(SURVYRS,BRDEATH==1)~GRADECAT,y=TRUE,model=TRUE,data=full)
model
anova(model) # check significance
model<-coxph(Surv(SURVYRS,BRDEATH==1)~ERPOSNE,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~LVNNE,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~PPNODECAT,y=TRUE,model=TRUE,data=full)
model
anova(model) # check significance
model<-coxph(Surv(SURVYRS,BRDEATH==1)~SIZECAT,y=TRUE,model=TRUE,data=full)
model

model<-coxph(Surv(SURVYRS,BRDEATH==1)~AB,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~BCL2,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~CK56,y=TRUE,model=TRUE,data=full)
```

```
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~GATA,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~Her1,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~ERS,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~CA9,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~P53,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~Her2,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~IGFB,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~PR,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~YB1,y=TRUE,model=TRUE,data=full)
model

# full1 is with local recurr (years) under locstat==1
full1<-cbind(try,GPEC.NCM.df$LOCSURV,GPEC.NCM.df$LOCSTAT)
colnames(full1)<-c(colnames(try),"LOCSURV","LOCSTAT")
summary (full1)
sum(full1$LOCSTAT==1) # how many with LRF

model<-coxph(Surv(LOCSURV,LOCSTAT==1)~AGECAT,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~SYS,y=TRUE,model=TRUE,data=full1)
model
anova(model) # check significance
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~GRADECAT,y=TRUE,model=TRUE,data=full1)
model
anova(model) # check significance
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~ERPOSNE,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~LVNNE,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~PPNODECAT,y=TRUE,model=TRUE,data=full1)
model
anova(model) # check significance
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~SIZECAT,y=TRUE,model=TRUE,data=full1)
model

model<-coxph(Surv(LOCSURV,LOCSTAT==1)~AB,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~BCL2,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~CK56,y=TRUE,model=TRUE,data=full1)
```

```
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~GATA,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~Her1,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~ERS,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~CA9,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~P53,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~Her2,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~IGFB,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~PR,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~YB1,y=TRUE,model=TRUE,data=full1)
model

# full2 is with distact recurr (years) under distat==1
full2<-cbind(t_try,GPEC.NCM.df$DISTSURV,GPEC.NCM.df$DISTSTAT)
colnames(full2)<-c(colnames(t_try),"DISTSURV","DISTSTAT")
summary (full2)

model<-coxph(Surv(DISTSURV,DISTSTAT==1)~AGECAT,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~SYS,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~GRADECAT,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~ERPOSNE,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~LVNNE,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~PPNODECAT,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~SIZECAT,y=TRUE,model=TRUE,data=full2)
model

model<-coxph(Surv(DISTSURV,DISTSTAT==1)~AB,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~BCL2,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~CK56,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~GATA,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~Her1,y=TRUE,model=TRUE,data=full2)
```

```
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~ERS,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~CA9,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~P53,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~Her2,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~IGFB,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~PR,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~YB1,y=TRUE,model=TRUE,data=full2)
model

####################
##  try CPH for full model analysis
##################


library(Design)

# Full is called BCSS model
full<-cbind(t_try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(full)<-c(colnames(t_try),"SURVYRS","BRDEATH")
summary (full)

ddist<-datadist(full, adjto.cat=c('first'))

options(datadist='ddist')

model<-cph(Surv(SURVYRS,BRDEATH==1)~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1,x=T,y=T,model=T,data=full)
model

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er

anova(model) # check significance

#fastbw(model, rule="p", type="residual", sls=.1)
#fastbw(model, rule="aic", type="individual", sls=.1)
# this one close to backward elimination
fastbw(model, rule="p", type="individual", sls=.1)

#reduced model after fastbw
```

```
model1<-cph(Surv(SURVYRS,BRDEATH==1)~GRADECAT+LVNNE+PPNODECAT+SIZECAT+
           BCL2+CK56+Her2+YB1,x=T,y=T,model=T,data=full)
anova(model1) # check significance

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model1), c(10))
fit_true$surv
fit_true$std.er

# cross tabulation, get counts for each categories of variables
# both "ctab" and "ftable" work

ctab(xtabs(BRDEATH~AGECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~SYS, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~GRADECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~ERPOSNE, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~LVNNE, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~PPNODECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~SIZECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~AB, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~BCL2, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~CK56, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~GATA, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~Her1, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~ERS, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~CA9, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~P53, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~Her2, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~IGFB, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~PR, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~YB1, full, full$BRDEATH==1))

##############
###reduced model check

plot(survfit(model1), ylim=c(.0, 1), xlab="Years", ylab="Proportion of survival")
summary(survfit(model1))

# PH check

attributes(model1)  # see the names in model1

detail <- coxph.detail(model1)
time <- detail$y [ ,2] # ordered times including censored ones
status <- detail$y[ ,3]  # censoring status
sch <- residuals(model1, "scaledsch")    # Schoenfeld residuals
summary(sch)
plot (time[status==1], sch[ ,1], xlab = "Ordered survival time",
```

```
      ylab = "Schoenfeld residual for GRADECAT=2" )
plot (time[status==1], sch[ ,7], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for BCL2" )

temp<- cox.zph(model1)
print(temp)  #check time varying
#win.graph(width = 10, height = 16, pointsize = 10, restoreConsole = TRUE)
win.graph(width = 5, height = 8, pointsize = 12)
par(mfrow=c(4,3))
plot(temp) #plot curves

# influencial observation
dfbeta <- residuals(model1, type="dfbeta")
win.graph(width = 10, height = 16, pointsize = 12)
par(mfrow=c(4,3))
for (j in 1:10) {
plot(dfbeta[,j], ylab=names(coef(model1))[j])
#abline(h=0, lty=2)
 }

# overall fit check (Cox-Snell residual)
rc<- abs(full$BRDEATH - model1$residuals) # Cox-Snell
rc
km.rc<-survfit(Surv(rc,full$BRDEATH==1) ~1)
S.km.rc <- summary(km.rc)
rcu <- S.km.rc$time # Cox-Snell residuals of uncensored points
surv.rc <- S.km.rc$surv
plot (rcu, -log(surv.rc), type="p", pch = "*",
  xlab="Cox-Snell residual rc", ylab = "Cumulative harzard on rc")
abline(a=0, b=1); abline(v=0); abline(h=0)

# stratifing BCL2
model_t<-cph(Surv(SURVYRS,BRDEATH==1)~SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT+
           strata(BCL2)+CK56+Her2+YB1,x=T,y=T,model=T,data=full)
model_t
survfit(model_t)

######################
# full1 model with all variables, locsurv as the outcome
######################

full1<-cbind(t_try,GPEC.NCM.df$LOCSURV,GPEC.NCM.df$LOCSTAT)
colnames(full1)<-c(colnames(try),"LOCSURV","LOCSTAT")
summary (full1)

ddist<-datadist(full1, adjto.cat=c('first'))

options(datadist='ddist')
```

```
model<-cph(Surv(LOCSURV,LOCSTAT==1)~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1,x=T,y=T,model=T,data=full1)
model

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er

anova(model) # check significance

# this one close to backward elimination
fastbw(model, rule="p", type="individual", sls=.1)


#reduced model after fastbw

model1<-cph(Surv(LOCSURV,LOCSTAT==1)~GRADECAT+PPNODECAT+
           IGFB+YB1,x=T,y=T,model=T,data=full1)
model1

anova(model1) # check significance

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model1), c(10))
fit_true$surv
fit_true$std.er

plot(survfit(model1), ylim=c(.0, 1), xlab="Years", ylab="Proportion of survival")
summary(survfit(model1))

# PH check

attributes(model1)  # see the names in model1

detail <- coxph.detail(model1)
time <- detail$y [ ,2] # ordered times including censored ones
status <- detail$y[ ,3]  # censoring status
sch <- residuals(model1, "scaledsch")    # Schoenfeld residuals
summary(sch)
plot (time[status==1], sch[ ,4], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for PPNODECAT=3" )
plot (time[status==1], sch[ ,2], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for GRADECAT=3" )
plot (time[status==1], sch[ ,5], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for IGFB" )

temp<- cox.zph(model1)
print(temp)  #time varying check
```

```
win.graph(width = 16, height = 16, pointsize = 12)
par(mfrow=c(2,3))
plot(temp) #plot curves

# influencial observation
dfbeta <- residuals(model1, type="dfbeta")
par(mfrow=c(2,3))
for (j in 1:6) {
plot(dfbeta[,j], ylab=names(coef(model1))[j])
#abline(h=0, lty=2)
 }

# overall fit check (Cox-Snell residual)
rc<- abs(full1$LOCSTAT - model1$residuals) # Cox-Snell

km.rc<-survfit(Surv(rc,full1$LOCSTAT==1) ~1)
S.km.rc <- summary(km.rc)
rcu <- S.km.rc$time # Cox-Snell residuals of uncensored points
surv.rc <- S.km.rc$surv
plot (rcu, -log(surv.rc), type="p", pch = "*",
  xlab="Cox-Snell residual rc", ylab = "Cumulative harzard on rc")
abline(a=0, b=1); abline(v=0); abline(h=0)

# stratifing YB1
model_t<-coxph(Surv(LOCSURV,LOCSTAT==1)~GRADECAT+PPNODECAT+
           IGFB+strata(YB1),x=T, y=T,model=TRUE,data=full1)
model_t
survfit(model_t)

######################
# full2 model with all variables, distsurv as the outcome
######################

full2<-cbind(t_try,GPEC.NCM.df$DISTSURV,GPEC.NCM.df$DISTSTAT)
colnames(full2)<-c(colnames(try),"DISTSURV","DISTSTAT")
summary (full2)

ddist<-datadist(full1, adjto.cat=c('first'))

options(datadist='ddist')

model<-cph(Surv(DISTSURV,DISTSTAT==1)~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1,x=T,y=T,model=T,data=full2)
model

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er
```

```
anova(model) # check significance

# this one close to backward elimination
fastbw(model, rule="p", type="individual", sls=.1)

#reduced model after fastbw

model1<-cph(Surv(DISTSURV,DISTSTAT==1)~SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT+
           BCL2+CA9+Her2+YB1,x=T,y=T,model=T,data=full2)
model1

anova(model1) # check significance

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model1), c(10))
fit_true$surv
fit_true$std.er

plot(survfit(model1), ylim=c(.0, 1), xlab="Years", ylab="Proportion of survival")
summary(survfit(model1))

# PH check

attributes(model1)  # see the names in model1

detail <- coxph.detail(model1)
time <- detail$y [ ,2] # ordered times including censored ones
status <- detail$y[ ,3]  # censoring status
sch <- residuals(model1, "scaledsch")    # Schoenfeld residuals
summary(sch)

plot (time[status==1], sch[ ,4], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for GRADCAT2" )
plot (time[status==1], sch[ ,6], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for LVNNE" )
plot (time[status==1], sch[ ,11], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for CA9" )

temp<- cox.zph(model1)
print(temp)  #display the results
win.graph(width = 12, height = 16, pointsize = 12)
par(mfrow=c(3,5))
plot(temp) #plot curves

# influencial observation
dfbeta <- residuals(model1, type="dfbeta")
par(mfrow=c(3,5))
for (j in 1:13) {
```

```
plot(dfbeta[,j], ylab=names(coef(model1))[j])
#abline(h=0, lty=2)
 }

# overall fit check (Cox-Snell residual)
rc<- abs(full2$DISTSTAT - model1$residuals) # Cox-Snell

km.rc<-survfit(Surv(rc,full2$DISTSTAT==1) ~1)
S.km.rc <- summary(km.rc)
rcu <- S.km.rc$time # Cox-Snell residuals of uncensored points
surv.rc <- S.km.rc$surv
plot (rcu, -log(surv.rc), type="p", pch = "*",
  xlab="Cox-Snell residual rc", ylab = "Cumulative harzard on rc")
abline(a=0, b=1); abline(v=0); abline(h=0)

# stratifing Her2
model_t<-coxph(Surv(DISTSURV,DISTSTAT==1)~SYS+GRADECAT+LVNNE+PPNODECAT+SIZECAT+
           BCL2+CA9+strata(Her2)+YB1,y=TRUE,model=TRUE,data=full2)
model_t
}
```

D.2.2 Using the dataset without missing (n=910)

```
##############################################################################
#
# AREG METHOD for MI, Coxph for survival model

# This code removes all observations with missings(n=910)
#
# examples written in SAcourse_project_new.doc and survdiff for missing values.doc
#
##############################################################################

# difference with MAY10: add fastbw, ....#
# add cross tabulation "ctab" from "catspec" or "ftable"

#Libraries

library(Design)
library(survival)
library(cat)
library(Hmisc)
library(timereg)
library(catspec)

#Prepare data

GPEC<-load("G:\\s-project\\liu\\imputBree")
```

```
#GPEC.df, Mark.m.df, Clin.m.df

dim (GPEC.df)

m.vars<-c("AB","BCL2","CK56","GATA","Her1","ERS","CA9","P53", "Her2", "IGFB","PR","YB1")
# X is missing

c.vars<-c("AGECAT","HISTCA","GRADECAT","ERPOSNE","LVNNE","SYS","PPNODECAT","SIZECAT")
#unknown or category 9 are missing. add SYS in

GPEC.cm.df <- GPEC.df[, c(m.vars,c.vars)]
# choose only clinical variables and markers

summary (GPEC.cm.df)

##remove the cases with missing clinical variables

miss<-(GPEC.df$GRADECAT=="9" | GPEC.df$ERPOSNE=="unknown " | GPEC.df$LVNNE=="9" |
       GPEC.df$SYS=="unknown" | GPEC.df$PPNODECAT=="unknown" | GPEC.df$SIZECAT=="unknown"
| GPEC.df$AB=="X" | GPEC.df$CK56=="X" |
       GPEC.df$BCL2=="X" | GPEC.df$GATA=="X" | GPEC.df$Her1=="X" | GPEC.df$ERS=="X" |
       GPEC.df$CA9=="X" | GPEC.df$P53=="X" | GPEC.df$Her2=="X" | GPEC.df$IGFB=="X" |
       GPEC.df$PR=="X" | GPEC.df$YB1=="X")
sum(miss==T)
GPEC.NCM.df<-GPEC.df[miss==FALSE,]
dim(GPEC.NCM.df)

######## Prepare the data try for modeling and imputation

#Combine M and C
try<-GPEC.NCM.df[,c(m.vars,c.vars)]

dim(try)
summary(try)

##set the levels of the variables in j
# all levels labeled c(1,2,NA) were originally 0,1,NA

for(i in 1:ncol(try)){
   if( i ==13 | i ==14)
       levels(try[,i])<-c(1, 2) # two levels for AGECAT and HISTCA
   else if (i ==15 | i ==19)
levels(try[,i])<-c(1, 2, 3, 9) # 4 levels for GRADECAT and PPNODECAT
   else
 levels(try[,i])<-c(1, 2, 9) # 3 levels for others
   try[,i]<-as.numeric(try[,i])
}
#summary(try)
try_numeric<-as.matrix(try) # keep the original one
```

```
# get categorical data
for(i in 1:ncol(try)){
    try[,i]<-as.factor(try[,i])
    if( i <13 )
 levels(try[,i])<-c(1,2, NA)
    else if (i ==15 |i ==19) # GRADECAT and PPNODECAT
 levels(try[,i])<-c(1,2,3)
    else if (i ==18) # SYS
 levels(try[,i])<-c(1,2,3,4)
    else
       levels(try[,i])<-c(1,2)
}
try_factor<-as.matrix(try) # keep the categorical one

summary(try)  # 20 variables

# Full is called BCSS model
full<-cbind(try,GPEC.NCM.df$SURVYRS,GPEC.NCM.df$BRDEATH)
colnames(full)<-c(colnames(try),"SURVYRS","BRDEATH")
summary (full)
sum(full$BRDEATH==1) # how many with BCSS

# single variable check; clinical variable should get same results each imputat

model<-coxph(Surv(SURVYRS,BRDEATH==1)~AGECAT,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~SYS,y=TRUE,model=TRUE,data=full)
model
anova(model) # check significance
model<-coxph(Surv(SURVYRS,BRDEATH==1)~GRADECAT,y=TRUE,model=TRUE,data=full)
model
anova(model) # check significance
model<-coxph(Surv(SURVYRS,BRDEATH==1)~ERPOSNE,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~LVNNE,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~PPNODECAT,y=TRUE,model=TRUE,data=full)
model
anova(model) # check significance
model<-coxph(Surv(SURVYRS,BRDEATH==1)~SIZECAT,y=TRUE,model=TRUE,data=full)
model

model<-coxph(Surv(SURVYRS,BRDEATH==1)~AB,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~BCL2,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~CK56,y=TRUE,model=TRUE,data=full)
model
```

```
model<-coxph(Surv(SURVYRS,BRDEATH==1)~GATA,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~Her1,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~ERS,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~CA9,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~P53,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~Her2,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~IGFB,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~PR,y=TRUE,model=TRUE,data=full)
model
model<-coxph(Surv(SURVYRS,BRDEATH==1)~YB1,y=TRUE,model=TRUE,data=full)
model


# full1 is with local recurr (years) under locstat==1
full1<-cbind(try,GPEC.NCM.df$LOCSURV,GPEC.NCM.df$LOCSTAT)
colnames(full1)<-c(colnames(try),"LOCSURV","LOCSTAT")
summary (full1)
sum(full1$LOCSTAT==1) # how many with LRF

model<-coxph(Surv(LOCSURV,LOCSTAT==1)~AGECAT,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~SYS,y=TRUE,model=TRUE,data=full1)
model
anova(model) # check significance
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~GRADECAT,y=TRUE,model=TRUE,data=full1)
model
anova(model) # check significance
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~ERPOSNE,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~LVNNE,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~PPNODECAT,y=TRUE,model=TRUE,data=full1)
model
anova(model) # check significance
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~SIZECAT,y=TRUE,model=TRUE,data=full1)
model

model<-coxph(Surv(LOCSURV,LOCSTAT==1)~AB,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~BCL2,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~CK56,y=TRUE,model=TRUE,data=full1)
```

```
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~GATA,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~Her1,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~ERS,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~CA9,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~P53,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~Her2,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~IGFB,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~PR,y=TRUE,model=TRUE,data=full1)
model
model<-coxph(Surv(LOCSURV,LOCSTAT==1)~YB1,y=TRUE,model=TRUE,data=full1)
model

# full2 is with distact recurr (years) under distat==1
full2<-cbind(try,GPEC.NCM.df$DISTSURV,GPEC.NCM.df$DISTSTAT)
colnames(full2)<-c(colnames(try),"DISTSURV","DISTSTAT")
summary (full2)
sum(full2$DISTSTAT==1) # how many with DRF

model<-coxph(Surv(DISTSURV,DISTSTAT==1)~AGECAT,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~SYS,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~GRADECAT,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~ERPOSNE,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~LVNNE,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~PPNODECAT,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~SIZECAT,y=TRUE,model=TRUE,data=full2)
model

model<-coxph(Surv(DISTSURV,DISTSTAT==1)~AB,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~BCL2,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~CK56,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~GATA,y=TRUE,model=TRUE,data=full2)
model
```

```
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~Her1,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~ERS,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~CA9,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~P53,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~Her2,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~IGFB,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~PR,y=TRUE,model=TRUE,data=full2)
model
model<-coxph(Surv(DISTSURV,DISTSTAT==1)~YB1,y=TRUE,model=TRUE,data=full2)
model

######################
# Multivariable model, survyrs as the outcome
# Using CPH
######################

ddist<-datadist(full, adjto.cat=c('first'))

options(datadist='ddist')

model<-cph(Surv(SURVYRS,BRDEATH==1)~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1,x=T,y=T,model=T,data=full)
model

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er

anova(model) # check significance

#fastbw(model, rule="p", type="residual", sls=.1)
# this one close to backward elimination
fastbw(model, rule="p", type="individual", sls=.1)


#reduced model after fastbw
model1<-cph(Surv(SURVYRS,BRDEATH==1)~GRADECAT+ERPOSNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+Her2+YB1,x=T,y=T,model=T,data=full)
model1

anova(model1) # check significance
```

```
# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model1), c(10))
fit_true$surv
fit_true$std.er

# cross tabulation, get counts for each categories of variables
# both "ctab" and "ftable" work

ctab(xtabs(BRDEATH~AGECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~SYS, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~GRADECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~ERPOSNE, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~LVNNE, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~PPNODECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~SIZECAT, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~AB, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~BCL2, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~CK56, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~GATA, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~Her1, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~ERS, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~CA9, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~P53, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~Her2, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~IGFB, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~PR, full, full$BRDEATH==1))
ctab(xtabs(BRDEATH~YB1, full, full$BRDEATH==1))

##############
###reduced model check

plot(survfit(model1), ylim=c(.0, 1), xlab="Years", ylab="Proportion of survival")
summary(survfit(model1))

# PH check

attributes(model1)  # see the names in model1

detail <- coxph.detail(model1)
time <- detail$y [ ,2] # ordered times including censored ones
status <- detail$y[ ,3]  # censoring status
sch <- residuals(model1, "scaledsch")    # Schoenfeld residuals
summary(sch)
plot (time[status==1], sch[ ,3], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for ERPOSNE=2" )
plot (time[status==1], sch[ ,7], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for AB=2" )

temp<- cox.zph(model1)
```

```
print(temp)  #check time varying
#win.graph(width = 10, height = 16, pointsize = 10, restoreConsole = TRUE)
win.graph(width = 10, height = 16, pointsize = 12)
par(mfrow=c(4,3))
plot(temp) #plot curves

# influencial observation
dfbeta <- residuals(model1, type="dfbeta")
win.graph(width = 10, height = 16, pointsize = 12)
par(mfrow=c(4,3))
for (j in 1:11) {
plot(dfbeta[,j], ylab=names(coef(model1))[j])
#abline(h=0, lty=2)
 }

# overall fit check (Cox-Snell residual)
rc<- abs(full$BRDEATH - model1$residuals) # Cox-Snell
rc
km.rc<-survfit(Surv(rc,full$BRDEATH==1) ~1)
S.km.rc <- summary(km.rc)
rcu <- S.km.rc$time # Cox-Snell residuals of uncensored points
surv.rc <- S.km.rc$surv
plot (rcu, -log(surv.rc), type="p", pch = "*",
  xlab="Cox-Snell residual rc", ylab = "Cumulative harzard on rc")
abline(a=0, b=1); abline(v=0); abline(h=0)

# stratifing erponse
model_t<-coxph(Surv(SURVYRS,BRDEATH==1)~GRADECAT+strata(ERPOSNE)+PPNODECAT+SIZECAT+AB+
           BCL2+CK56+Her2+YB1,x=T,y=T,model=T,data=full)
model_t
survfit(model_t)


######################
# full1 model with all variables, locsurv as the outcome
######################
full1<-cbind(try,GPEC.NCM.df$LOCSURV,GPEC.NCM.df$LOCSTAT)
colnames(full1)<-c(colnames(try),"LOCSURV","LOCSTAT")
summary (full1)

ddist<-datadist(full1, adjto.cat=c('first'))

options(datadist='ddist')


model<-cph(Surv(LOCSURV,LOCSTAT==1)~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1,x=T,y=T,model=T,data=full1)

model
```

```
# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model), c(10))
fit_true$surv
fit_true$std.er

anova(model) # check significance


#fastbw(model, rule="p", type="residual", sls=.1)
# this one close to backward elimination
fastbw(model, rule="p", type="individual", sls=.1)


#reduced model after fastbw

model1<-cph(Surv(LOCSURV,LOCSTAT==1)~GRADECAT+PPNODECAT+
           IGFB+YB1,x=T,y=T,model=T,data=full1)
model1

anova(model1) # check significance

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model1), c(10))
fit_true$surv
fit_true$std.er

# cross tabulation, get counts for each categories of variables
# both "ctab" and "ftable" work

ctab(xtabs(LOCSTAT~AGECAT, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~SYS, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~GRADECAT, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~ERPOSNE, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~LVNNE, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~PPNODECAT, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~SIZECAT, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~AB, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~BCL2, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~CK56, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~GATA, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~Her1, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~ERS, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~CA9, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~P53, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~Her2, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~IGFB, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~PR, full1, full1$LOCSTAT==1))
ctab(xtabs(LOCSTAT~YB1, full1, full1$LOCSTAT==1))
```

```
###############
###reduced model check

plot(survfit(model1), ylim=c(.0, 1), xlab="Years", ylab="Proportion of survival")
summary(survfit(model1))

# PH check

attributes(model1)  # see the names in model1

detail <- coxph.detail(model1)
time <- detail$y [ ,2] # ordered times including censored ones
status <- detail$y[ ,3]  # censoring status
sch <- residuals(model1, "scaledsch")    # Schoenfeld residuals
summary(sch)
plot (time[status==1], sch[ ,4], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for PPNODECAT=3" )
plot (time[status==1], sch[ ,2], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for GRADECAT=3" )
plot (time[status==1], sch[ ,5], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for IGFB" )

temp<- cox.zph(model1)
print(temp)  #time varying check
win.graph(width = 16, height = 16, pointsize = 12)
par(mfrow=c(2,3))
plot(temp) #plot curves

# influencial observation
dfbeta <- residuals(model1, type="dfbeta")
par(mfrow=c(2,3))
for (j in 1:6) {
plot(dfbeta[,j], ylab=names(coef(model1))[j])
#abline(h=0, lty=2)
 }

# overall fit check (Cox-Snell residual)
rc<- abs(full1$LOCSTAT - model1$residuals) # Cox-Snell

km.rc<-survfit(Surv(rc,full1$LOCSTAT==1) ~1)
S.km.rc <- summary(km.rc)
rcu <- S.km.rc$time # Cox-Snell residuals of uncensored points
surv.rc <- S.km.rc$surv
plot (rcu, -log(surv.rc), type="p", pch = "*",
  xlab="Cox-Snell residual rc", ylab = "Cumulative harzard on rc")
abline(a=0, b=1); abline(v=0); abline(h=0)

# stratifing YB1
```

```
model_t<-coxph(Surv(LOCSURV,LOCSTAT==1)~GRADECAT+PPNODECAT+
           IGFB+strata(YB1),x=T, y=T,model=TRUE,data=full1)
model_t
survfit(model_t)

######################
# full2 model with all variables, distsurv as the outcome
######################

full2<-cbind(try,GPEC.NCM.df$DISTSURV,GPEC.NCM.df$DISTSTAT)
colnames(full2)<-c(colnames(try),"DISTSURV","DISTSTAT")
summary (full2)

ddist<-datadist(full1, adjto.cat=c('first'))

options(datadist='ddist')

model<-cph(Surv(DISTSURV,DISTSTAT==1)~AGECAT+SYS+GRADECAT+ERPOSNE+LVNNE+PPNODECAT+SIZECAT+
           AB+BCL2+CK56+GATA+Her1+ERS+CA9+P53+Her2+IGFB+PR+YB1,x=T,y=T,model=T,data=full2)
model

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model_t), c(10))
fit_true$surv
fit_true$std.er

anova(model) # check significance

#fastbw(model, rule="p", type="residual", sls=.1)
# this one close to backward elimination
fastbw(model, rule="p", type="individual", sls=.1)

#reduced model after fastbw

model1<-cph(Surv(DISTSURV,DISTSTAT==1)~GRADECAT+PPNODECAT+SIZECAT+
           BCL2+P53+Her2+YB1,x=T,y=T,model=T,data=full2)

model1

anova(model1) # check significance

# use this one for overall 10 year survival probability
fit_true<-summary(survfit(model1), c(10))
fit_true$surv
fit_true$std.er

# cross tabulation, get counts for each categories of variables
# both "ctab" and "ftable" work
```

```
ctab(xtabs(DISTSTAT~AGECAT, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~SYS, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~GRADECAT, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~ERPOSNE, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~LVNNE, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~PPNODECAT, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~SIZECAT, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~AB, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~BCL2, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~CK56, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~GATA, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~Her1, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~ERS, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~CA9, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~P53, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~Her2, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~IGFB, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~PR, full2, full2$DISTSTAT==1))
ctab(xtabs(DISTSTAT~YB1, full2, full2$DISTSTAT==1))

###############
###reduced model check

plot(survfit(model1), ylim=c(.0, 1), xlab="Years", ylab="Proportion of survival")
summary(survfit(model1))

# PH check

attributes(model1)  # see the names in model1

detail <- coxph.detail(model1)
time <- detail$y [ ,2] # ordered times including censored ones
status <- detail$y[ ,3]  # censoring status
sch <- residuals(model1, "scaledsch")    # Schoenfeld residuals
summary(sch)

plot (time[status==1], sch[ ,2], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for GRADCAT3" )
plot (time[status==1], sch[ ,5], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for SIZECAT" )
plot (time[status==1], sch[ ,8], xlab = "Ordered survival time",
      ylab = "Schoenfeld residual for Her2" )

temp<- cox.zph(model1)
print(temp)  #display the results
win.graph(width = 16, height = 16, pointsize = 12)
par(mfrow=c(3,3))
plot(temp) #plot curves
```

```
# influencial observation
dfbeta <- residuals(model1, type="dfbeta")
par(mfrow=c(3,3))
for (j in 1:9) {
plot(dfbeta[,j], ylab=names(coef(model1))[j])
#abline(h=0, lty=2)
 }

# overall fit check (Cox-Snell residual)
rc<- abs(full2$DISTSTAT - model1$residuals) # Cox-Snell

km.rc<-survfit(Surv(rc,full2$DISTSTAT==1) ~1)
S.km.rc <- summary(km.rc)
rcu <- S.km.rc$time # Cox-Snell residuals of uncensored points
surv.rc <- S.km.rc$surv
plot (rcu, -log(surv.rc), type="p", pch = "*",
  xlab="Cox-Snell residual rc", ylab = "Cumulative harzard on rc")
abline(a=0, b=1); abline(v=0); abline(h=0)

# stratifing Her2
model_t<-coxph(Surv(DISTSURV,DISTSTAT==1)~GRADECAT+PPNODECAT+SIZECAT+
           strata(BCL2)+P53+Her2+YB1,y=TRUE,model=TRUE,data=full2)
model_t
```